# ProPythia - Descriptors Guide

Ana Marta Sequeira

March 2021

## 1 Protein feature extraction

Feature extraction is the first step to build a computational model for biological sequences. A biological sequence can be represented as a succession of L residues, where L is the length of the sequence [28].

The simplest method to describe a protein is its entire aminoacid sequence. This sequence can be used in sequence similarity search based tools (e.g. BLAST). However, this fails when the query protein does not have significant homology to proteins with known function [10].

Besides that, *Machine learning (ML)* methods and

*(deep learning)* models are based on vector mode rather than sequential and cannot perform directly on the sequence. This way, protein sequences should be converted into fixed length feature vectors that contain information regarding patterns and sequence-order effects on residues. When compared to DNA or RNA sequences, feature extraction methods for proteins raise some difficulties due to the diversity of aminoacids and the various structures and functions of proteins [3, 20]. Taking this into account, it is necessary to have methods that can properly identify protein characteristics from the primary sequences of proteins [17].

Feature extraction methods may include physicochemical, aminoacid composition, pseudo aminoacid composition, autocorrelation based features, composition, transition and distribution, conjoint triad and Quasi sequence order descriptors. In alternative, binary profiles can be generated. These descriptors are further detailed below.

### 1.1 Physicochemical descriptors

Physicochemical features are highly useful to represent and distinguish proteins or peptides of different structural, functional and interaction properties and have been widely used in protein prediction problems [18]. This group of descriptors includes mostly one dimension peptide representations such as length, charge, molecular weight, hydrophobic ratio, *grand average of hydropathy (GRAVY)*, aromacity score, isoelectric point, number of C,H,N,O and S atoms (atomic composition), number of each bond type and others. Boman index, a descriptor proposed by Boman can also be computed [6]. This descriptor computes the potential protein interaction index based in the amino acid sequence of a protein. The index is equal to the sum of the solubility values for all residues in a sequence, and it might give an overall estimate of the potential of a peptide to bind to membranes or other proteins [6]. Descriptors of this kind have been used in several studies to distinguish peptides [5, 21, 16].

### 1.2 Residue composition descriptors

The primary sequences of proteins are composed of 20 amino acids. This sequence can be described in several ways.

### 1.2.1 Aminoacid composition

*Aminoacid Composition (AAC)* represents the fraction of each amino acid type within a protein [3, 21, 10, 17]. The fraction of all 20 natural *aminoacid (aa)* are calculated as:

$$f(r) = \frac{Nr}{N} \qquad r = 1, 2, 3, ..., 20 \tag{1}$$

where Nr is the number of the amino acid type r and N is the length of the sequence.

The AAC is a simple but powerful feature, is computationally tractable and, as described in a paper by Roy et al [27], can be used to predict protein interactions with good performance. In previous machine learning models, it has been demonstrated that anticancer and non-anticancer peptides have significant differences considering the AAC [21, 17, 4].

### 1.2.2 Dipeptide composition

The *Dipeptide Composition (DPC)* describes the total number of dipeptides normalized by all the possible combinations of dipeptides present in the given peptide sequence. It returns a 400-dimensional descriptor [7] and is defined as:

$$f(r, s) = \frac{Nrs}{N - 1} \qquad r, s = 1, 2, 3, ..., 20 \tag{2}$$

where Nrs is the number of dipeptide represented by amino acid type r and s.

Adjoining DPC reflects the correlation between two adjoining amino acids. However, in a 3-Dimensional space, and taking into consideration the secondary structures of proteins, two amino acids with g-gap residues may be adjacent. A value of g of 0 is the adjoining DPC, g of 1 describes the correlation of two residues with one residue interval and so forth [11]. DPC composition, and g-gap DPC (reduce dimensionality) has demonstrated promising results in computational proteomics and has been used in anticancer peptides classification [31, 3, 21, 4].

### 1.2.3 Tripeptide composition

*Tripeptide composition (TPC)* describes the total number of tripeptides normalized by all the possible combinations of tripeptides present in the given peptide sequence [7], it gives a 8000 feature vector and is defined as:

$$f(r, s) = \frac{Nrst}{N - 2} \qquad r, s, t = 1, 2, 3, ..., 20 \tag{3}$$

where Nrst is the number of tripeptide represented by amino acid type r, s and t.

### 1.2.4 Reduced aminoacid composition

*Reduced amino acid composition (RAAC)* represents a solution to overcome the high dimensional feature vector issue. Here, the basic aas are grouped together into a smaller number of representative residues based on physicochemical properties. This way, RAAC allows for the minimization of the complexity vectors and enhances the ability to find structural similarity of the peptides. RAAC has been used for protein family characterization [3, 17].

Using AAC based methods, all the sequence order effects are lost; for example, correlations among the amino acids in proteins are ignored (residues apart in sequences may be neighbouring in protein 3-dimensional structure) [20, 10].

## 1.3 Autocorrelation based descriptors

Autocorrelation descriptors are defined based on the distribution of amino acid properties along the sequence. Autocorrelation descriptors include the Normalized Moreau-Broto autocorrelation descriptor, Moran autocorrelation and Geary autocorrelation Descriptor [7, 12].

They quantitatively measure the autocorrelation information of aminoacid residues, based on eight properties: hydrophobicity scale, average flexibility index, polarizability parameter, free energy of amino acid solution in water, residue accessible surface area, aa residue volume, steric parameters and relative mutability [7, 12]. All the aa indices are centralized and standardized before the calculation [7, 12].

### 1.3.1 Normalized Moreau-Broto autocorrelation descriptors

Moreau Broto autocorrelation descriptors for protein sequence may be defined as:

$$AC(d) = \sum_{i=1}^{N-d} P_i P_{i+d} \qquad d = 1, 2, 3, ..., nlag \tag{4}$$

where d is called the lag of the autocorrelation and Pi and Pi+d are the properties of the amino acids at position $i$ and $i + d$ , respectively. $nlag$ is the maximum value of the lag.

The normalized Moreau-Broto autocorrelation descriptors are defined as:

$$ATS(d) = \frac{AC(d)}{N - d} \qquad d = 1, 2, 3, ..., nlag \tag{5}$$

### 1.3.2 Moran autocorrelation descriptor

Moran autocorrelation descriptors to protein sequence may be defined as:

$$I(d) = \frac{\frac{1}{N-d}\sum_{i=1}^{N-d}(P_i - \overline{P})(P_{i+d} - \overline{P})}{\frac{1}{N}\sum_{i=1}^{N}(P_i - \overline{P})^2} \qquad d = 1, 2, 3, ..., 30 \tag{6}$$

where d is called the lag of the autocorrelation and $P_i$ and $P_{i+d}$ are the properties of the amino acids at position $i$ and $i + d$ , respectively. $\overline{P}$ is the average of the considered property P along the sequence, i.e.,

$$\overline{P} = \frac{\sum_{i=1}^{N} P_i}{N} \tag{7}$$

### 1.3.3 Geary autocorrelation descriptor

Geary autocorrelation descriptors application to protein sequence may be defined as:

$$C(d) = \frac{\frac{1}{2(N-d)}\sum_{i=1}^{N-d}(P_i - P_{i+d})^2}{\frac{1}{N-1}\sum_{i=1}^{N}(P_i - \overline{P})^2} \qquad d = 1, 2, 3, ..., 30 \tag{8}$$

where $d$ is called the lag of the autocorrelation and $P_i$ and $P_{i+d}$ are the properties of the amino acids at position $i$ and $i + d$ , respectively, $\overline{P}$ is the average of the considered property $P$ along the sequence.

For each amino acid index, there will be $30 \times nlag$ autocorrelation descriptors, being possible to calculate 240 features for each autocorrelation type [7, 12].

These features are especially useful for protein remote homology detection and fold recognition because they are able to extract the sequence patterns among proteins sharing low sequence similarities [20]. They are described to have good results in protein predictions problems [24, 14, 19].

## 1.4 Composition, transition and distribution

*Composition, Transisition, Distribution (CTD)* feature is composed of three descriptors, *Composition (C)*, *Transition (T)* and *Distribution (D)*, which are based on 7 physicochemical attributes: hydrophobicity, polarity, polarizability, charge, secondary structures, solvent accessibility and normalized Van der Waals volume. As described in table 1, the amino acids are divided in three classes according to their attribute with each amino acid being encoded by an index (1, 2 or 3) according to which class it belongs. After setting the amino acid classes, C is calculated. C is the global percentage for each encoded class in the sequence (number of amino acids of a particular property (such as hydrophobicity) divided by the total number of amino acids in a protein sequence). T represents the percent frequency with which class is followed by another (e.g. 1 followed by 3 or 3 followed by 1) in the encoded sequence. D characterizes the distribution patterns of amino acids of each class in the sequence. It represents the position percentages in the whole sequence for the first residue, 25% residues, 50% residues, 75% residues and 100% residues for a specific encoded class [7].

This feature can be applied in various biological problems, such as the prediction of antimicrobial peptides with high accuracy [5], and is described in literature as having good results in protein predictions problems [24, 14, 19].

Table 1: Amino acid attributes and the division of the amino acids into three groups for each attribute to calculate CTD descriptors. Adapted from *PyDPI* package manual [7].

| | Group 1 | Group 2 | Group 3 |
|---|---|---|---|
| Hydrophobicity | Polar<br>R,K,E,D,Q,N | Neutral<br>G,A,S,T,P,H,Y | Hydrophobic<br>C,L,V,I,M,F,W |
| Normalized van der waals | 0-2.78<br>G,A,S,T,P,D | 2.95-4.0<br>N,V,E,Q,I,L | 4.03-8.08<br>M,H,K,F,R,Y,W |
| Polarity | 4.9-6.2<br>L,I,F,W,C,M,V,Y | 8.0-9.2<br>P,A,T,G,S | 10.4-13.0<br>H,Q,R,K,N,E,D |
| Polarizability | 0-1.08<br>G,A,S,D,T | 0.128-0.186<br>C,P,N,V,E,Q,I,L | 0.219-0.409<br>K,M,H,F,R,Y,W |
| Charge | Positive<br>K,R | Neutral<br>A,N,C,Q,G,H,I,L,<br>M,F,P,S,T,W,Y | Negative<br>D,E |
| Secondary Structure | Helix<br>E,A,L,M,Q,K,R,H | Strand<br>V,I,Y,C,W,F,T | Coil G,N,P,S,D |
| Solvent accessibility | Buried<br>A,L,F,C,G,I,V,W | Exposed<br>R,K,Q,E,N,D | Intermediate<br>M,S,P,T,H,Y |

## 1.5 Conjoint triad descriptors

*Conjoint triad (CTriad)* descriptors were proposed in 2007 by Shen et al. to predict *Protein-Protein Interactions (PPI)* [26]. CTriad consider the properties of the aa and regard any three continuous aa as a unit. The 20 aa were clustered into seven classes according to dipoles and volumes of the side chains as they reflect electrostatic and hydrophobic interactions which are essential for protein-protein interactions and also for the interaction with other molecules, such as lipids [7, 12]. In this way, there are $7 \times 7 \times 7$ (343) different possible triads and the feature vector produced will reflect the frequency of each triad in the protein sequence. The triads can be differentiated according to the classes of aa, i.e., triads composed by three aa belonging to the same classes could be treated identically, as they may play similar roles [26].

Considering V the vector space of the sequence features, each feature $V_i$ represents a sort of triad type, $F$ the frequency vector corresponding to $V$ and $fi$ the frequency of type $V_i$ appearing

in the protein sequence, the detailed description for $(V, F)$ is illustrated in Fig 1 [7, 12].

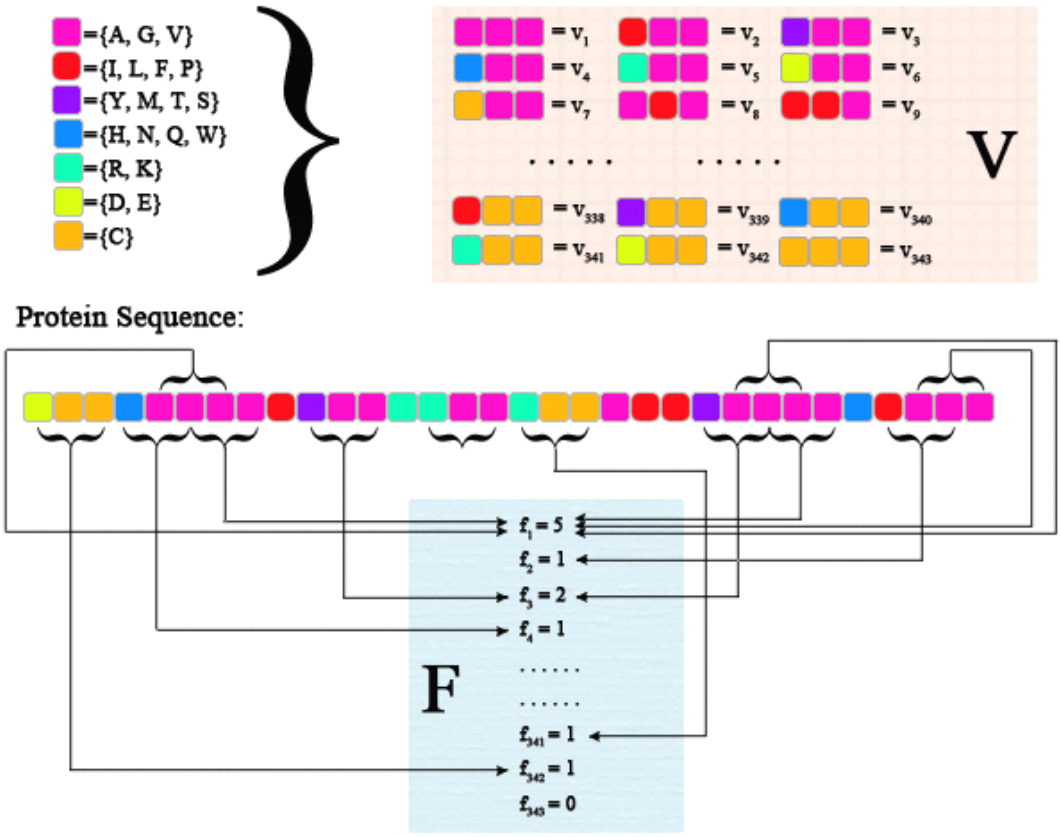


Figure 1: Schematic diagram for constructing the vector space (V,F) of protein sequence for Conjoint Triad descriptors. $V$ is the vector space of the sequence features; each feature $(V_i)$ represents a triad composed of three consecutive amino acids; F is the frequency vector corresponding to $V$, and the value of the ith dimension of $F(f_i)$ is the frequency that vi triad appeared in the protein sequence. From [7, 12].

The length of a protein, number of aa residues, influences the value of $f_i$. In general, a long protein would have a large value of $f_i$, which complicates the comparison between two heterogeneous proteins. Because of this, $f_i$ is normalized with the following equation:

$$d(i) = \frac{f_i - min\{f_1, f_2, f_3, ...f_{343}\}}{max\{f_1, f_2, f_3, ...f_{343}\}} \tag{9}$$

The numerical value of $d_i$ of each protein ranges from 0 to 1, which thereby enables the comparison between proteins.

Conjoint Triad have been used to study protein-protein interaction [26], levels of EC hierarchy and enzyme subfamily [30, 29].

## 1.6 Sequence order descriptors

Sequence order descriptors were proposed by K.C. Chou [8], being derived from both the Schneider-Wrede physicochemical distance matrix and the Grantham chemical distance matrix between each pair of the 20 aminoacids. The physicochemical properties computed include hydrophobicity,

hydrophilicity, polarity and side-chain volume. These features are able to represent aa distribution patterns of a specific physicochemical property along peptide sequence [24].

Here, we can distinguish two type of features, sequence order coupling numbers and quasi-sequence order. For a protein chain of $n$ aa residues $R_1, R_2, R_3 \ldots R_n$, the sequence order effect can be described through a set of sequence order coupling numbers that reflect the coupling mode between all of the most contiguous residues along a protein sequence [7, 12]. A schematic view is presented in Fig 2. The dth-rank sequence-order-coupling number is defined as:

$$\tau_d = \sum_{i=1}^{N-d} (d_{i,i+d})^2 \qquad d = 1, 2, 3, ..., maxlag \tag{10}$$

where $d_{i,i+d}$ is the 'physicochemical' distance between the two amino acids at position $i$ and $i+d$. $maxlag$ is the maximum lag and the length of the protein must be not less than $maxlag$.

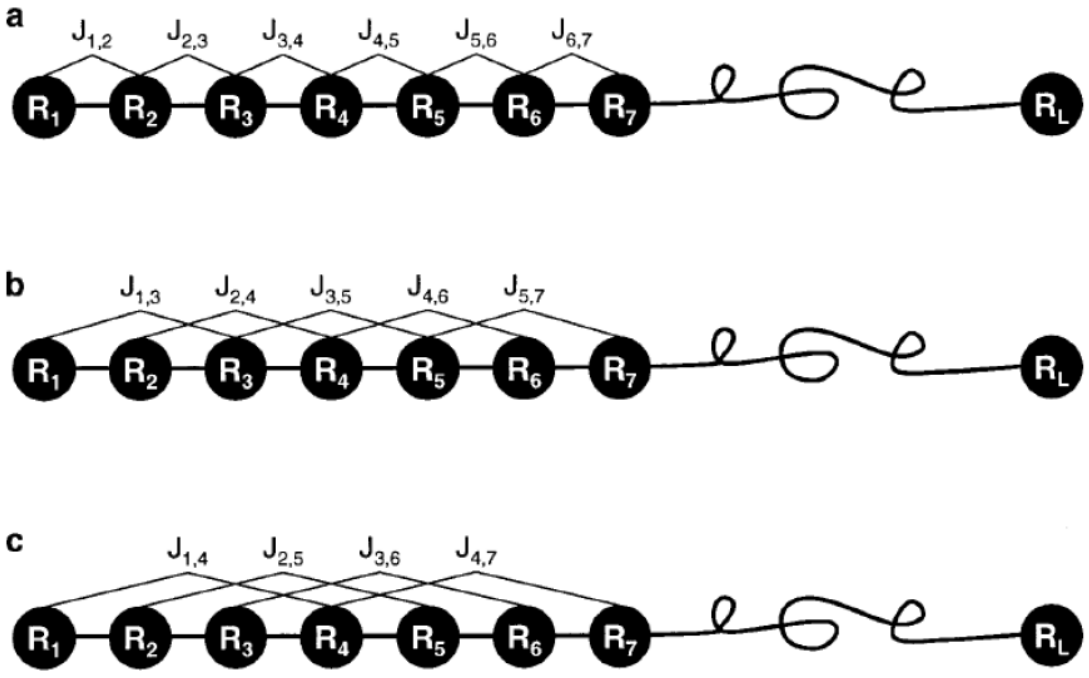


Figure 2: Schematic drawing to Sequence Order Coupling numbers of the (a) the 1st-rank, (b) the 2nd-rank, and (c) the 3rd-rank. (a) Reflects the coupling mode between all the most contiguous residues, (b) that between all the 2nd most contiguous residues, and (c) that between all the 3rd most contiguous residues. From [8].

Quasi-sequence order is derived from the coupling numbers but takes into account the frequency of each aa and the sequence order coupling number [7, 12].

For each aa a *Quasi Sequence order descriptors (QSO)* can be defined as:

$$X_r = \frac{f_r}{\sum_{r=1}^{20} f_r + W \sum_{d=1}^{maxlag} \tau_d)} \qquad r = 1, 2, 3, ..., 20 \tag{11}$$

where $f_r$ is the normalized occurrence for amino acid type $i$ and $w$ is a weighting factor (n=0.1). these are the first 20 quasi-sequence-order descriptors. The other 30 quasi sequence order are defined as:

$$X_r = \frac{w\tau_{d-20}}{\sum_{r=1}^{20} f_r + w \sum_{d=1}^{maxlag} \tau_d)} \qquad r = 21, 22, 23, ..., 20 + maxlag \tag{12}$$

The number of descriptors produced varies accordingly to the *maxlag* chosen.

QSO descriptors have been used to predict protein subcellular location [8], proteins binding affinity [1] or protein functional families [24].

## 1.7 Pseudo aminoacid composition descriptors

Chou's *Pseudo aminoacid composition (PAAC)* of a protein allows to deal with the aminoacid composition considering sequence order correlation. The sequence order correlation is calculated based on properties that play an important role in protein folding, interaction with both the environment and other molecules and function: the values of position, hydrophobicity, hydrophilicity or side chain mass of aa [3, 13, 20, 10]. For example, many helices in proteins are amphiphilic, this is, they are formed by the hydrophobic and hydrophilic aa according to a special order along the helix chain [9].

In PAAC descriptors, whereas the first 20 components reflect the conventional AAC, the remaining PAAC components reflect the correlation patterns, hence incorporating sequence order correlation patterns [10].

These descriptors englobe the PAAC, also called the type 1 pseudo-aminoacid composition and the *Amphiphilic pseudo aminoacid composition (APAAC)*, also called the type 2 pseudo-aminoacid composition.

These concepts proposed by Chou in 2001 and in 2005 have been extensively utilized in various fields of protein structure and function prediction [10, 24] such as homology detection, DNA-binding protein identification [3, 20], prediction of subfamily enzyme classes [9] and prediction of anticancer peptides [13].

## 1.8 Base class peptide descriptors

Base class peptide descriptors englobe the calculus of moment of a sequence and auto and cross correlation of aa values. For this calculus, several scales can be applied, like Eisenberg hydrophobicity consensus aminoacid scale, aminoacid side chain flexibilitiy scale, GRAVY hydrophobicity aminoacid scale, aminoacid side chain flexibilitiy scale, aminoacid polarity scale, amino acid transmembrane propensity scale and several others [22].

## 1.9 NLF and BLOSUM encodings

Other possibility is the use of substitution and scoring matrices such as BLOSUM or NLF matrices. BLOSUM matrices represent accepted mutations between amino acid pairs [15]. NLF matrices [23] encodes each aminoacid to a 18 len vector with physicochemical properties.

## 1.10 Binary profiles

In binary profiles, the peptide segment is presented as binary numerical numbers. In this features it is necessary to take into account that the length of the feature produced depends on the length of the given sequence and therefore, sometimes may be recommended to input sequences with equal length.

Considering aa composition, each aa is represented by a vector with 21 numerical values (20 units for 20 aas and a dummy variable if necessary adding non-natural aa to the sequence). For example A is presented by the vector (1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0) and C by (0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0).

Other way to implement a binary profile is to consider residue properties profile. For example, considering a property-profile for positive charge residues, an aa will be presented by "1" if it is positive charge otherwise "0". It is possible to create property profiles for 25 type of physicochemical properties like hydrophobicity, hydrophilicity or polarity. For example, hydrophobicity profile for amino acid sequence "DPARAAGAHQ" will be (0,1,1,0,1,1,0,1,0,0) as amino acid "P" and "A" are hydrophobic [25].

Binary profiles have been widely used for residue level annotation that includes prediction of protein's secondary structure as well as nucleotide or ligand binding sites in proteins [2].

# References


[1] Wajid Arshad Abbasi et al. "ISLAND: In-Silico Prediction of Proteins Binding Affinity Using Sequence Descriptors". In: (2017), pp. 1–14. eprint: 1711.10540.

[2] Piyush Agrawal et al. "In silico approach for prediction of antifungal peptides". In: *Frontiers in Microbiology* 9.FEB (2018), pp. 1–13.

[3] Shahid Akbar et al. "iACP-GAEnsC: Evolutionary genetic algorithm based ensemble classification of anticancer peptides by utilizing hybrid feature space". In: *Artificial Intelligence in Medicine* 79 (2017), pp. 62–70.

[4] Abdul Basit. "Identification of Anticancer Peptides Using Optimal Feature Space of Chou ' s Split Amino Acid Composition and Support Vector Machine". In: (2017).

[5] Pratiti Bhadra et al. "AmPEP: Sequence-based prediction of antimicrobial peptides using distribution patterns of amino acid properties and random forest". In: *Scientific Reports* 8.1 (2018), pp. 1–10.

[6] Hans G. Boman. "Antibacterial peptides: Basic facts and emerging concepts". In: *Journal of Internal Medicine* 254.3 (2003), pp. 197–215.

[7] Dong Sheng Cao et al. "PyDPI: Freely available python package for chemoinformatics, bioinformatics, and chemogenomics studies". In: *Journal of Chemical Information and Modeling* (2013).

[8] Kuo Chen Chou. "Prediction of protein subcellular locations by incorporating quasi-sequence-order effect". In: *Biochemical and Biophysical Research Communications* 278.2 (2000), pp. 477–483.

[9] Kuo Chen Chou. "Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes". In: *Bioinformatics* 21.1 (2005), pp. 10–19.

[10] Kuo-Chen Chou. "Pseudo Amino Acid Composition and its Applications in Bioinformatics, Proteomics and System Biology". In: *Current Proteomics* 6.4 (2009), pp. 262–274.

[11] Hui Ding et al. "Identification of bacteriophage virion proteins by the ANOVA feature selection and analysis". In: *Molecular BioSystems* 10.8 (2014), pp. 2229–2235.

[12] Jie Dong et al. "PyBioMed: a python library for various molecular representations of chemicals, proteins and DNAs and their interactions". In: *Journal of Cheminformatics* (2018). ISSN: 17582946. DOI: 10.1186/s13321-018-0270-2.

[13] Zohre Hajisharifi et al. "Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test". In: *Journal of Theoretical Biology* 341 (2014), pp. 34–40.

[14] David S. Horne. "Prediction of protein helix content from an autocorrelation analysis of sequence hydrophobicities". In: *Biopolymers* 27.3 (1988), pp. 451–477.

[15] Vanessa Isabell Jurtz et al. "An introduction to deep learning on biological sequence data: Examples and solutions". In: *Bioinformatics* 33.22 (2017), pp. 3685–3690. ISSN: 14602059. DOI: 10.1093/bioinformatics/btx531.

[16] Ernest Y. Lee et al. *Mapping membrane activity in undiscovered peptide sequence space using machine learning*. Vol. 113. 48. 2016, pp. 13588–13593.

[17] Feng Min Li and Xiao Qian Wang. "Identifying anticancer peptides by using improved hybrid compositions". In: *Scientific Reports* 6 (2016), pp. 1–6.

[18] Z. R. Li et al. "PROFEAT: A web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence". In: *Nucleic Acids Research* 34.WEB. SERV. ISS. (2006), pp. 32–37. ISSN: 03051048. DOI: 10.1093/nar/gkl305.

[19] Yunyun Liang. "Prediction of Protein Structural Class Based on Different Autocorrelation Descriptors of Position – Specific Scoring Matrix". In: 73 (2015), pp. 765–784. ISSN: 03406253.

[20] Bin Liu. "BioSeq-Analysis: a platform for DNA, RNA and protein sequence analysis based on machine learning approaches". In: *Briefings in Bioinformatics* January (2017), pp. 1–15.

[21] Balachandran Manavalan et al. "MLACP: machine-learning-based prediction of anticancer peptides". In: *Oncotarget* 8.44 (2017), pp. 77121–77136.

[22] Alex T. Müller et al. "modlAMP: Python for antimicrobial peptides". In: *Bioinformatics (Oxford, England)* 33.17 (2017), pp. 2753–2755.

[23] Loris Nanni and Alessandra Lumini. "A new encoding technique for peptide classification". In: *Expert Systems with Applications* 38.4 (2011), pp. 3185–3191. ISSN: 09574174. DOI: 10.1016/j.eswa.2010.09.005. URL: http://dx.doi.org/10.1016/j.eswa.2010.09.005.

[24] Serene A.K. Ong et al. "Efficacy of different protein descriptors in predicting protein functional families". In: *BMC Bioinformatics* 8 (2007), pp. 1–14.

[25] Akshara Pande et al. "Computing wide range of protein/peptide features from their sequence and structure". In: *bioRxiv* (2019), p. 599126.

[26] "Predicting protein-protein interactions based only on sequences information". In: *Proceedings of the National Academy of Sciences of the United States of America* 104.11 (2007), pp. 4337–4341.

[27] Sushmita Roy et al. "Exploiting amino acid composition for predicting protein-protein interactions". In: *PLoS ONE* 4.11 (2009).

[28] Susana Vinga. "Biological sequence analysis by vector-valued functions: revisiting alignment-free methodologies for DNA and protein classification". In: 2006, pp. 1–36.

[29] Yong Cui Wang et al. "Support vector machine prediction of enzyme function with conjoint triad feature and hierarchical context". In: *BMC Systems Biology* 5.SUPPL. 1 (2011), S6.

[30] Yong-Cui Wang et al. "Prediction of Enzyme Subfamily Class via Pseudo Amino Acid Composition by Incorporating the Conjoint Triad Feature". In: *Protein & Peptide Letters* 17.11 (2012), pp. 1441–1449.

[31] Lei Xu et al. "A novel hybrid sequence-based model for identifying anticancer peptides". In: *Genes* 9.3 (2018).