Progressive Feature Extraction of Biological Sequences for Machine Learning

Author A1, Author B, Shahid Bhat[1], Prashant Singh Rana[1]*
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:

PART 1

Progressive Feature Extraction for sequence "RRSFYRIF"

Sequence	aliphatic Index	boman Index	homent Index	molecular Weight	peptid Charge	Hydrophobibity	isoElectric Point	kidera Factors	instability index
R	25.32	56.20	44.88	6.34	85.45	17.46	60.74	13.75	46.84
RR	13.96	92.10	11.25	9.17	1.70	14.49	81.66	50.60	33.04
RRS	34.71	71.24	23.61	8.47	35.82	69.60	13.91	78.93	60.71
RRSF	35.47	49.07	64.57	18.55	57.85	14.32	20.35	21.44	15.47
RRSFY	13.60	4.99	22.82	35.12	14.29	46.43	61.64	45.77	56.33
RRSFYR	82.59	87.48	5.54	58.01	47.78	40.02	58.48	61.12	10.06
RRSFYRI	48.12	86.86	51.23	44.15	18.13	11.01	45.51	7.47	66.78
RRSFYRIF	21.68	37.29	5.21	74.72	20.73	9.17	2.35	63.08	48.71

PART 2

min	13.60	4.99	5.21	6.34	1.70	9.17	2.35	7.47	10.06
Max	82.59	92.10	64.57	74.72	85.45	69.60	81.66	78.93	66.78
SD	22.70	30.01	22.39	25.56	27.41	21.84	27.81	25.83	20.85

PART 3

Sequence	Min aliphatic Index	Max aliphatic Index	SD aliphatic Index	Min boman Index	Max boman Index	SD boman Index	Min homent Index	Max homent Index	SD homent Index	--	--	--	--
RRSFYRIF	13.60	82.59	22.70	4.99	92.10	30.01	5.21	64.57	22.39	--	--	--	--
ASDQWE	41.63	43.35	41.87	15.30	84.35	25.63	55.00	22.20	23.75	--	--	--	--
ZXCGFD	19.70	22.18	27.78	14.41	24.59	77.29	19.39	31.33	41.57	--	--	--	--
RTYR	33.21	37.26	20.76	26.32	19.12	56.01	61.36	68.93	96.39	--	--	--	--
NBVDGDGDFG	39.94	35.94	39.88	31.85	78.42	25.38	40.32	61.41	31.65	--	--	--	--
XXCV	48.34	15.95	83.10	18.95	18.97	17.85	43.67	11.34	20.35	--	--	--	--
WRWERRW	26.97	18.74	14.78	941.38	37.76	31.62	59.34	48.69	19.18	--	--	--	--

Physicochemical Properties Description

Feature Code	Count	Features	Description
F1	1	F1	
F2	2	F2_1 F2_2	
F3	5	F3_1 F3_2 F3_3 F3_4 F3_5	
F4	10	10	

Mathematical Functions

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