**Progressive Feature Extraction of Biological Sequences for Machine Learning**

**Author A1, Author B, Shahid Bhat1, Prashant Singh Rana1\***

Thapar Institute of Engineering and Technology, Patiala, Punjab, India.

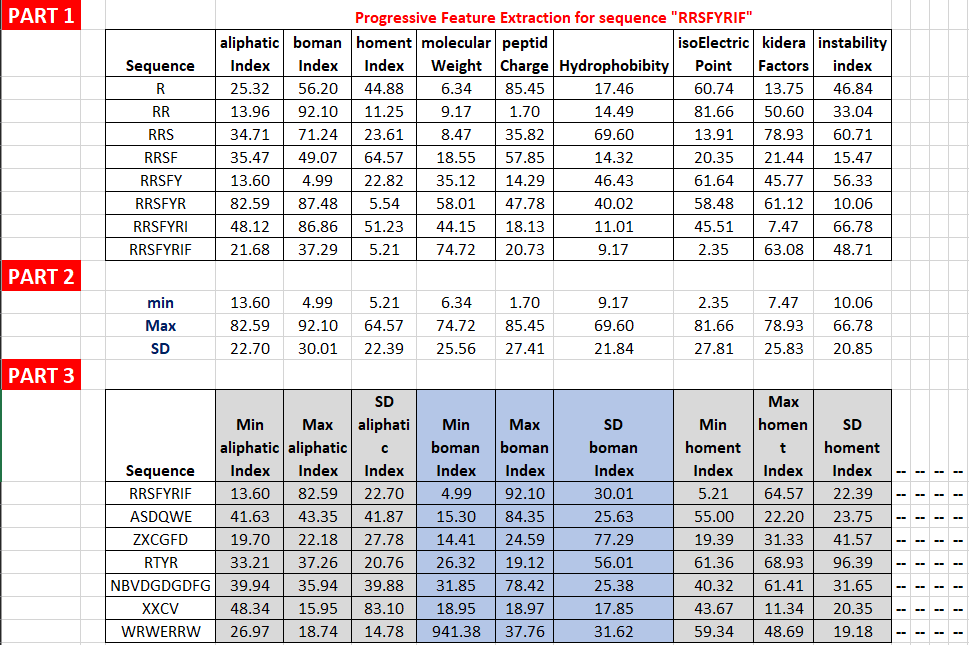
Email: { PSRana@gmail.com}

**Abstract:**

The availability of biological sequences (such as mRNA, circRNA, lncRNA, ncRNAs, epitopes, etc.) has increased massively in recent years. Therefore, new computational methods are needed to extract the significant discriminatory and relevant information to classify these sequences accurately using predictive methods such as Machine learning. Here, we extracted 56 physicochemical properties (such as aliphatic Index, boman Index, homent Index, molecular Weight, peptide Charge, Hydrophobicity, isoelectric point, kidera Factors, instability index, etc.) of a biological sequence in a progressive manner. There are three parts in progressive feature extraction: (i) First, the biological sequence is divided into incremental order (e.g. The sequence “RRSFYRIF” is divided as R, RR, RRS, RRSF, RRSFY, RRSFYR, RRSFYRI, RRSFYRIF) (ii) Second, extract all the 56 physicochemical for every part (iii) Merge all the calculated values using mathematical models (such as Entropy, Fourier and Complex Networks) in vertical order. However, this type of feature extraction technique calculates the same number of features for all the biological sequences even though they differ in length. This type of feature extraction may extract significant discriminatory and relevant information from biological sequences. As a case study, we analyze the 600 peptide sequences of varying length and try to classify as epitopes or non-epitopes. A web service is developed that extract the features of biological sequences in a progressive manner (www.mltool.in/ProgressiveFeatureExtraction).

**Keywords:** Progressive Features, Feature Extraction, Biological Sequences, Entropy, Fourier, Epitopes.

**Methodology:**



**Physicochemical Properties Description**

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature**  **Code** | **R Package and**  **Function used** | **Description** | **Features**  **Count** |
| PF1 | Peptides  aacomp() | * Compute the amino acid composition of a protein sequence * This function calculates the amount of amino acids of a particular class and classified as: Tiny, Small, Aliphatic, Aromatic, Non-polar, Polar, Charged, Basic and Acidic based on their size and R-groups using same function implemented in EMBOSS 'pepstat'. The output is a matrix with the number and percentage of amino acids of a particular class | 9\*2=18 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**Mathematical Functions**

|  |  |  |
| --- | --- | --- |
| **Feature Code** | **Parameters** | **Description** |
| F1 | 1 |  |
| F2 | 2 |  |
| F3 | 5 |  |
| F4 | 10 |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

Positive:

Browse File

Browse File

Negative:

psrana@gmail.com

Email:

Submit

**Test Cases:**

* Case I – All Sequence Length (Variable length)
* Case II – All Sequence Length (Length = 5)
* **Case III – All Sequence Length (Length = 10)**
* Case IV – All Sequence Length (Length = 15)
* Case V – All Sequence Length (min seq Length = 27)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Models | Sen | Spec | Prec | Recall | Accuracy |
| M1 |  |  |  |  |  |
| M2 |  |  |  |  |  |
| M3 |  |  |  |  |  |
| M4 |  |  |  |  |  |
| M5 |  |  |  |  |  |
| M6 |  |  |  |  |  |
| M7 |  |  |  |  |  |

**Table 1: All Sequence Length (Variable length)**

**Table 2:** All Sequence Length (Length = 5)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Models | Sen | Spec | Prec | Recall | Accuracy |
| M1 |  |  |  |  |  |
| M2 |  |  |  |  |  |
| M3 |  |  |  |  |  |
| M4 |  |  |  |  |  |
| M5 |  |  |  |  |  |
| M6 |  |  |  |  |  |
| M7 |  |  |  |  |  |