BIPEP: Sequence-based prediction of biofilm inhibitory peptides using combination of NMR and Physicochemical descriptors

Fereshteh Fallah Atanaki[a], Saman Behrouzi[a], Shohreh Ariaeenejad[b], Amin Boroomand[c] and Kaveh Kavousi[a]*

[a] Laboratory of Complex Biological Systems and Bioinformatics (CBB), Department of Bioinformatics, Institute of Biochemistry and Biophysics (IBB), University of Tehran, Tehran, Iran.

[b] Department of Systems and Synthetic Biology, Agricultural Biotechnology Research Institute of Iran (ABRII), Karaj, Iran

[c] School of Natural Sciences, University of California Merced, Merced, California, United States of America

[*] Corresponding author: Tel.: +98 21 88993978; Fax: +98 21 66404680. E-mail: kkavousi@ut.ac.ir

Supporting Information

Detailed information about computation of different feature vectors

Since the peptide sequences are the strings of amino acids, they need to be mapped onto numeric feature vectors before being used as an input in supervised learning classifiers. In this study, many different categories of features are computed.

Amino Acid Composition (AAC):

AAC defined as fraction of each of the amino acids present in a given peptide/protein sequence. This feature can be computed by using the following formula:

$$AAC(i) = \frac{\text{frequency of Amino Acid } i}{\text{length of peptide}} \quad \text{(S1)}$$

In above formula I can be any natural amino acid. This feature set has a length of 20 features.

Dipeptide composition (DPC):

DPC feature represents the total number of dipeptide divided by all the possible combinations of dipeptides present in the given protein/peptide sequence. DPC has a length of 400 (20 × 20) features which can be calculated using the following equation:

$$DPC(i) = \frac{\text{total number of dipeptide } i}{\text{total number of all possible dipeptide}} \quad \text{(S2)}$$

Composition, Transition and distribution (CTD):

These features developed by Dubchak et al in 1995 [1]. The first step is to break the amino acid into three different classes. The attributes used in the study include hydrophobicity, normalized van der Waals volume, polarity, and polarizability, as in the below table. The corresponding division is in the below table.

Table S1: different Amino Acid attributes and the Division of Amino Acid into three clusters

Property	**Class 1**	**Class 2**	**Class 3**
Hydrophobicity	Polar RKEDQN	Neutral GASTPHY	Hydrophobicity CLVIMFW
Normalized van der Waals volume	0-2.78 GASTPD	2.95-4.0 NVEQIL	4.03-8.08 MHKFRYW
Polarity	4.9-6.2 LIFWCMVY	8.0-9.2 PATGS	10.4-13.0 HQRKNED
Polarizability	0-1.08 GASDT	0.128-0.186 CPNVEQIL	0.219-0.409 KMHFRYW
Charge	Positive KR	Neutral ANCQGHILMFPSTW YV	Negative DE
Secondary structure	Helix EALMQKRH	Strand VIYCWFT	Coil GNPSD
Solvent accessibility	Buried ALFCGIVW	Exposed PKQEND	Intermediate MPSTHY
Surface tension	-0.20~0.16 GQDNAHR	-0.3~ -0.52 KTSEC	-0.98~ -2.46 ILMFPWYV
Protein-protein interface hotspot propensity -Bogan	High (5-21%) DHIKNPRWY	Medium (1.12-3.64%) EQSTGAMF	Low (0-0.83%) CLV
Protein-protein interface propensity -Ma	High (1.21-2.02) CDFMPQRWY	Medium (0.63-1.12) AGHVLNST	Low (0.14-0.29) EIK
Protein-DNA interface propensity -Schneider	High (4-30%) GKNQRSTY	Medium (1-3%) ADEFHILVW	Low (0-1%) CMP
Protein-DNA interface propensity -Ahmad	High (25-100%) GHKNQRSTY	Medium (5-18%) ADEFIPVW	Low (0-4%) CLM
Protein-RNA interface propensity -Kim	High (0.25-11) HKMRY	Medium (-0.25 –0.17) FGILNPQSVW	Low (-0.3 --0.8) CDEAT
Protein-RNA interface propensity -Ellis	High (1.18-2.07) HGKMRSYW	Medium (0.84-1.16) AFINPQT	Low (0.41-0.8) CDELV
Protein-RNA interface propensity -Phipps	High (0.95-1.8) HKMQRS	Medium (0.5-0.95) ADEFGLNPVY	Low (0-0.5) CITW
Protein-ligand binding site propensity -Khazanov	High (≥1.4) CFHWYM	Medium (0.79-1.21) DGILNSTV	Low (≤0.76) AEKPQR
Protein-ligand binding site propensity -Khazanov	High (477-1197) DEHRY	Medium (95-423) CFKMNQSTW	Low (<95) AGILPV
Molecular Weight	Low (75-105) AGS	Medium (115-155) CDEHIKLMNQPTV	High (165-204) FRWY
cLogP	-4.2 --3.3	-3.07 –2.26	-1.78 --1.05

	RKDNEQH	PYSTGACV	WMFLI
No of hydrogen bond donor in side chain	>1 HKNQR	1 DESTWY	0 ACGFILMPV
No of hydrogen bond acceptor in side chain	>1 DEHNQR	1 KSTWY	0 ACGFILMPV
Solubility in water	High (9-65 g/100g) ACGKRT	Medium (1.14-7.44 g/100g) EFHILMNPQSVW	Low (0.048-0.82 g/100g) DY
Amino acid flexibility index	Very flexible EGKNQS	Moderately flexible ADHIPRTV	Less flexible CFLMWY

Each ID of the Table S1 has a different "1" or "2", or "3" attribute that represents three different feature categories: "1: Composition (C) " 2: "Transition (T)", and 3: "Distribution (D)". These feature vectors computed from PyDPI 1.0. This package computes only some of the descriptors listed in the Table S1. So, in order to compute more features we changed the source code and extracted all 504 features. The new CTD code implements on our website in the feature selection part. Calculation details for a given attribute are as follows:

Composition:

For each encoded class in sequence, it is the global percent.

$$C_c = \frac{n_c}{\mathrm{N}} \quad c = 1,2,3 \quad \text{(S3)}$$

Where n_c the number of c in the encoded sequence and N is the length of the sequence.

Transition:

A transition from class 1 to 2 is the percent frequency with which 1 is followed by 2 or 2 is followed by 1 in the encoded sequence. Transition descriptor can be calculated as:

$$T_{rs} = \frac{n_{sp} + n_{ps}}{\mathrm{N\text{-}1}} \quad sp = 12,13,23 \quad \text{(S4)}$$

Where n_{sp}, n_{ps} is the numbers of dipeptide encoded as “sp” and “ps” respectively in the sequence and N is the length of the sequence.

Distribution:

Finally, distribution of each attribute in the sequence describes with distribution” feature. There are five “distribution” descriptors for each attribute and they are the position percents in the whole sequence for the first residue, 25% residues, 50% residues, 75% residues and 100% residues , respectively, for a specified encoded class.

The NMR based features:

The NMR based features for amino acids:

First, 34 features were calculated using NMR dataset[2] with respect to the following equations categorized as: Relative Spectral Power (RSP), Slow Wave Index (SWI), Harmonic Parameters, Hjorth, Entropy, Skewness, and Kurtosis.

Relative Spectral power (RSP): This feature is measured based on the Eq. S5

$$Relative\,Spectral\,Power\left(RSP\right)=\frac{\int_{-f_1}^{-f_0}S_X\left(f\right)df+\int_{f_0}^{f_1}S_X\left(f\right)df}{\int_{-\infty}^{\infty}S_X\left(f\right)df} \qquad \text{(S5)}$$

Where the numerator is the Absolute Spectral Power for the frequency (from $f_0\ Hz$ to $f_1 Hz$) of NMR signals normalized to the total power spectral density ($S_X(f)$). $S_X(f)$ is defined as $|X(f)|^2$ when $X(f)$ is the Fourier transform of signal $x(t)$.

Slow Wave Index (SWI): SWI is defined by Eqs. S6, S7, and S8, where BSP_{Alpha}, BSP_{Delta}, BSP_{Theta}, are the Sub-Band Spectral Power (Table S2)[3], and DSI, TSI and ASI are the Delta-Slow-wave Index, Theta-Slow-wave Index and the Alpha-Slow-wave Index, respectively.

$$DSI=\frac{BSP_{Delta}}{BSP_{Theta}+BSP_{Alpha}} \qquad (S6)$$

$$TSI=\frac{BSP_{Theta}}{BSP_{Delta}+BSP_{Alpha}} \qquad (S7)$$

$$ASI=\frac{BSP_{Aipha}}{BSP_{Theta}+BSP_{Delta}} \qquad (S8)$$

Table S2: Frequency sub-bands used in RSP computation.

Bands	Sub-bands
Delta	Delta 1 Delta 2
Theta	Theta 1 Theta 2
Alpha	Alpha 1 Alpha 2
Sigma	Sigma 1 Sigma 2
Beta	Beta 1 Beta 2

Harmonic Parameters: The harmonic parameters of center frequency (f_c), bandwidth (f_σ) and spectral value at center frequency (S_{f_c}), allow the analysis of a specific band in spectrums through Eqs (S9,S10, and S11)[4]:

$$f_c = \frac{\sum_{f_L}^{f_H} f S_X(f)}{\sum_{f_L}^{f_H} S_X(f)} \quad (S9)$$

$$f_\sigma = \frac{\sum_{f_L}^{f_H} (f - f_c)^2 S_X(f)}{\sum_{f_L}^{f_H} S_X(f)} \quad (S10)$$

$$S_{f_c} = S_X(f_c) \quad (S11)$$

Where, $S_X(f)$ is the PSD (power spectral density) of Fourier transform of $x(t)$ computed for $\{f_H, f_L\}$ band frequencies.

Hjorth parameter: The hidden information from time series signals are extracted through this feature according to the Activity, Mobility, and Complexity parameters represented in Eqs. S12, S13, and S14.

The activity parameter is the signal power, indicating the variance of a time function:

$$Activity = var(x(t)) \quad (S12)$$

Where $x(t)$ is the signal.

Mobility shows the mean frequency of the power spectrum:

$$Mobility = \sqrt{\frac{var\left(\frac{dx(t)}{dt}\right)}{var(x(t))}} \quad (S13)$$

The Complexity parameter compares the signals similarity to a pure sine wave, where the value converges into one if the signal is close to the main sine function:

$$complexity = \frac{Mobility\left(\frac{dx(t)}{dt}\right)}{Mobility(x(t))} \quad (S14)$$

Entropy: The relative degree of randomness is described by this feature and measured by Eq.S15.

$$H(x) = -\sum_{i=1}^{N} p(x_i) log_{10} p(x_i) \quad (S15)$$

x is a random variable with N possible outcomes and $p(x_i)$ is the probability outcome i.

Skewness: In probability theory and statistics, skewness indicates the imbalance and asymmetry of the data distribution mean value. The skewness value can be positive or negative, or even undefined, computed as follows:

$$x_{ske} = \frac{\sum_{n=1}^{N}\left(x(n)-x_m\right)^3}{(N-1)x_{sd}^3} \quad (S16)$$

Where, N is the length of signal$x$, x_m is the mean value and x_{std} is the standard deviation of x.

Kurtosis: Kurtosis is a statistical measure applied in describing the data distribution, or their skewness, of the observed data around the mean, expressed as:

$$x_{krt} = \frac{\sum_{n=1}^{N}\left(x(n)-x_m\right)^4}{(N-1)x_{sd}^4} \quad (S17)$$

All features extracted from NMR signals and their counts are tabulated in Table S3.

Table S3: Name and dimensionality of feature vectors extracted from NMR signals

Features	Dimension of feature vector
RSP	10
SWI	3
HP	15
Hjorth	3
Entropy	1
Skewness	1
Kurtosis	1

Clustering of amino acids based on their NMR features

Fuzzy c-means (FCM) clustering algorithm was run five times independently to cluster all natural amino acids except tyrosine into 2, 3, 4, 5, and 6 clusters based on the 34 features obtained from NMR (Figure S1). Due to the lack of C-NMR spectra, we manually assigned this amino acid to each of which cluster leading to the best performance in AMP classification task. FCM is a method of clustering which allows one object to be simultaneously clustered in more than one group with different membership scores [5,6].

Figure S1 Different amino acid clusters obtained from NMR signals based on FCM algorithm.

The NMR based features for peptides

The results of the above-mentioned clustering were applied to extract feature vectors for peptides. The pattern of composition (C), transition (T), and distribution (D) for the members of clusters along the peptide sequences were used to make the descriptors for each peptide[7].

"C" describes the global frequency of the members for each generated cluster in the peptide sequence. "T" is the percentage of transitions from the members of one cluster to another which occurs along the sequences. "D" describes the distribution of the members of each cluster in the sequence. Five descriptors were assigned to each cluster based on the position percent in the whole peptide primary structure; i.e., the first residue, 25% residues, 50% residues, 75% residues and 100% residues. Table S4 demonstrates number of features for constituent parts of NMR based descriptors [8]. For constructing the NMR based descriptor, by adjusting the number of clusters, five different clustering solutions were obtained. As a result, amino acids were grouped into 2, 3, 4, 5, and 6 clusters (see Figure S1), and on the basis of composition, transition, and distribution of amino acids along the peptide sequence, five different feature vectors were also calculated.

Table S4*:* The length of feature vectors based on number of clusters.

Features	**Composition (C)**	**Transition (T)**	**Distribution (D)**	**Feature vector length** $n(C) + \frac{n \times (n-1)}{2(T)} + 5 \times n(D)$
2 clusters	2	1	10	13
3 clusters	3	3	15	21
4 clusters	4	6	20	30
5 clusters	5	10	25	40
6 clusters	6	15	30	51

Table S5: Name and group of 150 best features.

Feature Name	Feature Group
Composition of R	AAC
Composition of N	AAC
Composition of D	AAC
Composition of E	AAC
Composition of K	AAC
Composition of F	AAC
Composition of S	AAC
Composition of T	AAC
Composition of W	AAC
Composition of RI	DPC
Composition of RS	DPC
Composition of NR	DPC
Composition of QA	DPC
Composition of LF	DPC
Composition of FN	DPC
Composition of FT	DPC
Composition of SG	DPC
Composition of SL	DPC
Composition of SF	DPC
Composition of ST	DPC
Composition of TQ	DPC
Composition of TF	DPC
PolarizabilityC1	CTD
PolarizabilityC3	CTD
SolventAccessibilityC2	CTD
SolventAccessibilityC3	CTD
SecondaryStrC1	CTD
SecondaryStrC3	CTD
ChargeC1	CTD
ChargeC2	CTD
ChargeC3	CTD
PolarityC2	CTD
PolarityC3	CTD
NormalizedVDWVC1	CTD
NormalizedVDWVC3	CTD
HydrophobicityC1	CTD
HydrophobicityC2	CTD
PPIHotspotPropBoganC1	CTD

PPIHotspotPropBoganC2	CTD
PPIPropMaC2	CTD
PPIPropMaC3	CTD
PRNAIPropKimC1	CTD
PRNAIPropKimC3	CTD
PRNAIPropEllisC1	CTD
PRNAIPropEllisC2	CTD
PRNAIPropPhippsC1	CTD
PRNAIPropPhippsC2	CTD
PLVBSKhazanovC2	CTD
PLVBSKhazanovC3	CTD
MolecularWeightC1	CTD
cLogPC1	CTD
cLogPC2	CTD
NoHydroBondDonorSideChainC1	CTD
NoHydroBondDonorSideChainC2	CTD
SolubilityInWaterC1	CTD
SolubilityInWaterC2	CTD
PolarizabilityT12	CTD
SolventAccessibilityT12	CTD
SolventAccessibilityT13	CTD
SolventAccessibilityT23	CTD
SecondaryStrT12	CTD
SecondaryStrT13	CTD
SecondaryStrT23	CTD
ChargeT12	CTD
ChargeT23	CTD
PolarityT12	CTD
PolarityT13	CTD
NormalizedVDWVT12	CTD
NormalizedVDWVT13	CTD
HydrophobicityT13	CTD
HydrophobicityT23	CTD
SurfaceTensionT13	CTD
PPIHotspotPropBoganT13	CTD
PDNAIPropSchneiderT12	CTD
PDNAIPropAhmadT23	CTD
PRNAIPropKimT12	CTD
PRNAIPropKimT23	CTD
PRNAIPropEllisT23	CTD
PRNAIPropPhippsT13	CTD

PLVBSKhazanovT12	CTD
PropPLPANBIntImaiT13	CTD
cLogPT13	CTD
cLogPT23	CTD
NoHydroBondDonorSideChainT13	CTD
NoHydroBondDonorSideChainT23	CTD
SolubilityInWaterT12	CTD
SolubilityInWaterT23	CTD
PolarizabilityD3001	CTD
PolarizabilityD3025	CTD
PolarizabilityD3100	CTD
SolventAccessibilityD1001	CTD
SolventAccessibilityD2075	CTD
SolventAccessibilityD2100	CTD
SecondaryStrD1025	CTD
SecondaryStrD2001	CTD
SecondaryStrD2025	CTD
SecondaryStrD3100	CTD
ChargeD1075	CTD
ChargeD1100	CTD
PolarityD1001	CTD
PolarityD2025	CTD
PolarityD3075	CTD
PolarityD3100	CTD
NormalizedVDWVD3001	CTD
NormalizedVDWVD3025	CTD
NormalizedVDWVD3100	CTD
HydrophobicityD1075	CTD
HydrophobicityD1100	CTD
SurfaceTensionD3001	CTD
PPIHotspotPropBoganD1001	CTD
PPIHotspotPropBoganD1025	CTD
PPIHotspotPropBoganD1075	CTD
PPIHotspotPropBoganD1100	CTD
PPIPropMaD1001	CTD
PRNAIPropKimD1075	CTD
PRNAIPropKimD1100	CTD
PRNAIPropKimD3025	CTD
PRNAIPropEllisD1100	CTD
PRNAIPropPhippsD1075	CTD
PRNAIPropPhippsD1100	CTD

PLBSPropKhazanovD1001	CTD
PLBSPropKhazanovD3001	CTD
PLBSPropKhazanovD3025	CTD
PLVBSKhazanovD3100	CTD
MolecularWeightD3001	CTD
cLogPD1075	CTD
cLogPD1100	CTD
NoHydroBondDonorSideChainD1075	CTD
NoHydroBondDonorSideChainD1100	CTD
AminoAcidFlexIndD2001	CTD
F3	NMR
F8	NMR
F9	NMR
F11	NMR
F12	NMR
F13	NMR
F21	NMR
F22	NMR
F23	NMR
F24	NMR
F36	NMR
F38	NMR
Aliphatic	PCP
Charged	PCP
Basic	PCP
PI	PCP
ChargeInPH8	PCP
Number of Carbon	PCP
Number of Hydrogen	PCP
Number of Nitrogen	PCP

Table S6: First independent dataset for validation**.**

Peptides for independent validation	**Sequence**
BIP1	ILSAIWSGIKSLF
BIP2	KTKKKLLKKT
BIP3	DGVKLCDVPSGTWSGHCGSSSKCSQQCKDREHFAYGGACHYQFP SVKCFCKRQC
BIP4	ALWKEVLKNAGKAALNEINNLV
BIP5	NKGCSACAIGAACLADGPIPDFEVAGITGTFGIAS
BIP6	FFRNLWKGAKAAFRAGHAAWRA

BIP7	CVNWKKILGKIIKVVK
BIP8	CWFWKWWRRRRR
BIP9	EVASFDKSKLK
BIP10	FFGKVLKLIRKIF
not_BIP1	DIIIIFPPFG
not_BIP2	TREWDG
not_BIP3	YTNGNWVPS
not_BIP4	MLDWKY
not_BIP5	MMDWHY
not_BIP6	GIFWEQ
not_BIP7	DVRSNKIRLWWENIFFNKK
not_BIP8	SAVDWWRL
not_BIP9	EFDWWNLG
not_BIP10	DIFKLVIDHISMKARKK

Table S7: Second independent dataset for validation**.**

Peptides for independent validation	**Sequence**
BIP_1	VRLIVAVRIWRR
BIP_2	QRWKKWKVLKLR
BIP_3	KVVWWKVIIKVL
BIP_4	KIWLLKLRQRQK
BIP_5	WRIKKQWIQIIV
BIP_6	VARWKIIIAKLW
BIP_7	VQWIQIVVWRKR
BIP_8	KVQIIKQLIAKK
BIP_9	ILVRWIRWRIQW
BIP_10	VIKVLIKRWLKL
BIP_11	RRIIKILLWKLR
BIP_12	KKWQLLIKWKLR
BIP_13	IWLRLKVVLKRK
BIP_14	IILKRVQVQKIK
BIP_15	KRIKKLLKVVLK
BIP_16	QQKVIRLLWKAK
BIP_17	KRLQWVKVKKIR
BIP_18	VLQIKKVLRLLL
BIP_19	RIWRRAWKARWK
BIP_20	KIVIRIILQVIK
BIP_21	KIKLIQKQLRIK
BIP_22	WWIKIVVIRVRR
BIP_23	VLKIKVKIWVVK

BIP_24	WKKVQWLKRLLL
BIP_25	IKIVRRAKIIIW
BIP_26	VIKWLLKILRAI
BIP_27	GLIIKIIKKRLW
BIP_28	IQIWIIRVIWRW
BIP_29	LLKLKQKGIVIA
BIP_30	IIKWIVVRQIRK
BIP_31	WLKRIVKVVVLK
BIP_32	KVIQWIIVRRVL
BIP_33	QWLVKWVIIKVV
BIP_34	VQRIIWLRVKIV
BIP_35	QQVKWWLIRWLA
BIP_36	RVLIKWKKVIVV
BIP_37	KVIKIVLVRVVK
BIP_38	IKWVLRKIVQII
BIP_39	IQRWWKVWLKVI
BIP_40	VKWKGKVIVVQL
BIP_41	LKLKAILKIIRV
BIP_42	LIVIQLLKKWWK
BIP_43	RVKAIKWRKIVV
BIP_44	IKIIWKALGQVI
BIP_45	GKLKIKVKLGIA
BIP_46	KGKIRKIVLIRR
BIP_47	WIIRWIKIWLKI
BIP_48	IVKKVKLIWGVK
BIP_49	IQLKLIWVKRKW
BIP_50	VAKVKKARWRLR
BIP_51	RQVRVKRWRARW
BIP_52	KIVQKKLRLVVI
BIP_53	QIIKVVWRAVII
BIP_54	QVVVKKKAIQVV
BIP_55	IRILVLRKAIIV
BIP_56	KKQKKIWRRILV
BIP_57	LWQLWLKLKLKG
BIP_58	LQRVIWQKWRKV
BIP_59	RRQWRGWVRIWL
BIP_60	RGARVIRWKLRR
BIP_61	IAWQLLWGWRVR
BIP_62	KRKQWKLWVRQI
BIP_63	KLLGILKQAIVV
BIP_64	WQGWAKIWVVRI

BIP_65	LKKIIVQAVGLI
BIP_66	IGQVVLVKIKIA
BIP_67	ALAIKVWIKILQ
BIP_68	VIAKIVLLRAGL
BIP_69	VKRVKQILWRLG
BIP_70	KRVQAKAWRLQR
BIP_71	RARQIRWLRKRV
BIP_72	KIQRRAWKQWRK
BIP_73	QQLRWKRVAKAI
BIP_74	KKAIKVVAIGRI
BIP_75	GRVLKIVWRKGR
BIP_76	VVGLRVRWVRLW
BIP_77	WAVRALKVKWAL
BIP_78	LKILIAQAKKGL
BIP_79	VWLAQKIGKWIW
BIP_80	AVAKWALKLWKQ
BIP_81	RGRLKQKWWRRL
BIP_82	VKGAIKRGIWVK
BIP_83	VIRAKAVWGWVK
BIP_84	KIWGLLKLGIAL
BIP_85	LAGLIVKWAGVR
BIP_86	AVKWLGWILAKK
BIP_87	VARAVQKRWRKK
BIP_88	IVKWIAQWKLVG
BIP_89	VKAKRWKWAQLA
BIP_90	LLIAGKWWKLAI
BIP_91	QKIGRAVIWKVK
BIP_92	RAIIKQRWQRRW
BIP_93	WVGVIIKWGLKL
BIP_94	KKIRQWGKAAAW
BIP_95	RLIQWGWKIWAV
BIP_96	QLRVAWKRAWWA
BIP_97	RARIGIWKKWWA
BIP_98	IQIQLVKRWAVI
BIP_99	KAVKKGRRAIVV
BIP_100	VLLRVGARIVVG
BIP_101	GAKIIRKVAQVA
BIP_102	RLAKRKGQAIWV
BIP_103	IKAAKAGQWRRV
BIP_104	ALLAGRKRAVAV
BIP_105	KAVAGARQRWAL

BIP_106	AIGAARAWRQWA
BIP_107	QLARLARVVWGL
BIP_108	AVIVRAAKGGAR
non_BIP1	YNPCLGFI
non_BIP2	YSTCSYYF
non_BIP3	DIIIIVGG
non_BIP4	ETIIIGGG
non_BIP5	LPYFAGCL
non_BIP6	SPNIFGQWM
non_BIP7	AVNACSSLF
non_BIP8	DPITRQWGD
non_BIP9	YTNGNWVPS
non_BIP10	GAKPCGGFF
non_BIP11	GGKVCSAYF
non_BIP15	GYRTCNTYF
non_BIP16	GYSTCSYYF
non_BIP17	KTKTCTVLY
non_BIP18	KYNPCANYL
non_BIP19	KYNPCASYL
non_BIP20	KYNPCLGFL
non_BIP21	KYNPCSNYL
non_BIP22	KYYPCFGYF
non_BIP23	NGKCVLVTL
non_BIP24	RIPTSTGFF
non_BIP25	SVKPCTGFA
non_BIP26	LVMCCVGIW
non_BIP27	NSPNIFGQWM
non_BIP28	ADPITRQWGD
non_BIP29	KAKTCTVLY
non_BIP30	VGARPCGGFF
non_BIP31	QNCPNIFGQWM
non_BIP32	QNHPNIFGQWM
non_BIP33	QNSPNIFGQWM
non_BIP34	QASPNIFGQWM
non_BIP35	ANSPNIFGQWM
non_BIP36	AASPNIFGQWM
non_BIP37	QNDPNIFGQWM
non_BIP38	QNSPNIFGQFM
non_BIP39	IAILPYFAGCL
non_BIP40	FHWWQTSPAHFS
non_BIP41	WPFAHWPWQYPR

non_BIP42	AFLPGGGGVALEAI
non_BIP43	DLRNIFLKIKFKKK
non_BIP44	SNLVECVFSLFKKCN
non_BIP45	EMRKPDGALFNLFRRR
non_BIP46	DKRLPYFFKHLFSNRTK
non_BIP47	EMRLPKILRDFIFPRKK
non_BIP48	EMRLSKFFRDFILQRKK
non_BIP49	ESRISDILLDFLFQRK
non_BIP50	STFFRLFNRSFTQALGK
non_BIP51	GLWEDLLYNINRYAHYIT
non_BIP52	SGSLSTFFRLFNRSFTQA
non_BIP53	SGTLSTFFRLFNRSFTQA
non_BIP54	DIRHRINNSIWRDIFLKRK
non_BIP55	SLSTFFRLFNRSFTQALGK
non_BIP56	SGSLSTFFRLFNRSFTQAGK
non_BIP57	CLGVGSCNDFAGCGYAIVCFW
non_BIP58	SGSLSTFFRLFNASFTQALGK
non_BIP59	MKKVNKALLFTLIMDILIIVGG
non_BIP61	SQKGVYASQRSFVPSWFRKIFRN
non_BIP62	TNRNYGKPNKDIGTCIWSGFRHC
non_BIP63	MKKISKFLPILILAMDIIIIVGG
non_BIP64	SINSQIGKATSSISKCVFSFFKKC
non_BIP65	SKNSQIGKSTSSISKCVFSFFKKC
non_BIP66	AGTKPQGKPASNLVECVFSLFKKCN
non_BIP67	SINSQIGKATSNLVECVFSLFKKCN
non_BIP68	WKAELAPGAVGALQAFLQLANAKIK
non_BIP73	NGWNN
non_BIP74	FPPFG
non_BIP75	SIFTLVA
non_BIP76	YKPWTNF
non_BIP77	YNPCANY
non_BIP78	DSACVYGF
non_BIP79	DSACVVGI
non_BIP80	TNGNWVPS
non_BIP81	EIIIIVGG
non_BIP82	GANPCALYY
non_BIP83	GVNASSSLF
non_BIP84	ESRVSRIILDFLFQRKK
non_BIP85	MAGNSSNFIHKIKQIFTHR
non_BIP86	NKSVIKGNPASNLAQCVFSFFKKC
non_BIP87	GWWEELLHETILSKFKITKALELPIQL

non_BIP89	VNYGNGVSCSKTKCSVNWGQAFQERYTAGINSFVSGVASGA GSIGRRP

Table S8: Performance of presented Model on separate feature vectors.

Feature sets	Sensitivity	Specificity	Accuracy	f1-score	AUC
AAC	**0.88**	**0.88**	**0.88**	**0.88**	**0.95**
DPC	**0.81**	**0.78**	**0.78**	**0.74**	**0.93**
CTD	**0.82**	**0.79**	**0.79**	**0.76**	**0.96**
NMR	**0.84**	**0.84**	**0.84**	**0.84**	**0.88**
PCP	**0.88**	**0.88**	**0.88**	**0.88**	**0.93**

Table S9: Performance of BIPEP with training sets of BioFIN and dPABBs

Evaluation Parameters	dPABBs datasets		BioFIN dataset	
	dPABBs	**BIPEP**	**BioFIN**	**BIPEP**
Accuracy	91.67%	94%	92.61%	93.39%
Sensitivity	88.75%	96%	90.85%	95.55%
Specificity	94.32%	94.72%	94.37%	95%
MCC	83%	87.74%	85%	87.23%

References

(1) Dubchak, I.; Muchnik, I.; Mayor, C.; Dralyuk, I.; Kim, S. Recognition of a Protein Fold in the Context of the SCOP Classification. *Proteins Struct. Funct. Bioinforma.* **1999**, *407* (February), 401–407.

(2) Shockravi, A.; Kavousi, K.; Rezania, J.; Jafari, R.; Norouzi Beirami, M. H.; Ariaeenejad, S.; Moosavi-Movahedi, Z.; Maghami, P.; Mortazavian, A. M.; Moosavi-Movahedi, A. A. Time – Frequency Approach in the Cluster Assignment of Amino Acids Based on Their NMR Profiles. *J. Iran. Chem. Soc.* **2017**.

(3) Yilmaz, A. S.; Alkan, A.; Asyali, M. H. Applications of Parametric Spectral Estimation Methods on Detection of Power System Harmonics. *Electr. Power Syst. Res.* **2008**.

(4) Tang, W.-C.; Lu, S.-W.; Tsai, C.-M.; Kao, C.-Y.; Lee, H.-H. Harmonic Parameters with HHT and Wavelet Transform for Automatic Sleep Stages Scoring. *Proc. World Acad. Sci. Enginnering Technol.* **2007**.

(5) Dunn, J. C. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *J. Cybern.* **1973**, *3* (3), 32–57.

(6) Bezdek, J. C. *Pattern Recognition with Fuzzy Objective Function Algorithms*; Springer US: Boston, MA, 1981.

(7) Huang, T.; Shi, X.-H.; Wang, P.; He, Z.; Feng, K.-Y.; Hu, L.; Kong, X.; Li, Y.-X.; Cai, Y.-D.; Chou, K.-C. Analysis and Prediction of the Metabolic Stability of Proteins Based on Their Sequential Features, Subcellular Locations and Interaction Networks. *PLoS One* **2010**, *5* (6), e10972.
(8) Bhadra, P.; Yan, J.; Li, J.; Fong, S.; Siu, S. W. I. AmPEP: Sequence-Based Prediction of Antimicrobial Peptides Using Distribution Patterns of Amino Acid Properties and Random Forest. *Sci. Rep.* **2018**, *8* (1), 1697.