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

Fereshteh Fallah Atanaki\textsuperscript{a}, Saman Behrouzi\textsuperscript{a}, Shohreh Ariaeenejad\textsuperscript{b}, Amin Boroomand\textsuperscript{c} and Kaveh Kavousi\textsuperscript{*a}

\textsuperscript{a} Laboratory of Complex Biological Systems and Bioinformatics (CBB), Department of Bioinformatics, Institute of Biochemistry and Biophysics (IBB), University of Tehran, Tehran, Iran.

\textsuperscript{b} Department of Systems and Synthetic Biology, Agricultural Biotechnology Research Institute of Iran (ABRII), Karaj, Iran

\textsuperscript{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:

\[ AAC(i) = \frac{\text{frequency of Amino Acid } i}{\text{length of peptide}} \]  

(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:

\[ DPC(i) = \frac{\text{total number of dipeptide } i}{\text{total number of all possible dipeptide}} \]  

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

\[ C_c = \frac{n_c}{N} \quad c = 1, 2, 3 \quad (S3) \]

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:

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

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 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

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

Where the numerator is the Absolute Spectral Power for the frequency (from \(f_0 \, \text{Hz}\) to \(f_1 \, \text{Hz}\)) 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 \(BSP_{\text{Alpha}}, BSP_{\text{Delta}}, BSP_{\text{Theta}}\), are the Sub-Band Spectral Power (Table S2), and DSI, TSI and ASI are the Delta-Slow-wave Index, Theta-Slow-wave Index and the Alpha-Slow-wave Index, respectively.

\[
DSI = \frac{BSP_{\text{Delta}}}{BSP_{\text{Theta}} + BSP_{\text{Alpha}}}
\] (S6)

\[
TSI = \frac{BSP_{\text{Theta}}}{BSP_{\text{Delta}} + BSP_{\text{Alpha}}}
\] (S7)

\[
ASI = \frac{BSP_{\text{Alpha}}}{BSP_{\text{Theta}} + BSP_{\text{Delta}}}
\] (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 ($f_\sigma$) and spectral value at center frequency ($S_{f_c}$), allow the analysis of a specific band in spectrums through Eqs (S9,S10, and S11):

$$f_c = \frac{\sum f_H fS_X(f)}{\sum f_H S_X(f)} \quad (S9)$$

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

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

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:

$$Activity = \text{var}(x(t)) \quad (S12)$$

Where $x(t)$ is the signal.

Mobility shows the mean frequency of the power spectrum:

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

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:

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

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:

$$x_{skw} = \frac{\sum_{n=1}^{N} (x(n) - x_m)^3}{(N-1)x_{std}^3}$$ (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:

$$x_{kurt} = \frac{\sum_{n=1}^{N} (x(n) - x_m)^4}{(N-1)x_{std}^4}$$ (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 \(^5,6\).
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. “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. 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 |
|----------|----------------|----------------|------------------|----------------------|
| 2 clusters | 2 | 1 | 10 | \( n(C) + \frac{n \times (n-1)}{2(T)} + 5 \times n(D) \) |
| 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           |
| 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           |
| Property                                                        | Value  |
|----------------------------------------------------------------|--------|
| PPIHotspotPropBoganC2                                          | CTD    |
| PPIPropMaC2                                                    | CTD    |
| PPIPropMaC3                                                    | CTD    |
| PRNAIPropKimC1                                                 | CTD    |
| PRNAIPropKimC3                                                 | CTD    |
| PRNAIPropEllisC1                                               | CTD    |
| PRNAIPropEllisC2                                               | CTD    |
| PRNAIPropPhippsC1                                              | CTD    |
| PRNAIPropPhippsC2                                              | CTD    |
| PLVBSKhasanovC2                                                | CTD    |
| PLVBSKhasanovC3                                                | 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    |
| PRNAIPropPhippsT13                                             | CTD    |
| Property                                      | Value |
|----------------------------------------------|-------|
| 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   |
| Peptides for independent validation | Sequence                                                                 |
|------------------------------------|--------------------------------------------------------------------------|
| BIP1                               | ILSAIWSGIKSLF                                                           |
| BIP2                               | KTKKKLLKKT                                                             |
| BIP3                               | DGVKLCDVPSGTWSGHCGSSSKCSQQCKDRAYFAYGGACHYQFP SVKCFCKRQC                |
| BIP4                               | ALWKEVLKNAAGKALNEINNLV                                                  |
| BIP5                               | NKGCSACAIJAGAACLADGPIPDFEVAIGITGTFGIAS                                  |
| BIP6                               | FFRNLWKGAKAARFRAGHAAWRA                                                 |

**Table S6**: First independent dataset for validation.
### Table S7: Second independent dataset for validation.

| Peptides for independent validation | Sequence                             |
|-------------------------------------|--------------------------------------|
| BIP_1                               | VRLIVAVRIWRR                         |
| BIP_2                               | QRWKKWKVLKLR                         |
| BIP_3                               | KVVWWKVIKVL                          |
| BIP_4                               | KIWLKLRQRQRQK                        |
| BIP_5                               | WRIKKQWIQIIIV                        |
| BIP_6                               | VARWKIIIAKLW                         |
| BIP_7                               | VQWIQIVVWRKR                         |
| BIP_8                               | KVQIIKQLIAKK                         |
| BIP_9                               | ILVRWIRWRIQW                         |
| BIP_10                              | VIKVLIKRWLKL                         |
| BIP_11                              | RRIIKILLWKLR                         |
| BIP_12                              | KKWQLLIKWKLR                         |
| BIP_13                              | IWLRLKVVLKRR                         |
| BIP_14                              | IILKRVQVQKIK                         |
| BIP_15                              | KRRIKKLLKVVLK                        |
| BIP_16                              | QQKVIRLLWKAK                         |
| BIP_17                              | KRLQWVKVKKKR                         |
| BIP_18                              | VLQIKKVVLRLLL                        |
| BIP_19                              | RIWRRAWKARWK                         |
| BIP_20                              | KIVIRIIQVIK                         |
| BIP_21                              | KIKLQIQQLRIK                         |
| BIP_22                              | WWIKIVVIRVRK                        |
| BIP_23                              | VLKIKVWIWVK                         |
| BIP_24 | WKKVQWLKRLLL   |
| BIP_25 | IKIVRRAKIIIW   |
| BIP_26 | VIKWLLKILRAI   |
| BIP_27 | GLIIKIIKRLW    |
| BIP_28 | IQIWIIRVIWRW   |
| BIP_29 | LLKLKQKGIVIA   |
| BIP_30 | IIKWIVVRQIRK   |
| BIP_31 | WLKRIVKVVVLK   |
| BIP_32 | KVIQWIIVRRL    |
| BIP_33 | QWLKVWIVIKV    |
| BIP_34 | VQRIIWLVRKIV   |
| BIP_35 | QQQFWWLIRWLA   |
| BIP_36 | RVLKIKKIVIVV   |
| BIP_37 | KVIKIVLVRVK    |
| BIP_38 | IKWVLKIVQII    |
| BIP_39 | IQRWWKVWLKVI   |
| BIP_40 | VKWKGKVIVVQL   |
| BIP_41 | LKLKAILKIIRV   |
| BIP_42 | LIVIQLLKKWWK   |
| BIP_43 | RVKAIKWRKIVV   |
| BIP_44 | IKIWKALGQVI    |
| BIP_45 | GKLKIKVKLGIA   |
| BIP_46 | KGKIRKIVLIRR   |
| BIP_47 | WIIRWIKIWLKI   |
| BIP_48 | IVKKVKLIWGVK   |
| BIP_49 | IQLKLWVKRKW    |
| BIP_50 | VAKVKKARWRLR   |
| BIP_51 | RQVRVKRWRARW   |
| BIP_52 | KIVQQKLRLVVI   |
| BIP_53 | QIIKVWRAVII    |
| BIP_54 | QVVVKKKAIQVV   |
| BIP_55 | IRILVRKAIIV    |
| BIP_56 | KKKKIIWRRILV   |
| BIP_57 | LWQLWLKLKLG    |
| BIP_58 | LQRVIWQKWRKV   |
| BIP_59 | RRQWRGWVRIWL   |
| BIP_60 | RGARVIRWKLR    |
| BIP_61 | IAWQLLWGWRVR   |
| BIP_62 | KRKQWKLWVRQI   |
| BIP_63 | KLLGGIWKQAIYV  |
| BIP_64 | WQGWAKIWWVRI   |
| BIP_65 | LKKIIVQAVGLI   |
| BIP_66 | IGQVVLVKIKIA   |
| BIP_67 | ALAIKVWIKILQ   |
| BIP_68 | VIAKIVLLRAGL   |
| BIP_69 | VKRVKQILWRLG   |
| BIP_70 | KRVQAKAWRRLQR  |
| BIP_71 | RARQIRWLRRKV   |
| BIP_72 | KIQRRRAWQWRK   |
| BIP_73 | QQLRWKRVAKAI   |
| BIP_74 | KKAIKVVAIGRI   |
| BIP_75 | GRVLLIKVRKGR   |
| BIP_76 | VVGLRVRWVRLW   |
| BIP_77 | WAVRALKVKWAL   |
| BIP_78 | LKILIAQAKKGL   |
| BIP_79 | VWLAQKIGKWIW   |
| BIP_80 | AVAKWALKLWKQ   |
| BIP_81 | RGRLQKWWRRRL   |
| BIP_82 | VKGAIKRGIWVK   |
| BIP_83 | VIRAKAVGWGVK   |
| BIP_84 | KIWGLLKLGIQL   |
| BIP_85 | LAGLIVKWAGVR   |
| BIP_86 | AVKWLGWILAKK   |
| BIP_87 | VARAVQKRWRKK   |
| BIP_88 | I VKWIAQWKLVG  |
| 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 | RARIGIWKWWWA   |
| BIP_98 | IQIQLVKRWAVI   |
| BIP_99 | KAVKKGRRAIVV   |
| BIP_100| VLLRVGARIUVG   |
| BIP_101| GAKIIRKVAQVA   |
| BIP_102| RLAKRKGQAIWV   |
| BIP_103| IKAATAGQWWRRV  |
| 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 | FHWWQTSFPHFS |
| non_BIP41 | WPFAHWPWQYPR |
| non_BIP42 | AFLPGGGGVALEAI |
| non_BIP43 | DLRNIFLKIKFKKK |
| non_BIP44 | SNLVECVFSLFKKCN |
| non_BIP45 | EMRKPDGALFNLFRRR |
| non_BIP46 | DKRLPYFFKHLFSNRTK |
| non_BIP47 | EMRLPKILRDIFIPRKK |
| non_BIP48 | EMRLSKFDRFILQQRKK |
| non_BIP49 | ESRISSDILLDFLFFQRK |
| non_BIP50 | STFFRLFNRSFTQALGK |
| non_BIP51 | GLWEDLLYINRYAHYIT |
| non_BIP52 | SGSLSTFFRLFNRSFTQA |
| non_BIP53 | SGTLSTFFRLFNRSFTQA |
| non_BIP54 | DIRHRINNSIWRDIFLKRK |
| non_BIP55 | SLSTFFRLFNRSFTQALGK |
| non_BIP56 | SGSLSTFFRLFNRSFTQAGK |
| non_BIP57 | CLGVGSCNDFAHCAYAIVCFW |
| non_BIP58 | SGSLSTFFRLFNRSFTQAGK |
| non_BIP59 | MKKVKNALLFTLIMDILIIVGG |
| non_BIP61 | SQKGVYASQRSFVPWSFRKIFRN |
| non_BIP62 | TNRNYGKPNKDIMGTCIWSGFRHC |
| non_BIP63 | MKKISKFLPILILAMDIIIIVGG |
| non_BIP64 | SINSQIGKATSSISKCVSFFKKC |
| non_BIP65 | SKNSQIGKSTSSISKCVSFFKKC |
| non_BIP66 | AGTKPQGKPSNLVECVSFFLKKCN |
| non_BIP67 | SINSQIGKAWSNLVECVSFFLKKCN |
| non_BIP68 | WKAELAPGAVGALQAFLQLANAKIK |
| non_BIP73 | NGWNN |
| non_BIP74 | FPPFG |
| non_BIP75 | SIFTLV |
| non_BIP76 | YKWPWTNF |
| non_BIP77 | YNPCANY |
| non_BIP78 | DSACVYYGF |
| non_BIP79 | DSACVVGI |
| non_BIP80 | TNGNWVPS |
| non_BIP81 | EIIIIVGG |
| non_BIP82 | GANPCALYY |
| non_BIP83 | GVNASSSLF |
| non_BIP84 | ESRVSRILLDFLFFQRKK |
| non_BIP85 | MAGNSSFIIHDKIFTH |
| non_BIP86 | NKSIVKGNPASNLACQVFSSFKKC |
| non_BIP87 | GWWEELLHETILSKFKITKALELPQIQL |
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.; Ariaenejad, 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. Engineering 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.