Progressive Feature Extraction of Biological Sequences for Machine Learning

Shahid Ahmad Bhat[1]

School of Mathematics
Thapar Institute of Engineering and Technology (Deemed to be university),
Patiala, Punjab, India.

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.

*Corresponding author
Email address: bhatshahid444@gmail.com (Shahid Ahmad Bhat)

1. Introduction

In recent years, due to advances in DNA sequencing, an increasing num ber of biological sequences have been generated by thousands of sequencing projects [1], creating a huge volume of data [2]. During the last decade, Machine Learning (ML) methods have shown broad applicability in computational biology and bioinformatics [3]. 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

2. Mathematical Models and Materials

In this section we describe the methodological approach used to achieve the proposed objectives are as follows:

2.1. Mathematical Models for Feature Extraction

In this section, 50 feature extraction approaches, 50 numerical mapping techniques with Fourier transform, Entropy measure are shown in the table [**?**]:

Mathematical Models	Function	Mathematical Formulation
M_1 : Statistical characteristics		
	F_1	$S[l] = \frac{\sum_{n=0}^{N-1} T[n]}{N-1}; l = 0, 1, 2, ..., N-1$
	F_2	$S[l] = \frac{1}{N-1}\sum_{n=0}^{N-1} w[n] \times T[n]$, where $l = 0, 1, 2, ..., N-1; w[n] \in [0, 1]$.
	F_3	$S[l] = \prod_{n=0}^{N-1}(T[n])^{N-1}; l = 0, 1, 2, ..., N-1$
	F_4	$S[l] = \prod_{n=0}^{N-1} w[n] \times (T[n])^{N-1}$, where $l = 0, 1, 2, ..., N-1; w[n] \in [0, 1]$.
	F_5	$S[l] = \frac{\sum_{n=0}^{N-1} w[n]}{\sum_{n=0}^{N-1} \frac{w[n]}{T[n]}}$, where $l = 0, 1, 2, ..., N-1; w[n] \in [0, 1]$.
	F_6	$S[l] = k + (\frac{T[1]-T[0]}{2T[1]-T[0]-T[2]}) \times N - 1$ where k =lower limit of sequence, $N-1$ = Size of the sequence, $T[1]$ = Frequency of the sequence, $T[0]$ = Preceding frequency sequence, $T[2]$ = Succeeding frequency sequence.
◇ Median	F_7	$S[l] = \begin{cases} \frac{T[\frac{N}{2}]+T[\frac{N+1}{2}]}{2}, & \text{If N is even,} \\ T[\frac{N+1}{2}] & \text{If N is odd.} \end{cases}$ $l = 0, 1, 2, ..., N-1.$
◇ Mean deviation	F_8	$S[l] = \frac{1}{N-1}\sum_{n=0}^{N-1} \lvert T[n] - F_1 \rvert$, where F_1 =Mean ; $l = 0, 1, 2, ..., N-1$.
◇ Standard deviation	F_9	$S[l] = \sqrt{\frac{1}{N-1}\sum_{n=0}^{N-1} \lvert T[n] - F_1) \rvert}$, where F_1 =Mean ; $l = 0, 1, 2, ..., N-1$.
◇ Variance	F_{10}	$S[l] = \sqrt{\frac{\sum_{n=0}^{N-1}(T[n]-F_1)^2}{N-1}}$, where F_1 =Mean ; $l = 0, 1, 2, ..., N-1$.
◇ Coefficient of variation	F_{11}	$S[l] = \frac{\sqrt{\frac{1}{N-1}\sum_{n=0}^{N-1} \lvert T[n]-F_1) \rvert}}{\frac{\sum_{n=0}^{N-1} T[n]}{N-1}}$; where F_1 =Mean ; $l = 0, 1, 2, ..., N-1$.
M_2 : Entropy Measures		
◇ Shannon Entropy	F_{12}	$S[l] = -\sum_{n=0}^{N-1} T[n] \times \log_2 T[n]$; $l = 0, 1, 2, ..., N-1$.
◇ Tsallis Entropy	F_{13}	$S[l] = \frac{1}{k-1}(1 - \sum_{n=0}^{N-1}(T[n])^k$; where $k \in R$ and $l = 0, 1, 2, ..., N-1$
	F_{14}	$S[l] = -\frac{1}{N-1}\sum_{n=0}^{N-1} T[n] \times \log_2 T[n]$; $l = 0, 1, 2, ..., N-1$.
	F_{15}	$S[l] = \frac{1}{2(N-1)\ln 2}\sum_{n=0}^{N-1}\Big(T[n] \times \ln 2 + T[n] \times \ln 2T[n] + (T[n]+1)\ln\frac{2}{T[n]+1}\Big)$; $l = 0, 1, 2, ..., N-1$.

	F_{16}	$S[l] = \frac{1}{2(N-1)} \sum_{n=0}^{N-1} (T[n]) \exp^{\frac{T[n]}{2}}$; $l = 0, 1, 2, ..., N-1$.
	F_{17}	$S[l] = \max_n T[n] + \sum_{n=0}^{N-1} T[n] \log_2 T[n]$; $l = 0, 1, 2, ..., N-1$.
	F_{18}	$S[l] = \min_n T[n] + \sum_{n=0}^{N-1} T[n] \log_2 T[n]$; $l = 0, 1, 2, ..., N-1$.
	F_{19}	$S[l] = -\log T[n] = -\log(F_{12})$; where F_{12} is Shannon Entropy and $l = 0, 1, 2, ..., N-1$.
M_3 : Entropy estimators		
$\diamond$ Linear entropy estimator	F_{20}	$S[l] = \frac{1}{\sqrt{2\pi(F_{10})^2}} \times \exp^{\frac{-(T[n])^2}{(F_{10})^2}}$; where F_{10} is Variance and $l = 0, 1, 2, ..., N-1$.
	F_{21}	$S[l] = \frac{1}{2} \ln 2\pi \exp(F_{10})^2$; where F_{10} is Variance and $l = 0, 1, 2, ..., N-1$.
$\diamond$ Kernal entropy estimator	F_{22}	$S[l] = \frac{1}{N-1} \sum_{n=0}^{N-1} (\|T[0] - T[n]\|)$, where $T[0]$ is initial vector, $\|\cdot\|$ is a norm and $l = 0, 1, 2, ..., N-1$.
	F_{23}	$S[l] = -\log(F_{22})$; where F_{12} is Kernal entropy estimator and $l = 0, 1, 2, ..., N-1$.
M_4 : Mathematical expectations		
$\diamond$ Expected values	F_{24}	$S[l] = E[T^+] - E[T^-]$, where $E[T^+] = \max_n\{T[n], 0\}$, and $E[T^-] = -\min_n\{T[n], 0\}$ and $l = 0, 1, 2, ..., N-1$.
$\diamond$ Generating function	F_{25}	$S[l] = E[(\exp^{tT[n]})^+] - E[(\exp^{tT[n]})^-]$, where $t \in \mathrm{R}$, $E[(\exp^{tT[n]})^+] = \max_n\{\exp^{tT[n]}, 0\}$, $E[(\exp^{tT[n]})^-] = -\min_n\{\exp^{tT[n]}, 0\}$, and $l = 0, 1, 2, ..., N-1$.
M_5 : Vector Norm		
$\diamond$ p-Norm	F_{26}	$S[l] = \|T\|_p = \sum_{n=0}^{N-1} (\|T[n]\|^p)^p$, where $p = 2$ is the l_2 norm and $l = 0, 1, 2, ..., N-1$.
$\diamond$ ∞-Norm	F_{27}	$S[l] = \|T\|_\infty = \max_n \|T[n]\|$, where $p = \infty$ is the ∞ norm and $l = 0, 1, 2, ..., N-1$.
$\diamond$ ∞-Norm	F_{28}	$S[l] = \|T\|_\infty = \max_n \|T[n]\|$, where $p = \infty$ is the ∞ norm and $l = 0, 1, 2, ..., N-1$.
$\diamond$ Global maxima	F_{29}	$S[l] = \|\max_n T[n]\|$;

⋄ Global minima	F_{30}	$l = 0, 1, 2, ..., N-1.$ $S[l] = \|\min_n T[n]\|;$ $l = 0, 1, 2, ..., N-1.$
M_6 : Fourier Transform ⋄ DFT	F_{31}	$S[l] = \sum_{n=0}^{N-1} T[n] \exp^{-\frac{i2\pi}{N-1}nk}$, where $\exp^{-\frac{i2\pi}{N-1}nk} = \cos(\frac{i2\pi}{N-1}nk) - i\sin(\frac{i2\pi}{N-1}nk)$ and $l = 0, 1, 2, ..., N-1.$
	F_{32}	$S[l] = \frac{k}{N-1} \sum_{n=0}^{N-1} T[n] \exp^{-\frac{i2\pi}{N-1}nk}$, where $\exp^{-\frac{i2\pi}{N-1}nk} = \cos(\frac{i2\pi}{N-1}nk) - i\sin(\frac{i2\pi}{N-1}nk)$ and $l = 0, 1, 2, ..., N-1.$
M_7 : Wavelet Transform ⋄ Cosin DWT	F_{33}	$S[l] = \frac{1}{2}(T[0] + (-1)T[N-1])$ $+ \sum_{n=0}^{N-1} T[n] \cos(\frac{\pi}{N-1}l);$ where $l = 0, 1, 2, ..., N-1.$
	F_{34}	$S[l] = \sum_{n=0}^{N-1} T[n] \cos[\frac{\pi}{N-1}(n+\frac{1}{2})l]$ where $l = 0, 1, 2, ..., N-1.$
	F_{35}	$S[l] = \frac{1}{2}T[0] + \sum_{n=0}^{N-1} T[n] \times$ $\cos[\frac{\pi}{N-1}(l+\frac{1}{2})n];$ where $l = 0, 1, 2, ..., N-1.$
	F_{36}	$S[l] = \sum_{n=0}^{N-1} T[n] \times$ $\cos[\frac{\pi}{N-1}(l+\frac{1}{2})(n+\frac{1}{2})];$ where $l = 0, 1, 2, ..., N-1.$
⋄ Sin DWT	F_{37}	$S[l] = \frac{1}{2}(T[0] + (-1)T[N-1])$ $+ \sum_{n=0}^{N-1} T[n] \sin(\frac{\pi}{N-1}l);$ where $l = 0, 1, 2, ..., N-1.$
	F_{384}	$S[l] = \sum_{n=0}^{N-1} T[n] \sin[\frac{\pi}{N-1}(n+\frac{1}{2})l]$ where $l = 0, 1, 2, ..., N-1.$
	F_{39}	$S[l] = \frac{1}{2}T[0] + \sum_{n=0}^{N-1} T[n] \times$ $\sin[\frac{\pi}{N-1}(l+\frac{1}{2})n];$ where $l = 0, 1, 2, ..., N-1.$
	F_{40}	$S[l] = \sum_{n=0}^{N-1} T[n] \times$ $\sin[\frac{\pi}{N-1}(l+\frac{1}{2})(n+\frac{1}{2})];$ where $l = 0, 1, 2, ..., N-1.$
M_8 : Wavelet coefficients ⋄ DWC	F_{41}	$S[l] = a_{0l} = \sum_{k=-\infty}^{+\infty} [h_{l-2k} a_{1k}$ $+(-1)^l h_{2k+1-n} d_{1k}];$ for every $l \in \mathbb{Z}$, a_1 and d_1 precisely reconstructed.
	F_{42}	$S[l] = \sum_{n=0}^{N-1} (-1)^n n^l T[n]$ where $l = 0, 1, 2, ..., N-1.$

M_9 : Z-transform $\diamond$ Bilateral ZT	F_{43}	$S[l] = \sum_{n=-\infty}^{\infty} T[n]Z^{-}n$ where $Z = A \cdot (\cos\phi + i\sin\phi)$, $A = \|Z\|$, i is imaginary unitl $n \in \mathbb{Z}$.
$\diamond$ Unilateral ZT	F_{44}	$S[l] = \sum_{n=0}^{\infty} T[n]Z^{-}n$ where $Z = A \cdot (\cos\phi + i\sin\phi)$, $A = \|Z\|$, i is imaginary unitl $n \in \mathbb{Z}$.
M_{10} : Positioning $\diamond$ Position of maxima	F_{45}	$S[l] = I[n]\{\max_n(T[n])\}$, where $I[n]$ is index and $n = 0, 2, ..., N-1$
$\diamond$ Position of minima	F_{46}	$S[l] = I[n]\{\min_n(T[n])\}$, where $I[n]$ is index and $n = 0, 2, ..., N-1$

3. Analyzing the drawbacks of Liu and Wang's intuitionistic fuzzy MAGDM method

This section, presents a brief review of Liu and Wang's intuitionistic fuzzy MAGDM method [*] as well as investigate its drawbacks. Let $A = \{A_1, A_2, ..., A_p\}$ be a collection of p alternatives which can be evaluated based on q attributes $B = \{B_1, B_2, ..., B_q\}$ in presence of a group $C = \{C_1, C_2, ..., C_s\}$ of s experts. The weight provided by expert C_t is represented by w_t such that $w_t \geq 0 (t = 1, 2, ..., s)$, $\sum_{t=1}^{s} w_t = 1$ and the weight corresponding to B_k attribute is represented by w_k such that $w_k \geq 0 (k = 1, 2, ..., q)$, $\sum_{k=1}^{q} w_k = 1$. Let $M^t = (m_{ik}^t)_{p\times q}$ where, $m_{ik}^t = \langle(\tau_{ik}^t, \eta_{ik}^t)\rangle$ be an IFV of value of alternative A_i with respect to attribute B_k provided by expert C_t, $1 \leq i \leq p$ and $1 \leq k \leq q$. The general steps of Liu and Wang's intuitionistic fuzzy MAGDM method [*] are as follows:

Step 1: Construct the standardize decision matrix $M^t = (m_{ik}^t)_{p\times q}$, where $m_{ik}^t = \langle(\tau_{ik}^t, \eta_{ik}^t)\rangle$ if B_k is benefit attribute or $m_{ik}^t = \langle(\eta_{ik}^t, \tau_{ik}^t)\rangle$ if B_k is cost attribute, $1 \leq i \leq p$ and $1 \leq k \leq q$.

Step 2: Determine the support $Sup(m_{ik}^t, m_{ik}^e) = 1 - d(m_{ik}^t, m_{ik}^e)$, were $d(m_{ik}^t, m_{ik}^e) = \frac{|\tau_{ik}^t - \tau_{ik}^e| + |\eta_{ik}^t - \eta_{ik}^e| + |\pi_{ik}^t - \pi_{ik}^e|}{2}$, $e \neq t; (t, e = 1, 2, ..., s)$, $1 \leq i \leq p$ and $1 \leq k \leq q$.

Step 3: Determine the support $TT(m_{ik}^t)$ and the weights ϕ_{ik}^t, where $TT(m_{ik}^t) = \sum_{e=1, e\neq t}^{s} (m_{ik}^t, m_{ik}^e)$ and $\phi_{ik}^t = \frac{w_t(1+TT(m_{ik}^t))}{\sum_{t=1}^{s} w_t(1+TT(m_{ik}^t))}$.

Step 4: Using the IFEIPWA operator to transform all decision matrices $M^t = (m_{ik}^t)_{p\times q}$ into a single comprehensive decision matrix $M = (m_{ik})_{p\times q}$.

Step 5: Using the IFEIWA operator to transform the matrix $M = (m_{ik})_{p\times q}$, obtained in Step 3, into a vector $M = (m_{ik})_{p\times 1}$ in order to obtain all attribute values corresponding to each alternative.

Step 6: Determine the values SV $S(M_i)$ and the AV $A(M_i)$ corresponding to each alternative $A = \{A_1, A_2, ..., A_p\}$.
Step 7: Finally, rank all the alternatives based on $S(M_i)$ and $A(M_i)$ obtained in Step 6, three cases arise:
Case I: Higher the value of $S(M_i)$, better the ranking order of corresponding alternative A_i. If the values of $S(M_i)$ are same go to the case II.
Case II: Compare the obtained $A(M_i)$ values and the larger the value of $A(M_i)$, better the ranking order of corresponding alternative A_i. If the values of $A(M_i)$ are same go to the case III.
Case III: If $S(M_i) = A(M_i)$, the the corresponding alternatives $A_i; (i = 1, 2, ..., p)$ are equal.

However, it has been analyzed that there are some drawbacks in Liu and Wang's intuitionistic fuzzy MAGDM method [*], where the ranking order of the alternatives cannot be obtain in some real life situations. To show the same, we consider three decision making problems to highlight the drawbacks of Liu and Wang's intuitionistic fuzzy MAGDM method [*].

4. Front matter

5. Bibliography styles

There are various bibliographyx^2 styles available. You can select the style of your choice in the preamble of this document. These styles are Elsevier styles based on standard styles like Harvard and Vancouver. Please use BibTEX to generate your bibliography and include DOIs whenever available.

Here are two sample references: [1, 2, 3].

References

[1] R. Feynman, F. Vernon Jr., The theory of a general quantum system interacting with a linear dissipative system, Annals of Physics 24 (1963) 118–173. doi:10.1016/0003-4916(63)90068-X.

[2] P. Dirac, The lorentz transformation and absolute time, Physica 19 (1- –12) (1953) 888–896. doi:10.1016/S0031-8914(53)80099-6.

[3] A. Singh, A. Kumar, S. Appadoo, Modified approach for optimization of real life transportation problem in neutrosophic environment, Mathematical Problems in Engineering 2017.