Computational Biology and Bioinformatics
2024, Vol. 12, No. 1, pp. 18-31
https://doi.org/10.11648/j.cbb.20241201.13

Research Article

A Study on Novel Amino Acid Pair Features for Protein Evolutionary Classifications

Xiaogeng Wan[1, *], Xinying Tan[2], Jun Cao[3]

[1]Department of Mathematics, Beijing University of Chemical Technology, Beijing, China
[2]The Fourth Medical Center, PLA General Hospital, Beijing, China
[3]Faculty of Environment and Life, Beijing University of Technology, Beijing, China

Abstract

Protein evolutionary classification from amino acid sequence is one of the hot research topics in computational biology and bioinformatics. The amino acid composition and arrangement in a protein sequence embed the hints to its evolutionary origins. The feature extraction from an amino acid sequence to a numerical vector is still a challenging problem. Traditional feature methods extract protein sequence information either from individual amino acids or kmers aspects, which have general performance with limitations in classification accuracy. To further improve the accuracy in protein evolutionary classifications, six new features defined on separated amino acid pairs are proposed for protein evolutionary classification analysis, where composition and arrangement as well as physical properties are considered for the different combinations of separated amino acid pairs. Different from general consideration of amino acid pairs, the new features account for the features of separated amino acid pairs with spatial intervals in the sequence, which may deeper reflect the spatial relationships and characters between the amino acid in pairs. In test of the performances of the new features, five standard protein evolutionary classification examples are employed, where the new features proposed are compared with classical protein sequence features such as averaged property factors (APF), natural vector (NV) and pseudo amino acid composition (PseAAC) as well as kmer versions of these features. The area under precision-recall curve (AUPRC) analysis shows that the new features are efficient in evolutionary classifications, which outperform traditional protein sequence features that are based on individual amino acids and kmers. Parameter analysis on the novel separated amino acid pair features and kmer features show that the features of some medium or longer length of amino acid pair intervals and kmers may achieve higher classification accuracy in evolutionary classifications. From this analysis, the newly proposed separated amino acid pairs with spacial intervals are proved to be efficient units in extracting protein sequences features, which may interpret richer evolutionary information of protein sequences than individual amino acids and kmers.

Keywords

Protein Sequence, Features, Amino Acid Pair, Evolutionary Classification

[*]Corresponding author: wxgbj88@sina.com (Xiaogeng Wan)

Received: 14 August 2024; **Accepted:** 7 September 2024; **Published:** 23 September 2024

1. Introduction

Protein sequence similarity analysis is a hot topic in bioinformatics research [1-6]. The protein sequence similarity methods are usually categorized into alignment-based and alignment-free approaches. Alignment-based approaches may attain high computational complexity and poor accuracy in dealing with sequences of low identity, whereas alignment-free approaches manage to overcome these drawbacks, which tend to have wider application in protein evolution and functional studies [7, 8]. Alignment-free methods usually map a protein sequence into a vector in real space [7], and treat the distance between these vectors as sequence similarity. These alignment-free approaches greatly improve the speed of sequence comparison and are more effective in handling large data [7, 8], thus gain increasing attention in analyzing biological sequences.

The mapping from a protein sequence to a numerical vector, i.e. feature extraction [7], is a key step and a highly challenging task for alignment-free approaches. Many methods have been developed for protein sequence feature extraction. Typical methods such as the averaged property factors (APF) [9], natural vector (NV) [10], PseAAC [11], moment vector [12, 13], Pse-in-One [14]. These methods extract the amino acid composition, arrangements and physical properties characters of protein sequences, and are proved effective in traditional evolutionary analysis. The natural vector method converts each genetic sequence into a unique point in finite dimensional real space [10]. The pseudo amino acid composition (PseAAC) [11] characterizes the amino acid order in protein sequence using a series of correlation factors, which attains wide applications in protein evolutionary analysis [11]. Despite the composition and arrangement features, physical and chemical properties of amino acids also have wide application and high importance in protein evolution studies [7, 15-17]. Randić summarized in [18], ordering amino acids according to their physical and chemical properties may show better insights in protein similarity analysis than simple amino acid alphabetical orders. The averaged property factors (APF) [9] employs the sequence average of ten important physical properties to successfully classify the different CAT groups of CATH database. The FEGS features, extracted from a series of graphical curve of physicochemical properties, are proved efficient in protein evolutionary classifications [7].

Traditional feature methods extract protein sequence features based on individual amino acids. However, studies on local sequence units such as kmers found that the kmer features are able to construct phylogenetic trees in a much faster way [19]. Therefore, many sequence features are developed based on kmers [19-28]. Zhang et. al. have proposed a novel kmer natural vector to capture the kmer features in genetic sequences [19]. The K-string dictionary feature proposed by Yu et. al. in [20] significantly reduces the representation of proteins by using a lower dimensional frequency vector. Christine et. al. have developed a software tool named Snekmer for recording proteins into kmer vectors and performing protein family classifications on nitrogen cycling families datasets [21]. Ghandi et. al. have proposed a new improved R package to train the gapped-kmer SVM classifiers for protein sequences [22]. Liu et. al. have developed a computational methods based on auto-cross covariance transformation with kmer composition and ensemble learning to identify the DNA-binding proteins [23]. Wen has proposed a kmer sparse matrix, which attains one-to-one correspondence with the biological sequence [24].

Although kmer methods have advantages, but they more or less neglect the inter relationships between amino acids within kmers [29, 30]. Moreover, the feature extraction from a sequence to a numerical vector or matrix is still a challenging problem in computational biology. To consider the inter relationships between amino acids, six novel protein sequence features are proposed accounting the distribution and physical property characters of separated amino acid pairs with spacial intervals in protein sequences. To test the usefulness of the new methods, the new features are applied on five standard protein evolutionary classification datasets, and compare the newly proposed features with traditional features based on individual amino acids and kmers. Analysis on precision-recall curves and AUPRC values proves the efficiency of the newly proposed separated amino acid pair features, which attain the overall highest classification accuracy in standard evolutionary classification problems and outperform traditional features based on individual amino acids and kmers. Parameter analyses on kmer and separated amino acid pair features demonstrate the parameter influences to the classification accuracy, which also indicates optimal parameters for these features.

2. Materials and Methods

Lets first introduce some novel kmer and separated amino acid pair features that cover not only composition and arrangements, but also physical property characters of the kmers and separated amino acid pairs.

2.1. Kmer features

Let $a_1 a_2 \cdots a_L$ denotes a protein sequence, k is a positive integer parameter for the length of the kmers. Splits the whole protein sequence into non-overlapping connected kmer segments, since the length of protein sequence may not necessarily be integer multiple of k, thus the last segment may be shorter than a kmer. The following kmer features are defined on the kmer segments of the protein sequence.

2.1.1. The Kmer Composition Number Feature (KN Feature)

The KN feature is a $20 \cdot K$ dimensional vector accounting the composition number and order of appearance for the different amino acids in kmers, which can be denoted by

$$V_{KN} = (n_{1,A}, n_{1,R}, ..., n_{1,V}, n_{2,A}, n_{2,R}, ..., n_{2,V}, n_{K,A}, n_{K,R}, ..., n_{K,V}),$$

where $n_{i,j}$ stands for the count of the j-type amino acids appeared at the i-th position of the kmers in the protein sequence, j=A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V denote the twenty kinds of amino acids.

2.1.2. The Kmer Mean Distance Feature ($K\mu$ Feature)

The $K\mu$ feature accounts the average distance of the twenty kinds of amino acids at different positions of the kmers, which is a $20 \cdot K$ dimensional vector denoted by

$$V_{K\mu} = (\mu_{1,A}, \mu_{1,R}, ..., \mu_{1,V}, \mu_{2,A}, \mu_{2,R}, ..., \mu_{2,V}, \mu_{K,A}, \mu_{K,R}, ..., \mu_{K,V}),$$

where $\mu_{i,j} = (\sum_{s=1}^{n_{i,j}} d[i, j][s]) / n_{i,j}$ stands for the average distance of the j-type amino acids appeared at the i-th position of the kmer, $n_{i,j}$ stands for the number of j-type amino acid appeared at the i-th position of the kmer as defined before, $d[i, j][s]$ denotes the geometric distance from the s-th j-type amino acids at the i-th positions of the kmer to the initial amino acid (origin) of the sequence, $s = 1, 2, ..., n_{i,j}$.

2.1.3. The Kmer Central Distance Moment Feature (KD Feature)

The KD feature is also a $20 \cdot K$ dimensional vector accounts the second order normalized central distance moments for the positional distribution of the twenty types of amino acids in kmers, it can be expressed by

$$V_{KD} = (D_2^{1,A}, D_2^{1,R}, ..., D_2^{1,V}, D_2^{2,A}, D_2^{2,R}, ..., D_2^{2,V}, D_2^{K,A}, D_2^{K,R}, ..., D_2^{K,V}),$$

where the element $D_2^{i,j} = \sum_{s=1}^{n_{i,j}} \frac{(d[i, j][s] - \mu_{i,j})^2}{n_{i,j} \cdot L}$, the $n_{i,j}$, $\mu_{i,j}$ and $d[i, j][s]$ are defined as before, i=1, 2,..., K denote the i-th position in the kmers, $s = 1, 2, ..., n_{i,j}$.

2.1.4. The Kmer Distance Feature (KF feature)

The KF feature is a $20 \cdot K$ dimensional vector accounts the proposition and order of appearance for the different kinds of amino acids in kmers, it is notated by

$$V_{KF} = (f_{1,A}, f_{1,R}, ..., f_{1,V}, f_{2,A}, f_{2,R}, ..., f_{2,V}, f_{K,A}, f_{K,R}, ..., f_{K,V}),$$

where the element $f_{i,j} = \frac{n_{i,j}}{\sum_{j=1}^{20} n_{i,j}}$ denotes the frequency of the j-type amino acids appeared at the i-th position of the kmers in the given protein sequence.

2.1.5. The Kmer Physical Property Feature (KP Feature)

The KP feature describes the mean physical property values at different positions of the kmers, it can be represented by a $12 \cdot K$ dimensional vector

$$V_{KP} = (p_{1,1}, p_{1,2}, ..., p_{1,12}, p_{2,1}, p_{2,2}, ..., p_{2,12}, p_{K,1}, p_{K,2}, ..., p_{K,12}),$$

where $p_{i,m} = \frac{\sum_{j=1}^{20} p_j^{(m)} \cdot n_{i,j}}{\sum_{j=1}^{20} n_{i,j}}$ is the mean property value for the m-th property factor at the i-th position of the kmers in the sequence, $p_j^{(m)}$ is the m-th physical property value for the j-type amino acids, m=1, 2,..., 12 is the index for the 12 important physical properties as listed in Table S1.

2.2. Separated Amino Acid Pair Features

2.2.1. The Separated Amino Acid Pair Composition Number Feature (SN Feature)

SN feature extracts the composition numbers of λ-spaced amino acid pairs. Let $a_1 a_2 \cdots a_L$ represents a protein sequence, λ is the length parameter for the intervals between the separated amino acid pairs), the SN feature accounts the composition of the separated amino acid pairs a_i and $a_{i+\lambda+1}$ ($i = 1, 2, ..., L - \lambda - 1$) with intermediate λ positional interval. When $\lambda = 0$, a_i and $a_{i+\lambda+1}$ become adjacent, here the separated amino acid pairs are considered with $\lambda = 1, 2, ..., 20$. The SN feature is a 400 dimensional vector represented by:

$$V_{SN} = (n_{A*A}^{\lambda}, n_{A*R}^{\lambda}, ..., n_{V*V}^{\lambda}),$$

where n_{h*k}^{λ} represents the number of the separated amino acid pair $k*h$ with λ positional intervals, here h, k stand for the twenty types of amino acids.

2.2.2. The Separated Amino Acid Pair Mean Distance Feature ($S\mu$ Feature)

Follow the same notations, the $S\mu$ feature extracts the geometric mean distance for the separated amino acid pairs

a_i and $a_{i+\lambda+1}$ with λ-positional interval, $\lambda = 1, 2, ..., 20$. The $S\mu$ feature can be expressed by the 400 dimensional vector:

$$V_{S\mu} = (\mu_{A*A}^{\lambda}, \mu_{A*R}^{\lambda}, ..., \mu_{V*V}^{\lambda}),$$

where $\mu_{k*h}^{\lambda} = \frac{T_{k*h}^{\lambda}}{n_{k*h}^{\lambda}}$ stands for the mean distance from the separated amino acid pair $k*h$ ($*$ represents the λ positional interval, k, h denote the twenty types of amino acids) to the initial amino acid (origin) in the sequence, where $T_{k*h}^{\lambda} = \sum_{i=1}^{n_{k*h}^{\lambda}} s^{\lambda}[k*h][i]$ is the sum of distances between each separated amino acid pair $k*h$ and the origin, $s^{\lambda}[k*h][i]$ denotes the distance between the i-th $k*h$ pair and the origin.

2.2.3. The Separated Amino Acid Pair Central Distance Moment Feature (SD Feature)

The SD feature extracts the second order normalized central moments for the separated amino acid pairs a_i and $a_{i+\lambda+1}$ with λ positional interval in the protein sequence, $\lambda = 1, 2, ..., 20$. The SD feature can be expressed by a 400 dimensional vector:

$$V_{SD} = (D_2^{A*A}, D_2^{A*R}, ..., D_2^{V*V}),$$

where D_2^{h*k} is the second order normalized central moment defined by $D_2^{h*k} = \sum_{i=1}^{n_{h*k}^{\lambda}} \frac{(s^{\lambda}[k \cdot h][i] - \mu_{h*k}^{\lambda})^2}{(n_{h*k}^{\lambda})^{j-1} \cdot n}$, the n_{k*h}^{λ}, μ_{k*h}^{λ} and $s^{\lambda}[k \cdot h][i]$ are defined as above, and k, h denote the 20 types of amino acids.

2.2.4. The Separated Amino Acid Pair Frequency Feature (SF Feature)

The SF feature is a 400 dimensional vector, denotes the proportion of the separated amino acids pairs a_i and $a_{i+\lambda+1}$ with λ positional interval, $\lambda = 1, 2, ..., 20$, it can be expressed by:

$$V_{SF} = (f_{A*A}^{\lambda}, f_{A*R}^{\lambda}, ..., f_{V*V}^{\lambda}),$$

where $f_{k*h}^{\lambda} = \frac{n_{k*h}^{\lambda}}{\sum_{k,h} n_{k*h}^{\lambda}}$ is the frequency for the separated amino acid pair k and h.

2.2.5. The Separated Amino Acid Pair Physical Property Feature I (SPI Feature)

Consider both composition and sequence arrangements, as well as physical property features of the separated amino acid pairs with λ positional interval in the protein sequence, the novel $400+\mu$ dimensional feature vector SPI (integer $\mu \geq 0$) is defined as:

$$V_{SPI} = (x_1, ..., x_{400}, x_{401}, ..., x_{400+\mu}),$$

where

$$x_u = \begin{cases} \frac{f_u}{\sum_{i=1}^{400} f_i + \omega \sum_{j=1}^{\mu} \theta_j}, & 1 \leq u \leq 400 \\ \frac{\omega \theta_{u-400}}{\sum_{i=1}^{400} f_i + \omega \sum_{j=1}^{\mu} \theta_j}, & 401 \leq u \leq 400+\lambda \end{cases},$$

and f_u represents the normalized frequency for the 400 combinations of the separated amino acid pairs, and $\theta_j = \frac{1}{12} \sum_{m=1}^{12} (\frac{p_i^{(m)} + p_{i+\lambda+1}^{(m)}}{2} - \frac{p_{i+j}^{(m)} + p_{i+j+\lambda+1}^{(m)}}{2})^2$ is the j-tier average property factor of the sequence, $p_i^{(m)}$ stands for the m-th property value of amino acid a_i in the sequence. SPI uses the 12 important physical properties (Supplementary Table S1) of amino acids as used in the APF and PseAAC features, μ is a positive integer no larger than the sequence length, ω represents a weight factor in charge of the amino acid arrangement effect, here $\omega = 0.05$ is used as the same as in PseAAC and rPseAAC features (refers to the refined PseAAC with 12 physical properties involved in the correlation factors, the 12 physical properties are the inclusion of the physical properties used in APF and PseAAC features, which are listed in Supplementary Table S1). When $\mu = 0$, SPI is merely the frequency of the 400 kinds of amino acids pairs with λ positional interval. When $\mu > 0$, the initial 400 components x_u ($1 \leq u \leq 400$) reflects the composition effects, while the last μ components indicate the sequence arrangement effects. Here, we use the medium value $\mu = 10$ as PseAAC and rPseAAC do in our analysis.

2.2.6. The Separated Amino Acid Pair Physical Property Feature II (SPII Feature)

The SPII feature also accounts the composition, sequence arrangements and physical property features of the separated amino acid pairs with λ positional interval in the protein sequences. The SPII is a $400+\mu$ dimensional feature vector (integer $\mu \geq 0$) can be expressed by:

$$V_{SPII} = (x_1, ..., x_{400}, x_{401}, ..., x_{400+\mu}),$$

where

$$x_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{400} f_i + \omega \sum_{j=1}^{\mu} \theta_j}, & 1 \le u \le 400 \\ \dfrac{\omega \theta_{u-400}}{\sum_{i=1}^{400} f_i + \omega \sum_{j=1}^{\mu} \theta_j}, & 401 \le u \le 400 + \lambda \end{cases},$$

f_u denotes the normalized frequency for the 400 combinations of separated amino acid pairs, and $\theta_j = \frac{1}{12}\sum_{m=1}^{12}(\sqrt{p_i^{(m)} \cdot p_{i+\lambda+1}^{(m)}} - \sqrt{p_{i+j}^{(m)} \cdot p_{i+j+\lambda+1}^{(m)}})^2$ stands for the j-tier sequence correlation factor. SPII uses the same 12 physical property factors and parameters ($\mu = 10$, $\omega = 0.05$) as SPI and rPseAAC features do.

3. Results

In this study, five classic evolutionary classification datasets, namely the 50 beta globins, the 27 antifreeze proteins, the 40 coronavirus spike proteins, the 25 transferrin sequences (TFs) from 25 vertebrates, the 52 influenza virus protein sequences, are used to validate the effectiveness of the new methods. Evolutionary classification analysis are performed on the five datasets using the novel kmer and separated amino acid pair features defined above, and also compare the efficiency of these features with traditional amino acid feature methods based on individual amino acids such as averaged property factors (APF) [9], natural vector (NV) [10], PseAAC [11], and rPseAAC (refers to the refined PseAAC with 12 physical properties involved in the correlation factors). These features are compared in terms of their precision-recall curves and AUPRC values [31].

The first example is comprised of 50 beta globin sequences belonging to four groups [7], namely, the Aves, Reptilia, Pisces, and Mammals. Accession numbers and taxonomic information of these beta globin proteins can be found in Supplementary Table S2. The AUPRC values for all different features are presented in Table 1. In this table, the separated amino acid pair features show overall the highest AUPRC values among all features, while the kmer features attain the lowest AUPRC values than other features. Among all the features, the rPseAAC (AUPRC=91.92%), SN (Mean AUPRC=91.98%), $S\mu$ (Mean AUPRC=96.04%), SF (Mean AUPRC=92.17%), SPI (Mean AUPRC=91.73%), SPII (Mean AUPRC=92.52%) features perform better than other features. The neighbor-joining tree for these high accuracy features are presented in Figures 1-6. In these figures, the neighbor-joining tree of the separated amino acid pair features clearly separate the mammals and non-mammals, the rPseAAC presents certain mistakes. The separated amino acid pair features such as SN, $S\mu$, SF and SPII also correctly clustered the Aves, Reptilia, Pisces into different branches. The $S\mu$ not only correctly classify the four main groups (Aves, Reptilia, Pisces, and Mammals), but also correctly classify the Chondrichthyes and Actinopterygii in the branch of fishes, as well as the orders and families such as Canidae, Primate, Rodentia, Proboscidea, and Perissodactyla, Artiodactyla, Ruminantia, in the branch of mammals. The rPseAAC, SN, SPII features correctly clustered the species in Anseriformes and Galliformes order. The other features present more or less mistakes in classifying the orders and families in the big branch of mammals.

Table 1. *AUPRC values for different features.*

AUPRC values (%)							
Features		**E1**	**E2**	**E3**	**E4**	**E5**	**Mean**
NV		83.49	69.75	86.56	82.73	91.06	82.72
APF		90.10	70.46	83.67	84.95	89.98	83.83
PseAAC		88.93	67.20	66.62	86.17	95.24	80.83
rPseAAC		91.92	70.70	70.86	85.41	95.68	82.91
Mean		88.61	69.53	76.93	84.81	92.99	82.57
KN	Mean	71.47	69.20	73.14	77.22	95.24	77.25
	Max	73.64	80.37	84.86	82.76	99.01	84.13
$K\mu$	Mean	73.45	70.76	69.86	75.84	95.19	77.02
	Max	76.13	76.05	79.52	86.31	99.66	83.53
KD	Mean	66.62	66.51	69.36	75.38	95.28	74.63

AUPRC values (%)							
Features		**E1**	**E2**	**E3**	**E4**	**E5**	**Mean**
KF	Max	73.10	77.55	74.46	84.89	98.93	81.79
	Mean	71.47	58.80	64.41	77.59	95.12	73.48
KP	Max	73.76	69.82	70.21	83.11	98.95	79.17
	Mean	72.15	57.10	64.61	75.00	95.26	72.82
	Max	73.65	64.91	74.43	80.90	99.71	78.72
Mean		71.03	64.47	68.28	76.21	95.22	75.04
SN	Mean	91.98	78.17	80.61	94.21	98.08	88.61
	Max	95.24	82.95	82.95	96.48	98.77	91.28
$S\mu$	Mean	96.04	82.61	79.15	95.40	89.34	88.51
	Max	97.24	90.23	87.15	97.59	97.14	93.87
SD	Mean	80.04	67.54	76.12	89.18	95.12	81.60
	Max	91.12	81.83	81.56	94.59	97.46	89.31
SF	Mean	92.17	83.04	79.02	94.91	97.46	89.32
	Max	95.28	87.92	87.19	96.97	98.44	93.16
SPI	Mean	91.73	80.18	76.22	87.30	97.29	86.54
	Max	94.67	86.90	85.68	92.07	98.16	91.50
SPII	Mean	92.52	82.73	78.41	93.12	97.51	88.86
	Max	95.49	88.76	87.59	95.43	98.68	93.19
Mean		90.75	79.05	78.25	92.35	95.80	87.24
Total mean		83.45	68.21	75.26	84.65	94.92	80.98

This table shows the AUPRC values for the different features of the five examples. For the kmer and separated amino acid pair features, both mean and maximum AUPRC values are computed over different parameters. The last column shows the average AUPRC values over different types of features.

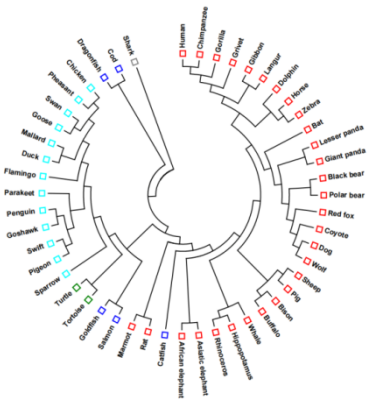

Figure 1. *Evolutionary analysis for the beta globin sequences by rPseAAC features.*

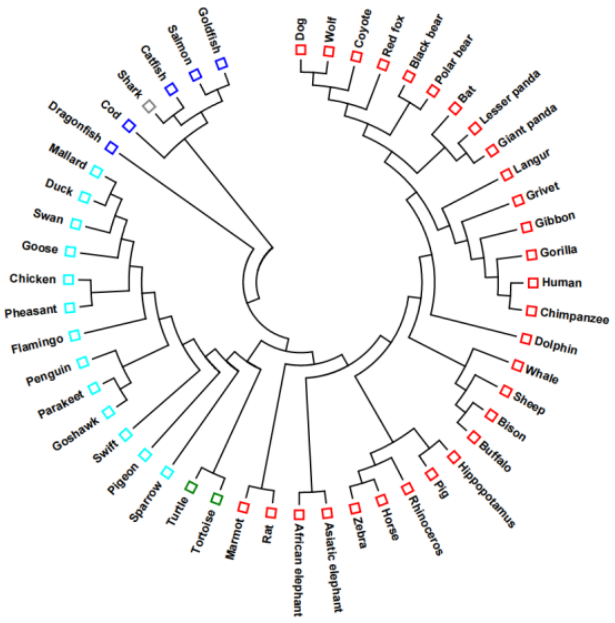

Figure 2. *Evolutionary analysis for the beta globin sequences by SN features.*

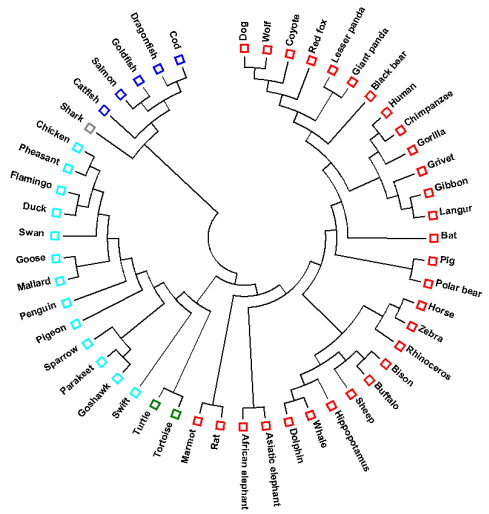

Figure 3. *Evolutionary analysis for the beta globin sequences by Sμ features.*

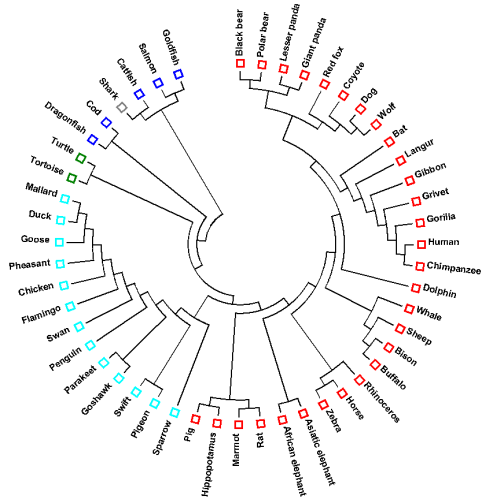

Figure 4. *Evolutionary analysis for the beta globin sequences by SF features.*

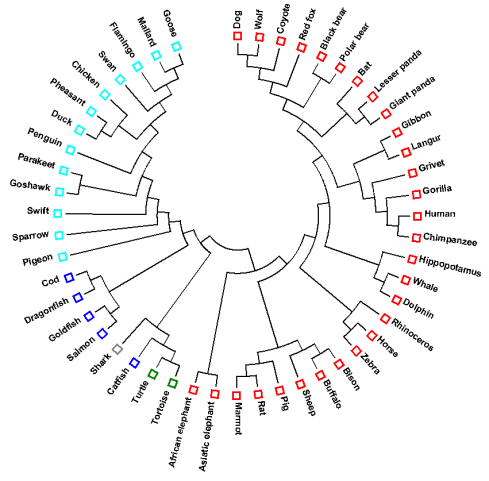

Figure 5. *Evolutionary analysis for the beta globin sequences by SPI features.*

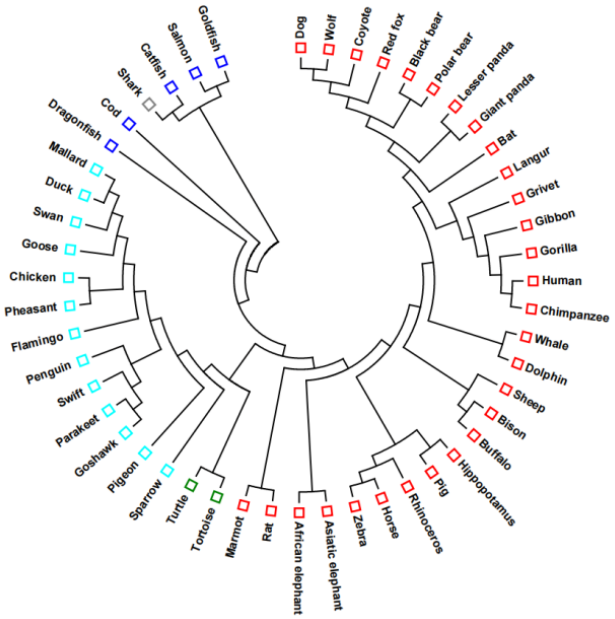

Figure 6. *Evolutionary analysis for the beta globin sequences by SPII features.*

The second example contains 27 antifreeze protein sequences (AFPs) [7]. The accession numbers and taxonomic information of the 27 proteins are listed in Supplementary Table S3. Comparing the AUPRC values as shown in Table 1, the SN (mean AUPRC=78.17%), $S\mu$ (mean AUPRC= 82.61%), SF (mean AUPRC=83.04%), SPI (mean AUPRC=80.18%), SPII (mean AUPRC=82.73%) features perform better than other features. The neighbor-joining tree of the high classification accuracy features are presented in Supplementary Figures S1-S5. In these figures, majority of the taxonomies are well classified with a few exceptions. The AFPs belonging to Choristoneura fumiferana (CF) and Dorcus curvidens binodulosus (DCB) are correctly classified by the $S\mu$, SF and SPI, SPII features. The Dendroides canadensis (DC) are well classified by the SF, SPI and SPII features. The Microdera dzhungarica punctipennis (MDP) are correctly classified by $S\mu$. Majority of these features show closer relations for Choristoneura fumiferana (CF), Dorcus curvidens binodulosus (DCB) and Dendroides canadensis (DC) than Tenebrio molitor (TM), which relations agree with the early evolutionary discoveries [7]. Traditional amino acid features and kmer features present poorer classification accuracy than the separated amino acid pair features.

The third example consists of 40 corona virus spike protein sequences [7]. These corona virus spike protein sequences belong to three groups (alpha, beta and gamma corona virus groups). The accession numbers and taxonomic information of this dataset are presented in Supplementary Table S4. The beta corona virus proteins can further be classified into subgroups, namely, the spike proteins of beta corona virus 1, murine corona virus, SARS-CoV and SARS-CoV-2 (see Supplementary Table S4). As shown in

Table 1, the NV (AUPRC=86.56%), APF (AUPRC=83.67%), SN (mean AUPRC=80.61%), $S\mu$ (mean AUPRC=79.15%), SF (mean AUPRC=79.02%), SPI (mean AUPRC=76.22%), and SPII (mean AUPRC=78.41%) features show better performance than the other features. The PseAAC and majority of the kmer features attain comparatively lower classification accuracy for the 40 corona virus proteins. The neighbor-joining tree for those features with high classification accuracy features are plotted in Supplementary Figures S6-S12. In these figures, the APF, $S\mu$, SF, SPI, SPII features not only correctly cluster the corona virus into three different groups, i.e. the alpha, beta and gamma groups, but also accurately classify the subgroups. The NV and SN features correctly classify majority of the corona virus with an error that the NY-PV08438 in the beta group is error classified to the gamma group. Moreover, the neighbor-joining tree of the APF, $S\mu$, SF, SPI features clearly show that the 2019-nCoVs (i.e. SARS-CoV-2) have closer relations with SARS-CoVs than with the beta corona virus 1 and murine corona viruses, which agree with early discoveries.

The fourth example is made up of 25 transferrin sequences (TFs) from vertebrates [7]. The accession numbers of this dataset are presented in Supplementary Table S5. The 25 transferrin sequences contain three main groups: amphibian, fish and mammal. The group of mammals can be further divided into subgroups of the transferrin (TF) and the lactoferrin (LF) proteins. Examining the AUPRC values as shown in Table 1, the SN (mean AUPRC=94.21%), $S\mu$ (mean AUPRC=95.40%), SD (mean AUPRC=89.18%), SF (mean AUPRC=94.91%), SPI (mean AUPRC=87.30%), and SPII (mean AUPRC=93.12%) features outperform the other features. The Kmer features attain comparatively lower classification accuracy for the 25 transferrin sequences than other features. The neighbor-joining tree for those features with high classification accuracy are plotted in Supplementary Figures S13-S18. All these high accuracy features clearly separate the fish and mammal transferrin sequences into separate branches, and the LFs and TFs of mammals are also correctly separated into different clusters. The SN, $S\mu$, SF, SPI and SPII also correctly classify the Salmo, Salvelinus, Oncorhynchus taxon into different clusters.

The fifth example is consisted of 52 influenza virus proteins belonging to six different influenza A virus subtypes differentiated by their hemagglutinin (H) and neuraminidase (N) types [10, 17]. The accession numbers and taxonomic information of this dataset are presented in Supplementary Table S6. As shown in Table 1, the SN (mean AUPRC=98.08%), SF (AUPRC=97.46%), SPI (mean AUPRC=97.29%), SPII (mean AUPRC=97.51%) features show better performance than other features, the kmer features also work well in this example. The neighbor-joining tree for these high classification accuracy features are presented in Supplementary Figures S19-S22. In these figures, all six influenza A virus subtypes are correctly classified, with an exception for the SPI features where the H11N9 with mallard host show closer relations with the H7N9 proteins than the other H11N9 proteins. The six virus subtypes are clustered into two main branches, namely one branch of H7N9 and H11N9, and the other branch of H1N1, H5N1, H3N2 and H7N3. The virus subtypes of the same neuraminidase type tend to show closer relations, e.g. H1N1 and H5N1, H7N9 and H11N9, which are further clustered by means of their geographical locations and hosts.

To present an over view comparison between the different features, the precision-recall curves for all features and parameters are shown in Figures 7 and 8 and Supplementary Figures S23-S26. Figure 7 presents the precision-recall curves for the traditional features defined on individual amino acids, while Figure 8 and Supplementary Figures S23-S26 show the impact of parameter variations for the kmer and separated amino acid pair features. From the precision-recall curves and AUPRC values for all five examples, all features are effective. When compare the different features among their feature extraction units, i.e. individual amino acid, kmers and separated amino acid pairs, although the different categories of features extract protein sequence features follow the same fashion, i.e. they all extract the composition, arrangements and same physical properties of amino acids, however, they attain different efficiency and accuracy in the classifications. The newly proposed separated amino acid pair features show the overall best performances with the highest average classification accuracy (mean AUPRC=82.57%) than the traditional features (mean AUPRC=82.57%) and kmer features (mean AUPRC=75.04%). The kmer features perform the worst among all types of features. This implies that the composition and arrangement of separated amino acid pairs with spacial intervals may better interpret the evolutionary relationship between protein sequences than the composition and arrangement features of individual amino acids and kmers. Local sequence segments such as kmers or individual amino acids may not fully reflect the evolutionary relations between protein sequences. The reason for these outcomes is may be because the spatial intervals of separated amino acid pairs may describe wider scope of the amino acid distributions, and hence extract more efficient characters. The composition and arrangement of the separated amino acid pair features with wider scope of amino acid distributions may better reflect the evolutionary relations between protein sequences.

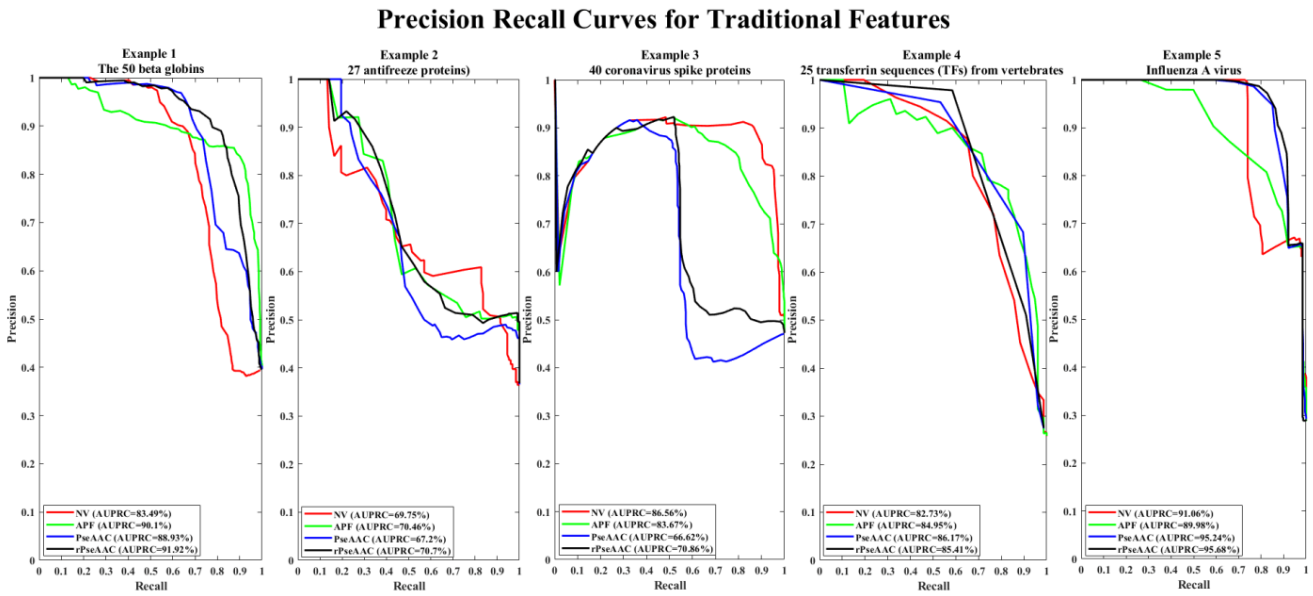

Figure 7. *Precision-recall curves for traditional features.*

This figure shows the precision-recall curves for the traditional features based on individual amino acids for all five examples.

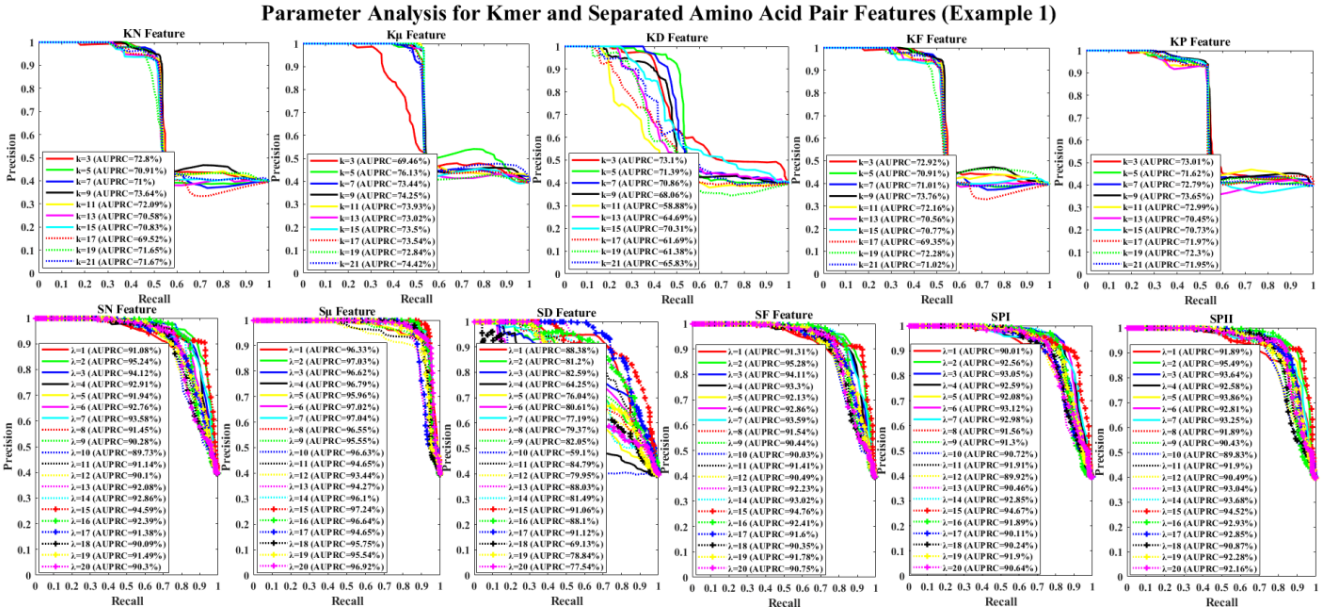

Figure 8. *Parameter analysis for separated amino acid pair features and kmer features (Example 1).*

This figure presents the precision-recall curves for the separated amino acid pair features and kmer features in example 1 with varying parameters. Each precision-recall curve is labeled with the corresponding parameter choice and the AUPRC value.

When comparing the performances within each main category of features, the SF feature (mean AUPRC=89.32%) shows the overall highest classification accuracy among all separated amino acid pair features. Ranking the average AUPRC values in descent order, the most efficient SF feature is followed by the SPII (mean AUPRC=88.86%) and SN (mean AUPRC=88.61%) features, the SD (mean AUPRC= 81.60%) feature attains the lowest classification accuracy among all separated amino acid pair features. The traditional individual amino acid features also perform well in the evolutionary classifications. The kmer features show worse performance than the separated amino acid pair features and the traditional individual amino acid features. However, when

comparing the kmer features, the KN (mean AUPRC=77.25%) and *Kμ* (mean AUPRC=77.02%) features show better performance among all kmer features.

The impact of parameter changes on the kmer and separated amino acid pair features can be observed from Figure 8, Supplementary Figures S23-S26 and Tables 2-3. Tables 2 and 3 separately present the AUPRC values for the kmer and separated amino acid pair features with different parameters. From these figures and tables, the kmer features show the best performance when k=7, which also perform well when k=11 and 19. For the separated amino acid pair features, they attain the best performance when $\lambda = 11$, which also perform well when $\lambda = 15$ and 19. This implies that the kmers and separated amino acid pair intervals with medium or longer length may better interpret the evolutionary relations of protein sequences.

Table 2. *AUPRC values for Kmer features.*

k	AUPRC values (%)					
	KN	**Kμ**	**KD**	**KF**	**KP**	**Mean**
3	75.81	76.95	76.82	74.41	73.80	75.56
5	75.03	75.26	75.52	72.85	73.30	74.39
7	79.69	78.08	79.44	76.64	74.98	77.77
9	74.62	76.59	76.17	72.31	71.25	74.19
11	77.07	79.44	74.10	74.08	74.28	75.79
13	75.74	74.87	73.68	71.30	70.39	73.20
15	79.86	77.47	72.06	74.29	72.40	75.22
17	77.45	76.12	72.03	70.82	71.78	73.64
19	78.21	78.07	73.56	75.07	74.31	75.84
21	79.06	77.34	72.91	73.01	71.77	74.82
Mean	77.25	77.02	74.63	73.48	72.83	75.04
Max	79.86	79.44	79.44	76.64	74.98	77.77

This table shows the average AUPRC values for the kmer features with varying k values for different examples. The last two rows separately show the average and maximum AUPRC values for the different kmer features over all k parameters, while the last column presents the average AUPRC values for each parameter k over all types of kmer features.

Table 3. *AUPRC values for separated amino acid pair features.*

λ	AUPRC values (%)						
	SN	**Sμ**	**SD**	**SF**	**SPI**	**SPII**	**Mean**
1	88.16	87.84	81.13	89.35	84.90	88.28	86.61
2	89.37	87.96	79.80	91.09	85.21	90.02	87.24
3	88.87	90.33	83.19	88.93	86.00	88.13	87.58
4	87.93	87.80	78.95	87.75	86.30	88.21	86.16
5	89.36	87.50	79.49	90.74	88.30	90.73	87.69
6	88.84	87.72	80.04	88.59	87.56	88.26	86.84
7	88.13	89.18	79.90	90.02	87.26	89.63	87.35
8	87.89	88.22	81.82	87.47	86.23	87.81	86.57

λ	AUPRC values (%)						
	SN	**Sμ**	**SD**	**SF**	**SPI**	**SPII**	**Mean**
9	87.78	86.48	85.71	88.37	86.84	88.13	87.22
10	88.42	91.44	80.83	88.00	85.70	87.96	87.06
11	88.84	88.98	85.49	90.66	87.65	91.08	88.78
12	88.75	89.76	82.40	89.51	84.74	88.04	87.20
13	88.08	88.67	84.19	89.98	85.91	89.36	87.70
14	89.20	88.45	81.83	88.32	84.62	87.57	86.67
15	89.92	90.18	81.27	90.90	87.94	89.43	88.27
16	88.30	87.49	83.06	87.49	86.03	87.95	86.72
17	88.21	86.46	80.82	89.90	86.13	87.73	86.54
18	88.04	89.24	79.37	89.75	88.63	88.79	87.30
19	89.68	89.02	80.93	89.59	89.12	90.20	88.09
20	88.41	87.37	81.80	90.00	85.79	89.86	87.20
Mean	88.61	88.50	81.60	89.32	86.54	88.86	87.24
Max	89.92	91.44	85.71	91.09	89.12	91.08	88.78

This table shows the average AUPRC values for the separated amino acid pair features with varying parameters over different examples. The last two rows separately show the average and maximum AUPRC values for the separated amino acid pair features over all λ parameters, while the last column presents the average AUPRC values for each parameter λ over all types of the separated amino acid pair features.

4. Discussion

In this study, six novel protein sequences features are proposed, which account the composition, arrangements, and physical property characters for separated amino acid pairs in protein sequence. To test the effectiveness and accuracy of the new features, five standard evolutionary classification datasets [7, 10, 17, 32] are used in the simulation studies. The taxonomic classification results found by the new features agree with the early discoveries to a large extend [7, 10, 17]. Among all three categories of features, the new features tend to present overall the best performance when compare with the kmer features and classic traditional features such as the averaged property factors (APF) [9], the natural vector (NV) [10], the pseudo amino acid compositions (PseAAC) [11], and also the refined version of the PseAAC (rPseAAC). The Precision-Recall Curves are drawn for each simulation example, where the AUPRC values [31] are computed to compare the accuracy of the different features.

Among the three category of features, i.e. the traditional features, kmer features, and the separated amino acid pair features, the newly proposed separated amino acid pair features show the best performance in all categories of features. The separated amino acid pair features attain the highest classification accuracy (mean AUPRC=87.24%), while the traditional features defined on individual amino acids also perform well but with comparatively lower classification accuracy (mean AUPRC=82.57%), whereas the kmer features show the overall lowest classification accuracy (mean AUPRC=75.04%).

For the traditional features, the APF, NV, and PseAAC features, cover nearly all aspects of an amino acid sequences (amino acid composition, sequence arrangements, as physical property characters). The APF features particularly focus on the physical properties of amino acids [9], while the NV features account for both amino acid composition and sequence arrangements [10], the PseAAC feature is essentially a compositional feature with weighted considerations of the $\lambda-$tier correlations between physical property sequences [11]. All these features are defined for individual amino acids of the twenty kinds [9-11]. The NV, APF and the refined PseAAC feature perform better than the original PseAAC feature. This may be because of the fact that the former three features consider more complex situations of the protein sequences.

The kmer features (similar to the kmer natural vector feature) [19] show consistent style with the separated amino acid pair features in their definitions. The kmer features cover the general composition, arrangements and physical prop-

erties characters of kmers [19-28], where the KN and Kμ features perform better than the other kmer features, which may indicate the composition number and mean distance of the kmers bear richer evolutionary information of protein sequences than other kmer characters.

The separated amino acid pair features, account for the composition, arrangements and physical property characters of the separated amino acid pairs, highly outperform the traditional and kmer features, this may imply that the composition and arrangement of separated amino acid pairs with spacial intervals may capture the amino acid distribution characters in a wider scope, which can catch more effective evolutionary information of a given protein sequence. By focusing on the general composition, arrangement and physical properties of the different sequential units (individual amino acid, kmers and separated amino acid pairs), the separated amino acid pairs with spatial intervals are found to be perfect sequential units that characterize richer evolutionary information of protein sequence than basic units such as individual amino acids and kmers. When comparing among the six newly proposed separated amino acid pair features, the composition features such as SF and SN, as well as the SPII feature, involve both composition and physical properties of the separated amino acid pairs, show better performance than the arrangement features.

When performing parameter analysis on the kmer and separated amino acid pair features, the features with some medium or larger parameters present optimum classification accuracy in the evolutionary classifications. As to the kmers, although the potential number of combinations of the kmers may grow when K increases, however the counts for the true appearance of such kmers may decrease in a real protein sequence [32]. In our analysis, the features of shorter kmers show general performance in the evolutionary classifications, whereas the features of some medium or longer kmers, e.g. the 7-mers (k=7), 11 and 19-mers, attain overall higher classification accuracy than the features of other lengths of kmers. For the separated amino acid pair features, spatial intervals with length $\lambda = 11$, 15 and 19 show the overall best performance than the separated amino acid pair features with other lengths of intervals.

This studies show that the distribution and physical property features of the separated amino acid pairs outperform the traditional individual amino acid features and kmer features. These newly proposed separated amino acid pair features are efficient in protein evolutionary classification studies, which may have wider usages in application studies. Additionally, more studies can be engaged to explore more characters on the separated amino acid pairs to develop more efficient features for protein evolutionary classification analysis.

5. Conclusion

The newly proposed separated amino acid pair features are efficient in protein evolutionary classification studies, which outperform traditional protein sequence features based on individual amino acids and kmers. The distribution and physical property characters of the separated amino acid pairs may attain better interpretation for the evolutionary relationship between protein sequences.

Abbreviations

APF	Averaged Property Factors
NV	Natural Vector
PseAAC	Pseudo Amino Acid Composition
rPseAAC	Refined Pseudo Amino Acid Composition
AUPRC	Area Under Precision-Recall Curves
PR Curves	Precision-Recall Curves
KN	The Kmer Composition Number Feature
Kμ	The Kmer Mean Distance Feature
KD	The Kmer Central Distance Moment Feature
KF	The Kmer Frequency Feature
KP	The Kmer Physical Property Feature
SN	The Separated Amino Acid Pair Composition Number Feature
Sμ	The Separated Amino Acid Pair Mean Distance Feature
SD	The separated Amino Acid Pair Central Distance Moments Feature
SF	The Separated Amino Acid Pair Frequency Feature
SPI	The separated amino Acid Pair Physical Property Feature I
SPII	The separated amino Acid Pair Physical Property Feature II

Supplementary Material

The supplementary material can be accessed at https://doi.org/10.11648/j.cbb.20241201.13

Acknowledgments

We express our gratitude to Beijing University of Chemical Technology for providing library resources to support this study.

Author Contributions

Xiaogeng Wan: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Writing – original draft, Writing – review & editing

Xinying Tan: Data curation, Resources, Validation, Writing – review & editing

Jun Cao: Funding acquisition, Resources, Validation, Writing – review & editing

Data Availability Statement

The accession numbers of the five standard evolutionary datasets are provided in Supplementary Material Tables.

Conflicts of Interest

The authors proclaim that there is no competing interest.

References

[1] Gupta, M. K, Niyogi, R., Misra, M. A. A 2D graphical representation of protein sequence and their similarity analysis with probabilistic method. *Match-commun. Math*. Co. 2014, 72(2), 519–532. https://doi.org/10.5483/BMBRep.2008.41.3.217

[2] He, P. A new graphical representation of similarity/dissimilarity studies of protein sequences. *SAR QSAR in Environ. Res*. 2010, 21(5-6), 571-580. https://doi.org/10.1080/1062936x.2010.510481

[3] Hu, J., Huang, G. Similarity/dissimilarity analysis of protein sequences by a new graphical representation. *Curr. Bioinf*. 2013, 8, 539–544. https://doi.org/10.2174/1574893611308050003

[4] Li, Z., Geng, C., He, P., Yao, Y. A novel method of 3D graphical representation and similarity analysis for proteins. *Match.* 2014, 71(1), 213-226.

[5] Liu, Y., Li, D., Lu, K., Jiao, Y., He P. P-H Curve, a Graphical Representation of Protein Sequences for Similarities Analysis. *Match-commun. Math. Co*. 2013, 70(1), 451–566.

[6] Yao, Y., Dai, Q., Li, C., He, P., Nan X. Analysis of similarity/dissimilarity of protein sequences. *Proteins: Struct., Funct., Bioinf*. 2008, 73(4), 864-871.

[7] Mu, Z., Yu, T., Liu, X., Zheng, H., Wei, L., Liu, J. FEGS: a novel feature extraction model for protein sequences and its applications. *BMC Bioinf*. 2021, 22(1), 297. https://doi.org/10.1186/s128 59-021-04223-3

[8] Zielezinski, A., Vinga, S., Almeida, J., Karlowski, W. M. Alignment-free sequence comparison: benefts, applications, and tools. *Genome Biol*. 2017, 18(1), 186. https://doi.org/10.1186/s13059-017-1319-7

[9] Rackovsky, S. Sequence physical properties encode the global organization of protein structure space. *Proc. Natl. Acad. Sci*. 2009, 106(34), 14345–14348. https://doi.org/10.1073/pnas.0903433106

[10] Yu, C., Deng, M., Cheng, S. Y., Yau, S. C., He, R. L., Yau, S. S.-T. Protein space: A natural method for realizing the nature of protein universe. *J. of Theor. Biol*. 2013, 318, 197–204. https://doi.org/10.1016/j.jtbi.2012.11.005

[11] Shen, H., Chou, K. PseAAC: A flexible web server for generating various kinds of protein pseudo amino acid composition. *Anal. Biochem*. 2008, 373, 386-388. https://doi.org/10.1016/j.ab.2007.10.012

[12] Yau, S. S.-T, Yu, C., He, R. L. A protein map and its application. *DNA Cell Biol*. 2008, 27, 241-250. https://doi.org/10.1089/dna.2007.0676

[13] Yu, C., Cheng, S. Y., He, R. L., Yau, S. S.-T. Protein map: An alignment-free sequence comparison method based on various properties of amino acids. *Gene*. 2011, 486(1–2), 110–118. https://doi.org/10.1016/j.gene.2011.07.002

[14] Liu, B., Liu, F., Wang, X., Chen, J., Fang, L., Chou, K. Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Res.* 2015, 43(W1), W65-W71. https://doi.org/10.1093/ nar/gkv458

[15] He, P., Zhang, Y., Yao, Y., Tang, Y., Nan, X. The graphical representation of protein sequences based on the physicochemical properties and its applications. *J. Comput. Chem*. 2010, 31, 2136–2142.

[16] Wu, Z., Xiao, X., Chou, K. C. 2D-MH: a web-server for generating graphic representation of protein sequences based on the physicochemical properties of their constituent amino acids. *J. Theor. Biol*. 2010, 267, 29–34. https://doi.org/10.1016/j.jtbi. 2010.08.007

[17] Yu, J., Qu, A., Tang, H., Wang, F., Wang C., Wang, H., Wang, J., Zhu H. A novel numerical model for protein sequences analysis based on spherical coordinates and multiple physicochemical properties of amino acids. *Biopolymers*. 2019, 110, e23282. https://doi.org/10.1002/bip.23282

[18] Randić, M. 2-D graphical representation of proteins based on physicochemical properties of amino acids. *Chem. Phys. Lett*. 2008, 440(4-6), 291–295. https://doi.org/10.1016/j.cplett.2007.04.037

[19] Zhang, Y., Wen, J., Yau, S. S.-T. Phylogenetic analysis of protein sequences based on a novel k-mer natural vector method. Genomics. 2019, 111, 1298–1305. https://doi.org/10.1016/j. ygeno.2018.08.010

[20] Yu, C., He, R. L., Yau, S. S.-T. Protein sequence comparison based on K-string dictionary. Gene. 2013, 529(2), 250-256. https://doi.org/10.1016/j.gene.2013.07.092

[21] Chang, C. H., Nelson, W. C., Jerger, A., Wright, A. T., Egbert, R. G., McDermott, J. E. Snekmer: a scalable pipeline for protein sequence fingerprinting based on amino acid recording. *Bioinform Adv*. 2023, 3(1), vbad005. https://doi.org/10.1093/bioadv/vbad005

[22] Ghandi, M., Mohammad-Noori, M., Ghareghani, N., Lee, D., Garraway, L., Beer, M. A. GkmSVM: an R package for gapped-kmer SVM. *Bioinformatics.* 2016, 32(14), 2205-2207. https://doi.org/10.1093/bioinformatics/btw203

[23] Liu, B., Wang, S., Dong, Q., Li, S., Liu, X. Identification of DNA-binding proteins by combining auto-cross covariance transformation and ensemble learning. *IEEE T. on Nanobiosci*. 2016, 15(4), 328-334. https://doi.org/10.1109/TNB.2016.2555951

[24] Wen, J., Zhang, Y., Yau, S. S.-T. K-mer Sparse matrix model for genetic sequence and its applications in sequence comparison. *J. Theor. Biol*. 2014, 363, 145-150. https://doi.org/10.1016/j. jtbi.2014.08.028

[25] Kim, T. K., Bunron, L. Fast Global Alignment Technique Using Kmer-Distance and Parallelism. *BigDAS '15: Proceedings of the 2015 International Conference on Big Data Applications and Services Jeju Island Republic of Korea*. 2015. https://doi.org/10.1145/2837060.2837094

[26] Liu, Y., Wang, X., Liu, B. IDP–CRF: Intrinsically Disordered Protein/Region Identifification Based on Conditional Random Fields. *Int J Mol Sci*. 2018, 19(9), 2483. https://doi.org/10.3390/ijms19092483

[27] Wen, J., Chan, R. H. F., Yau, S. C., He, R. L., Yau, S. S.-T. K-mer natural vector and its application to the phylogenetic analysis of genetic sequences. *Gene*. 2014, 546(1), 25-34. https://doi.org/10.1016/j.gene.2014.05.043

[28] Naznin, F., Sarker, R., Essam, D. Two Hybrid Algorithms for Multiple Sequence Alignment. *AIP Conf. Proc*. 2010, 1210(1), 69-83. https://doi.org/10.1063/1.3314271

[29] Yang, X. W., Wang, T. M. A novel statistical measure for sequence comparison on the basis of k-word counts. *J. Theor. Biol*. 2013, 318, 91–100. https://doi.org/10.1016/j.jtbi.2012.10.035

[30] Yu, H. J. Segmented K-mer and its application on similarity analysis of mitochondrial genome sequences. *Gene*. 2013, 518, 419–424. https://doi.org/10.1016/j.gene.2012.12.079

[31] Tian K., Zhao X., Zhang Y., Yau S. Comparing protein structures and inferring functions with a novel three-dimensional Yau-Hausdorff method. *J. Biomol. Struct. Dyn.* 2019, 37(16), 4151-60. https://doi.org/10.1080/07391102.2018.154 0359

[32] Morikawa N. Discrete differential geometry of n-simplices and protein structure analysis. Applied Mathematics. 2014, 5(16), 2458-2463. https://doi.org/10.4236/am.2014.516237