MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **93** (2025) 579–597
ISSN: 0340–6253
doi: 10.46793/match.93-3.579J

A Novel Feature Extraction Model for Protein Sequence Comparison

Jian Jin, Jie Feng*

School of Science, Minzu University of China, Beijing 100081, China
15339694679@163.com, fengjie0536@163.com

(Received January 19, 2024)

Abstract

In this paper, we introduce a novel feature extraction model for protein sequence comparison. First we cluster 20 natural amino acids into 8 groups based on their physicochemical properties using K-Means algorithm, then a 36-dimensional feature vector is extracted from the frequency, the mean absolute error of the position of amino acids in reduced amino acid sequences, and the order information of 20 amino acids in the original sequences. Finally, the Euclidean distance is used to measure the similarity and evolutionary distance between protein sequences. The test indicates that our method is fast and accurate for classifying and inferring the phylogeny of proteins.

1 Introduction

Biological sequence comparison is an important research directions in computational biology and bioinformatics. Many other research works, such as molecular evolution, protein structure prediction and gene recognition are built upon this work. Using similarity analysis methods to study the similarity and differences in gene or protein sequences between different species can further reveal their structural and functional characteristics.

*Corresponding author: fengjie0536@163.com

Sequence comparison methods are generally divided into alignment-based methods and alignment-free methods. Among them, BLAST and ClustaW are the most widely used alignment-based methods [1–3].The results of these programs provide approximate solutions to protein sequence comparison problems. On the other hand, many alignment-free methods have been proposed for sequence comparison. The alignment-free methods do not compare base pairs. It takes sequence as a whole and converts it into graphical representations [4–16] or numerical vectors [17–25] for analysis and comparison. For example, Mu [15] constructed a CGR curve for protein sequences based on the physicochemical properties of amino acids, and then transformed the curve into a multidimensional feature vector using the distribution of points in the CGR image; Qi [16] proposed a three-dimensional graphical representation of protein sequences based on 10 physicochemical properties of 20 amino acids and a BLOSUM62 matrix, where the values in the BLOSUM62 matrix represent the probability of one amino acid being replaced by other amino acids; He [17] classified 20 amino acids into 8 categories based on their three physicochemical properties, and extracted the number, the average position and the variation of the position of amino acids; Li [18] mapped each amino acid into a vector based on the physicochemical properties of amino acids and the Hungarian algorithm. Then protein sequences were represented as time series in eleven dimensional space and the DTW algorithm was applied to calculate the distance between two time series to measure the similarity of protein sequences.

In this study, we introduce a new alignment-free method for protein sequence comparison. First we cluster 20 natural amino acids into 8 groups based on their physicochemical properties using K-Means algorithm, then a 36-dimensional feature vector is extracted from the frequency, the mean absolute error of the position of amino acids in reduced amino acid sequences, and the order information of 20 amino acids in the original sequences. The similarity between protein sequences is measured by Euclidean distance and the phylogenetic trees are constructed for three data sets. The test indicates that our method is fast and accurate for classifying and inferring the phylogeny of proteins.

2 Feature vectors of protein sequences

2.1 Feature extraction based on amino acid quantity and position

In this paper, we consider 10 major physicochemical properties of amino acids including the value of the dissociation constant ($pK_a(COOH)$ and $pK_a(NH_3^+)$), the isoelectric point(pI), the hydrophilicity(Hyd), the solubility(Sol), the molecular weight(Mw), the polar requirement(Pr), the chemical composition of side chains(Cc), the hydrophobicity(Hb) and the sidechain mass(Scm), the values of these properties are listed in Table 1.

Table 1. 10 physicochemical properties of 20 amino acids.

Amino acid	$pK_a(COOH)$	$pK_a(NH_3^+)$	$pI(25°C)$	Hyd	Sol	Mw	Pr	Cc	Hb	Scm
A	2.34	9.69	6.01	1.8	167.2	89.079	7	0	0.62	15
C	1.96	10.28	5.07	2.5	0	121.145	4.8	2.75	0.29	47
D	1.88	9.6	2.77	-3.5	5	133.089	13	1.38	-0.9	59
E	2.19	9.67	3.22	-3.5	8.5	147.116	12.5	0.92	-0.74	73
F	1.83	9.13	5.48	2.8	27.6	165.177	5	0	1.19	91
G	2.34	9.6	5.97	-0.4	249.9	75.052	7.9	0.74	0.48	1
H	1.82	9.17	7.59	-3.2	0	155.141	8.4	0.58	-0.4	82
I	2.36	9.68	6.02	4.5	34.5	131.16	4.9	0	1.38	57
K	2.18	8.95	9.74	-3.9	739	146.17	10.1	0.33	-1.5	73
L	2.36	9.6	5.98	3.8	21.7	131.16	4.9	0	1.06	57
M	2.28	9.21	5.74	1.9	56.2	149.199	5.3	0	0.64	75
N	2.02	8.8	5.41	-3.5	28.5	132.104	10	1.33	-0.78	58
P	1.99	10.96	6.48	1.6	1620	115.117	6.6	0.39	0.12	42
Q	2.17	9.13	5.65	-3.5	7.2	146.131	8.6	0.89	-0.85	72
R	2.17	9.04	10.76	-4.5	855.6	174.188	9.1	0.65	-2.53	101
S	2.21	9.15	5.68	-0.8	422	105.078	7.5	1.42	-0.18	31
T	2.11	9.62	5.87	-0.7	13.2	119.105	6.6	0.71	-0.05	45
V	2.32	9.62	5.97	4.2	58.1	117.133	5.6	0	1.08	43
W	2.38	9.39	5.89	-0.9	13.6	204.213	5.2	0.13	0.81	130
Y	2.2	9.11	5.66	-1.3	0.4	181.176	5.4	0.2	0.26	107

Based on these physicochemical property data, we then sort 20 natural amino acids into groups using K-Means clustering algorithm. In order to eliminate the effect of the inconsistency of the magnitudes, the data is first subjected to minimum and maximum normalization. The elbow diagram and silhouette coefficient diagram are plotted to determine the optimal number of cluster groups , the results are shown in Figures 1 and 2.

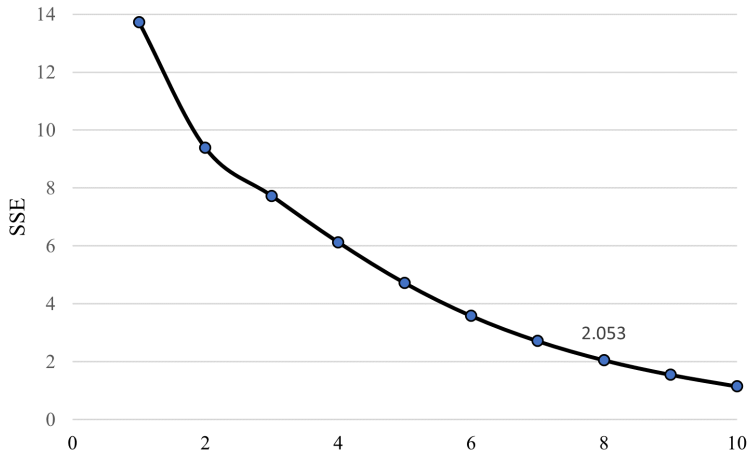

Figure 1. Diagram of the elbow.

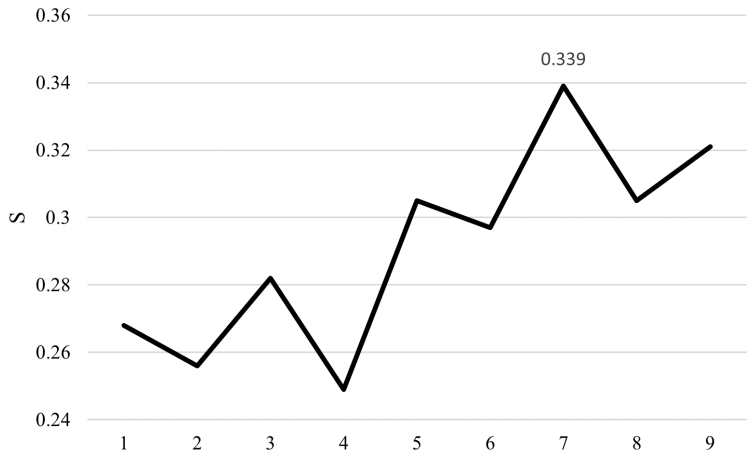

Figure 2. Diagram of the silhouette coefficient.

From Figure 2, we can see that the silhouette coefficient reached its highest value when grouped into 8 categories, and at this time, the SSE value is less than 4 from Figure 1, the clustering effect is good. In summary, we divided 20 amino acids into 8 classes and the results are shown in Table 2. According to Table 2, a 20-letter protein primary sequence can

be converted into a 8-letter reduced amino acid sequence.

Table 2. Amino acid clustering results

Amino acids	Denote
W Y	α
D E H N Q	β
A I L M V	γ
G S T	δ
C	ϵ
P	θ
K R	η
F	ε

2.1.1 Feature extraction based on the number of amino acids

For a reduced amino acid sequence $H = h_1h_2h_3...h_n$, where $h_i \in \Omega, i = 1, 2, ..., n$, $\Omega = \{\alpha, \beta, \gamma, \delta, \epsilon, \theta, \eta, \varepsilon\}$, the frequency of occurrence of each amino acid was calculated using the following formula:

$$R_i = \frac{n_i}{n}. \quad i = \alpha, \beta, \gamma, \delta, \epsilon, \theta, \eta, \varepsilon \tag{1}$$

where n_i denotes the number of amino acids of class i in the sequence, n denotes the length of the sequence. Then we extract a 8-dimensional feature vector $R = (R_\alpha, R_\beta, R_\gamma, R_\delta, R_\epsilon, R_\theta, R_\eta, R_\varepsilon)$ for each protein primary sequence.

2.1.2 Feature extraction based on amino acid location information

For a protein sequence of length n, $H = h_1h_2h_3...h_n$, we then extract the location information of each amino acid. For example, for the amino acid α, we define its mean absolute error of position as follows:

$$M_\alpha = \frac{\sum_{i=1}^{n} f_\alpha(h_i)|i - \mu_\alpha|}{n_\alpha} \tag{2}$$

$$f_\alpha(h_i) = \begin{cases} 0, & h_i \neq \alpha \\ 1, & h_i = \alpha \end{cases} \qquad i = 1, 2, ..., n \tag{3}$$

where n_α denotes the number of occurrences of α, $\mu_\alpha = \frac{\sum_{i=1}^{n} f_\alpha(h_i) \cdot i}{n_\alpha}$ is the average position of α [17]. According to these definitions, the mean absolute error of each amino acid position can be calculated. Finally ,an 8-dimensional feature vector $M = (M_\alpha, M_\beta, M_\gamma, M_\delta, M_\epsilon, M_\theta, M_\eta, M_\varepsilon)$ is obtained.

Take the sequence fragment MTMHTTMTTLTLTSL as an example, its reduced amino acid sequence is $\gamma\delta\gamma\beta\delta\delta\gamma\delta\delta\gamma\delta\gamma\delta\delta\gamma$. For the amino acid γ, which occurs in positions 1, 3, 7, 10, 12, and 15, it occurs a total of six times, then its average position is $\mu_\gamma = (1+3+7+10+12+15)/6 = 8$, and the mean absolute error of position is $M_\gamma = (|1-8| + |3-8| + \ldots + |15-8|)/6 = 13/3$. The mean absolute error of the other amino acid positions can be calculated in the same way, with M being 0 for amino acids that do not appear in the sequence.

2.2 Feature extraction based on amino acid order information

For a protein primary sequence with length $2L+1$, $S = s_{-L} \ldots s_0 \ldots s_L$, where $s_j \in \Omega$, $j = -L, \ldots, 0, \ldots, L$, Ω ={A,C,D,E,...,T,V,W,Y}. Shi et al [26] proposed position weighted amino acid composition (PWAA) to avoid losing sequence order information. This method has been used in many protein sites prediction, which can effectively extract the residual position information near the target position, thus improving the prediction accuracy of the target. The formula is as follows:

$$C_i = \frac{1}{L(L+1)} \sum_{j=-L}^{L} f_i(s_j) \left(j + \frac{|j|}{L} \right) \tag{4}$$

$$f_i(s_j) = \begin{cases} 0, & s_j \neq i \\ 1, & s_j = i \end{cases} \tag{5}$$

where $i = A, C, D, E, \ldots, T, V, W, Y$, $\quad j = -L, \ldots, 0, \ldots, L$, L denotes

the number of amino acids upstream or downstream from the central site in the protein sequence.

However, the method has some limitations. It is only applicable to protein fragments with odd sequence lengths. Therefore, we made some modifications on Shi's method by combining the information of the number of occurrences and positions of each amino acid, and the new method is applicable to protein sequences of arbitrary length.

For a protein sequence of length n, $S = s_1 s_2 s_3 ... s_n$, we define the formula as follows:

$$C_i^* = \frac{1}{L_i(L_i+1)} \sum_{j=1}^{n} f_i(s_j) \left(j - \mu_i^* + \frac{|j - \mu_i^*|}{L_i} \right) \tag{6}$$

$$f_i(s_j) = \begin{cases} 0, & s_j \neq i \\ 1, & s_j = i \end{cases} \tag{7}$$

$$\mu_i^* = Round(\mu_i) = Round \left(\frac{\sum_{j=1}^{n} f_i(s_j) \cdot j}{n_i} \right) \tag{8}$$

Where n_i denotes the number of the ith amino acid, μ_i^* denotes the integer value obtained after rounding the average position, and L_i denotes the number of the ith amino acid appearing in the average position μ_i^* and its upstream. It should be noted that the meaning of L here is different from that in Equation 4.

We use the sequence fragment MTMHTTMTTLTLTSL to explain the method. As we can see, the amino acid M occurs at positions 1, 3 and 7, according to Equation (8), $\mu_M^* = Round((1+3+7)/3) = 4$, $L_M = 2$. According to equation (6), $C_M^* = (1 - 4 + \frac{|1-4|}{2} + 3 - 4 + \frac{|3-4|}{2} + 7 - 4 + \frac{|7-4|}{2})/(2 * 3) = 5/12$. Similarly, C_T^*, C_H^*, C_L^* and C_S^* can be calculated. If the amino acid did not appear in the sequence, its C^* value is 0. Therefore for a protein primary sequence, the above equations lead to a final 20-dimensional feature vector $C^* = (C_A^*, C_C^*, \ldots, C_Y^*)$.

Based on three sets of feature vectors R, M and C^* extracted above, a 36-dimensional protein sequence feature vector can be obtained, denoted as $(R_\alpha, R_\beta, \ldots, R_\varepsilon, M_\alpha, M_\beta, \ldots, M_\varepsilon, C_A^*, C_C^*, \ldots, C_Y^*)$. It contains the num-

ber, position and order information of amino acids, and at the same time, the first two feature vectors R and M are obtained from the clustering results of the physicochemical properties of amino acids, so they also contain physicochemical property features of amino acids.

In addition, if we consider 22 amino acids (plus Selenocysteine and Pyrrolysine). We can first cluster 22 amino acids into m classes using K-Means algorithm based on their physicochemical properties, then calculate the frequency, the mean absolute error of the position of amino acids in reduced amino acid sequences, and the order information of 22 amino acids in the original sequences. Finally, a $2m$+22-dimensional feature vector will be extracted from each protein sequence.

3 Protein sequence similarity and evolutionary analysis

To illustrate the utility of the above feature vectors of protein sequences, we will apply it to the comparison of protein primary sequences. In order to eliminate the effect of the inconsistency of the magnitude between the features, it is also necessary to perform the minimum-maximum normalization to it firstly. Then the similarities between two protein sequences are computed by using the Euclidean distance. The smaller the Euclidean distance is, the more similar the sequences are. Finally, the UPGMA algorithm is used to construct the evolutionary tree of biological sequences.

3.1 Sequences of transferrin (TFs) from 24 vertebrate species

The first dataset is the sequences of transferrin (TFs) from 24 vertebrate species ([27]), the detailed information are provided in Table 3. A phylogenetic tree is constructed for this data set and the result is shown in Figure 3. As can be seen in Figure 3, all transferrin (TF) and lactoferrin (LF) protein sequences were categorized accurately and formed to four branches. The first branch was the transferrin sequences of all fishes, the

second branch is the mammalian transferrin sequence, the third branch is the amphibian transferrin sequence and the fourth branch is all the lactoferric (LF) protein sequences. In the first branch, TFs belonged to Salmo (Brown trout TF, Atlantic salmon TF), Salvelinus (Lake trout TF, Japanese char TF, Brook trout TF), and Oncorhynchus (Sockeye salmon TF, Rainbow trout TF, Chinook salmon TF, Coho salmon TF, Amago salmon TF) were classified accurately, which is in agreement with the known evolutionary relationships ([15]).

Table 3. The concise information for 24 TF protein sequences.

Sequence name	Species	Accession no.	Length
Human TF	Homo sapiens	S95936	698
Rabbit TF	Oryctolagus coniculus	X58533	695
Rat TF	Rattus norvegicus	D38380	698
Cow TF	Bos Taurus	U02564	704
Buffalo LF	Bubahts arnee	AJ005203	708
Cow LF	Bos Taurus	X57084	708
Goat LF	Copra hircus	X78902	708
Camel LF	Camehts dromedaries	AJ131674	708
Pig LF	Sus scrofa	M92089	704
Human LF	H. sapiens	NM_002343	710
Mouse LF	Mus musculus	NM_008522	707
Possum TF	Trichosurus vulpecula	AF092510	711
Frog TF	Xenopus laevis	X54530	702
Japanese flounder TF	Pctralichthys olivaceiis	D88801	685
Atlantic salmon TF	Salmo salar	L20313	690
Brown trout TF	Salmo trutta	D89091	691
Lake trout. TF	Salvelimts namaycush	D89090	691
Brook trout TF	Sahelinus fontinalis	D89089	691
Japanese char TF	Sahelinus phius	D89088	691
Chinook salmon TF	Oncorhynchus tshawytscha	AH008271	677
Coho salmon TF	Oncorhynchus kisuich	D89084	691
Sockeye salmon TF	Oncorhynchus nerka	D89085	691
Rainbow trout TF	Oncgrhynchus mykiss	D89083	691
Amago salmon TF	Oncorhynchus masou	D89086	691

In contrast, we compare the phylogenetic tree in Figure 3 with that constructed by conventional ClustalW (showed in Figure 4) and other alignment-free methods. From Figure 4, we can see that the transferrin (TF) and lactoferrin (LF) protein sequences were not separated completely, and so did the amphibians and mammals. In Ref. [27], the Japanese flounder TF and other fish transferrin sequences were not clustered together, and the Possum TF was closer to other fish transferrin protein sequences

than the Japanese flounder TF. In Ref.([28]), Possum TF was not clustered with other mammalian transferrin sequences and formed a separate branch.

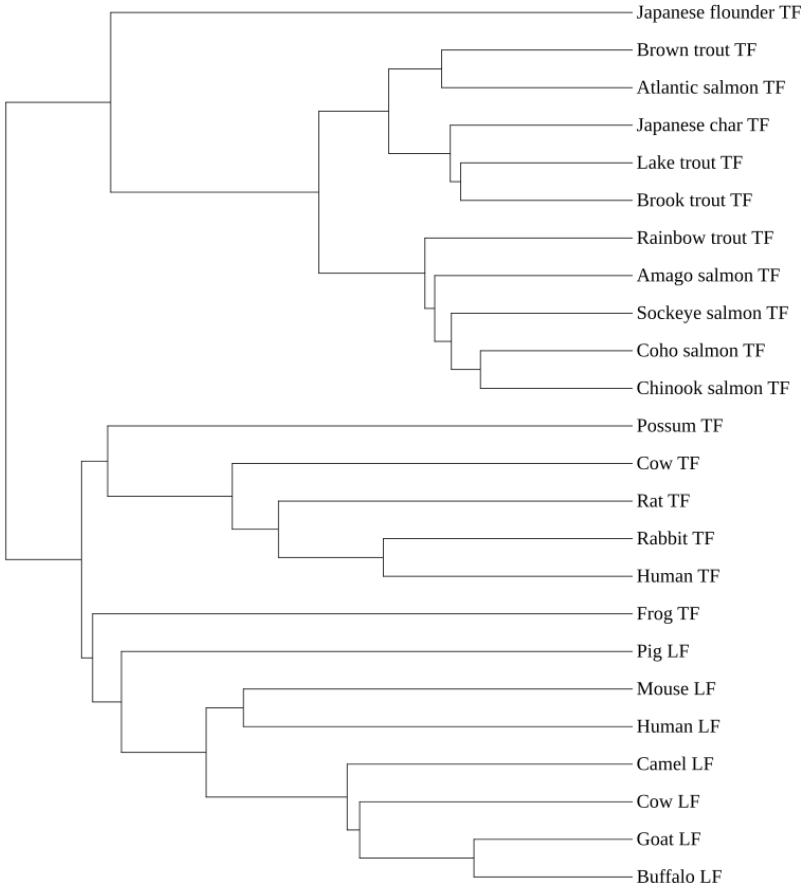

Figure 3. Phylogenetic tree of 24 transferrin sequences

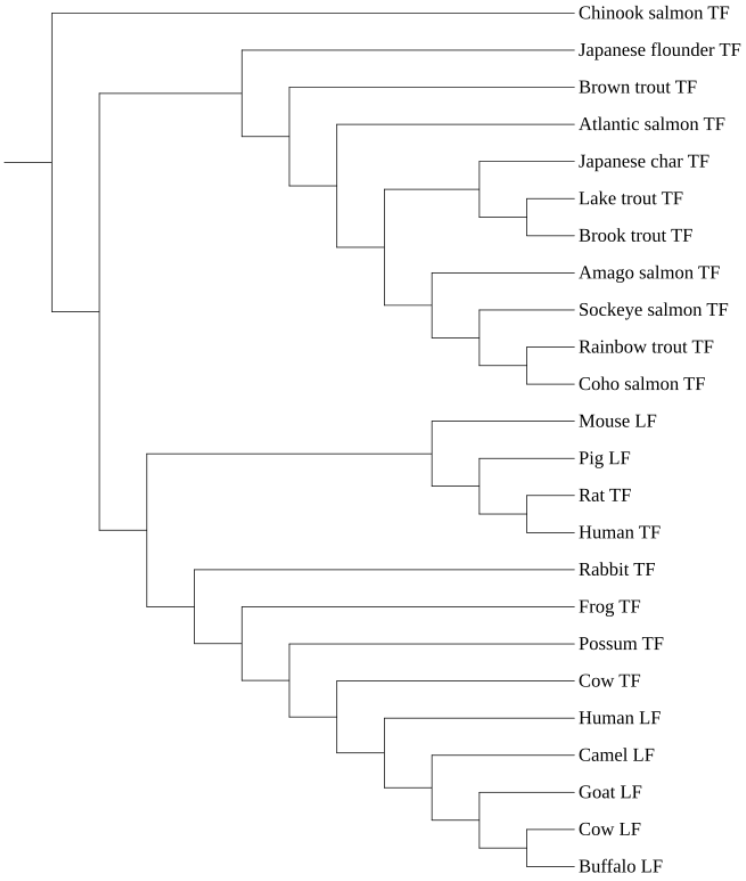

Figure 4. Phylogenetic tree of 24 transferrin (TFs) sequences constructed by ClustalW

3.2 Sequences of 35 coronavirus spike proteins

The second dataset consists of 35 coronavirus spike proteins, and their information is shown in Table 4. Spike protein is a transmembrane glycoprotein of SARS-CoV-2 with petal shaped protrusions outside the envelope, which can bind to cell receptors and allow the genetic material of the virus to invade host cells. It is a part of the Covid-19 virus that can circulate around the body and bind to ACE2 in the body, causing damage to cells, tissues, and organs ([29]).

Table 4. The information of 35 coronavirus spike proteins.

ID(NCBI)	Abbreviation	Name	Group
P10033	FIPV-1146	Feline infectious peritonits virus strain 79-1146	I
Q66928	FCoV-1683	Feline coronavirus strain 79-1 683	I
Q91AV1	PEDVC	Porcine epidemic diarrhea virus strain CV777	I
Q9DY22	TGEVT	Transmissible gastroenteritis virus strain TO14	I
P18450	TGEVF	Porcine transmissisble gastroenterits coronavirus strain FS772/70	I
P36300	CECoV	Canine enteric coronavirus strain INSAVC-1 I	I
Q9J3E7	MHVM	Murine hepatitis virus strain ML-10	II
Q83331	MHVB	Murine hepatitis virus strain Berkeley	II
P11224	MHVA	Murine hepatitis virus strain A59	II
O55253	MHVD	Murine hepatitis virus strain DVM	II
Q9IKD1	RtCoV	Rat coronavirus strain 681	II
P25190	BCoVF	Bovine coronavirus strain F15	II
P15777	BCoVM	Bovine coronavirus strain Mebus	II
Q9QAR5	BCoVL	Bovine coronavirus strain LSU-94LSS-051	II
Q91A26	BCoVT	Bovine enteric coronavirus 98TXSF-110-ENT	II
P36334	HCoV-OC43	Human coronavirus strain OC43	II
Q82666	IBV	Infectious bronchitis virus	III
P05135	IBV-6/82	Avian infectious bronchitis virus strain 6/82	III
P12722	IBVD	Avian infectious bronchitis virus strain D274	III
Q64930	IBVC	Infectious bronchitis virus strain CU-T2	III
Q82624	IBVA	Infectious bronchitis virus strain Ark99	III
P11223	IBVB	Avian infectious bronchitis virus strain Beaudette	III
Q98Y27	IBVH	Infectious bronchitis virus strain H52	III
AAP41037	Tor2	SARS coronavirus Tor2	IV
AAP30030	BJ01	SARS coronavirus BJ01	IV
AAR91586	NS-1	SARS coronavirus NS-1	IV
AAP51227	GD01	SARS coronavirus GD01	IV
AAP33697	Frankfurt 1	SARS coronavirus Frankfurt 1	IV
AAP13441	Urbani	SARS coronavirus Urbani	IV
AAQ01597	TC1	SARS coronavirus Taiwan TC1	IV
AAU81608	CDC	SARS Coronavirus CDC 200301157	IV
AAS00003	GZ02	SARS coronavirus GZ02	IV
AAR86788	QXC1	SARS coronavirus ShanghaiQXC1	IV
AAR23250	Sino1-11	SARS coronavirus Sino1-11	IV
AAT76147	TJF	SARS coronavirus TJF	IV

A phylogenetic tree is constructed using our method for this protein sequence dataset and the result is shown in Figure 5. As can be seen in Figure 5, the 35 coronavirus spike proteins were accurately categorized into four groups, and this is in agreement with the results obtained by other authors ([30, 31]).

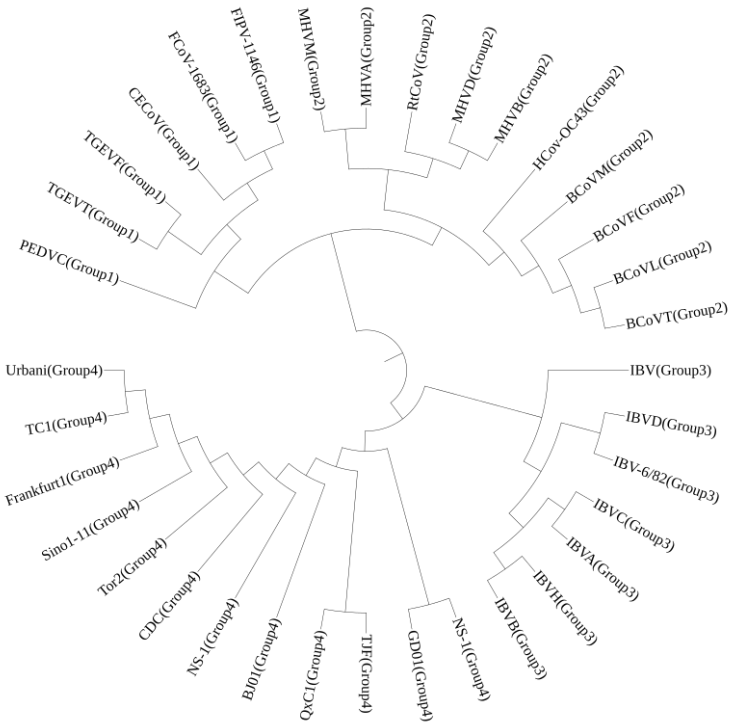

Figure 5. Phylogenetic tree of 35 coronavirus spike proteins

3.3 Sequences of 115 human rhinoviruses and 3 HEV-C viruses

The third dataset is 115 human rhinoviruses (HRV) and three sequences of HEV-C viruses, the detailed information is provided in Table 5. The HRV viruses are subdivided into HRV-A viruses, HRV-B viruses and HRV-C viruses ([32]). Furthermore, Palmenberg et al. ([33]) have proposed that HRV-A 45, HRV-A 95 and HRV-A 08 can be formed as a fourth category, named HRV-D, since it has some RNA elements that are not typical of other HRV-A strains.

The phylogenetic tree of 115 human rhinoviruses and 3 HEV-C viruses constructed by our method is shown in Figure 6, in which 3 HEV-C viruses, 26 HRV-B viruses, 6 HRV-C viruses, 3 HRV-D and 80 HRV-A viruses were

clustered correctly, and the results support the study of Palmenberg et al.([33]).

Table 5. The concise information for protein sequences of 115 human rhinoviruses and 3 HEV-C viruses.

ID(NCBI)	Abbreviation	ID(NCBI)	Abbreviation	ID(NCBI)	Abbreviation
AF499637	HEV.cva-13	FJ445117	A.hrv-13-f03	FJ445157	A.hrv-81
AF546702	HEV.cva-21	FJ445118	A.hrv-18	FJ445158	A.hrv-81-f06
AY751783	A.hrv-39	FJ445119	A.hrv-19	FJ445159	A.hrv-81-f07
DQ473485	B.hrv-03	FJ445120	A.hrv-20	FJ445160	A.hrv-82
DQ473486	B.hrv-06	FJ445121	A.hrv-21	FJ445161	B.hrv-83
DQ473488	B.hrv-48	FJ445122	A.hrv-22	FJ445162	B.hrv-84
DQ473489	B.hrv-70	FJ445123	A.hrv-25	FJ445163	A.hrv-85
DQ473490	B.hrv-04	FJ445124	B.hrv-26	FJ445164	B.hrv-86
DQ473491	A.hrv-41	FJ445125	A.hrv-29	FJ445165	A.hrv-89-f09
DQ473492	A.hrv-73	FJ445126	A.hrv-31	FJ445166	A.hrv-89-f08
DQ473493	A.hrv-15	FJ445127	A.hrv-32	FJ445167	A.hrv-90
DQ473494	A.hrv-74	FJ445128	A.hrv-33	FJ445168	B.hrv-91
DQ473496	A.hrv-49	FJ445129	A.hrv-40	FJ445169	B.hrv-92
DQ473497	A.hrv-23	FJ445130	B.hrv-42	FJ445170	A.hrv-95
DQ473499	A.hrv-44	FJ445131	A.hrv-43	FJ445171	A.hrv-96
DQ473500	A.hrv-59	FJ445132	A.hrv-45	FJ445172	B.hrv-97
DQ473504	A.hrv-88	FJ445133	A.hrv-47	FJ445173	A.hrv-98
DQ473505	A.hrv-36	FJ445134	A.hrv-49-f04	FJ445174	B.hrv-99
DQ473506	A.hrv-46	FJ445135	A.hrv-50	FJ445175	A.hrv-100
DQ473507	A.hrv-53	FJ445136	A.hrv-51	FJ445176	A.hrv-07
DQ473508	A.hrv-28	FJ445137	B.hrv-52-f10	FJ445177	A.hrv-09
DQ473510	A.hrv-75	FJ445138	A.hrv-54	FJ445178	A.hrv-10
DQ473511	A.hrv-55	FJ445139	A.hrv-54-f05	FJ445179	A.hrv-30
EF077279	C.nat001	FJ445140	A.hrv-56	FJ445180	A.hrv-38
EF077280	C.nat045	FJ445141	A.hrv-57	FJ445181	A.hrv-64
EF173414	A.hrv-11	FJ445142	A.hrv-58	FJ445182	A.hrv-76
EF173415	A.hrv-12	FJ445143	A.hrv-60	FJ445183	A.hrv-78
EF173420	B.hrv-17	FJ445144	A.hrv-61	FJ445184	A.hrv-89
EF173423	B.hrv-37	FJ445145	A.hrv-62	FJ445185	A.hrv-94
EF173425	B.hrv-93	FJ445146	A.hrv-63	FJ445186	B.hrv-27
EF186077	C.qpm	FJ445147	A.hrv-65	FJ445187	B.hrv-35
EF582385	C.c024	FJ445148	A.hrv-66	FJ445188	B.hrv-52
EF582386	C.c025	FJ445149	A.hrv-67	FJ445189	A.hrv-34
EF582387	C.c026	FJ445150	A.hrv-68	FJ445190	A.hrv-24
FJ445111	A.hrv-01	FJ445151	B.hrv-69	L05355	B.hrv-14
FJ445112	B.hrv-05	FJ445152	A.hrv-71	L24917	A.hrv-16
FJ445113	A.hrv-08	FJ445153	B.hrv-72	V01149	HEV.pv-1m
FJ445114	A.hrv-09-f01	FJ445154	A.hrv-77	X02316	A.hrv-02
FJ445115	A.hrv-09-f02	FJ445155	B.hrv-79		
FJ445116	A.hrv-13	FJ445156	A.hrv-80		

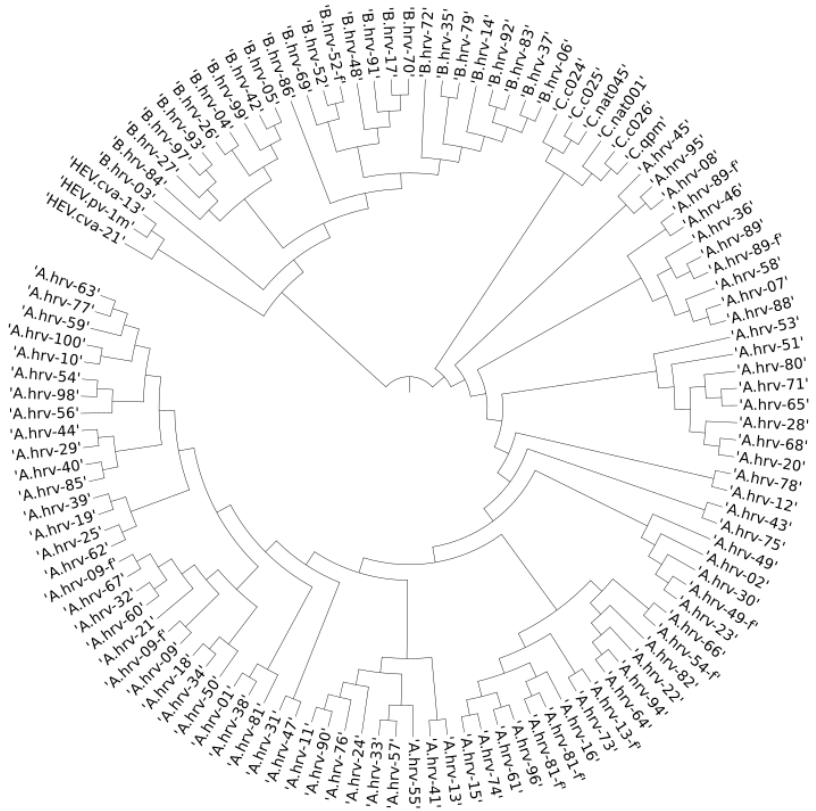

Figure 6. Phylogenetic tree of protein sequences of 115 human rhinoviruses and 3 HEV-C viruses

4 Conclusion

In this paper, we proposed a new alignment-free method for protein sequence comparison. We extract a 36-dimensional feature vector for each protein sequence containing the frequency, the mean absolute error of the position of amino acids in reduced amino acid sequences, and the order information of 20 amino acids in the original sequences. Finally, the validation was carried out on three datasets, and the results demonstrated the effectiveness and applicability of our method. In addition, our approach does not require complicated calculation. The method is more simple,

convenient and fast. The novel feature extraction method proposed in this article can be further applied to protein subcellular localization, protein post-translational modifications and other related issues.

References

[1] S. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool, *J Mol. Biol.* **215** (1990) 403–410.

[2] S. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.* **25** (1997) 3389–3402.

[3] J. D. Tompson, D. G. Higgins, T. J. Gibson, CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic Acids Res.* **22** (1994) 4673–4680.

[4] Y. H. Pan, D. Qian, P. Zhu, Graphical transformation and similarity clustering analysis for protein sequences, *Life Sci. Res.* **22** (2018) 191–228.

[5] Y. P. Zhang, P. A. He, Graphical representation of protein sequences and its applications, *J. Zhejiang Sci. Tech. Univ.* **27** (2010) 308–314.

[6] Y. H. Yao, S. J. Yan, H. M. Xu, J. N. Han, X. Y. Nan, P. A. He, Q. Dai, Similarity/dissimilarity analysis of protein sequences based on a new spectrum-like graphical representation, *Evol. Bioinf.* **10** (2014) 87–96.

[7] H. Y. Wu, Y. S. Zhang, W. Chen, Z. C. Mu, Comparative analysis of protein primary sequences with graph energy, *Physica A* **437** (2013) 249–262.

[8] W. B. Hou, Q. H. Pan, M. F. He, A new graphical representation of protein sequences and its applications, *Physica A* **444** (2016) 996–1002.

[9] D. D. Sun, C. R. Xu, Y. S. Zhang, A novel method of 2D graphical representation for proteins and its application, *MATCH Commun. Math. Comput. Chem.* **75** (2016) 431–446.

[10] M. Randić, J. Zupan, A. T. Balaban, D. Vikić-Topić, D. Plavsić, Graphical representation of proteins, *Chem. Rev.* **111** (2011) 790–862.

[11] P. A. He, S. N. Xu, Q. Dai, Y. H. Yao, A generalization of CGR representation for analyzing and comparing protein sequences, *Int. J. Quantum Chem.* **116** (2016) 476–482.

[12] J. Li, P. Koehl, 3D representations of amino acids – applications to protein sequence comparison and classification, *Comput. Struct. Biotech. J.* **11** (2014) 47–58.

[13] H. L. Hu, Z. Li, H. W. Dong, T. H. Zhou, Graphical representation and similarity analysis of protein sequences based on fractal interpolation, *IEEE/ACM Trans. Comput. Biol. Bioinf.* **14** (2017) 182–192.

[14] A. Czerniecka, D. Bielińska-Wąż, P. Wąż, T. Clark, 20D-dynamic representation of protein sequences, *Genomics* **107** (2016) 16–23.

[15] Z. Mu, T. Yu, E. Qi, J. Liu, G. Li, DCGR: feature extractions from protein sequences based on CGR via remodeling multiple information, *BMC Bioinf.* **20** (2019) #351.

[16] Z. H. Qi, K. C. Li, J. L. Ma, Y. H. Yao, L. Y. Liu, Novel method of 3-dimensional graphical representation for proteins and its application, *Evol. Bioinf.* **14** (2018) 1–8.

[17] L. He, Y. K. Li, R. L. He, S. S. T. Yau, A novel alignment-free vector method to cluster protein sequences, *J. Theor. Biol.* **427** (2017) 41–52.

[18] C. Li, Q. Dai, P. A. He, A time series representation of protein sequences for similarity comparison, *J. Theor. Biol.* **538** (2022) #111039.

[19] K. C. Chou, Some remarks on protein attribute prediction and pseudo amino acid composition, *J. Theor. Biol.* **273** (2011) 236–247.

[20] W. Chen, H. Lin, K. C. Chou, Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences, *Mol. Biosys.* **11** (2015) 2620–2634.

[21] X. H. Xie, Z. G. Yu, G. S. Han, V. Anh, Whole-proteome based phylogenetic tree construction with inter-amino-acid distances and the conditional geometric distribution profiles, *Mol. Phylogen. Evol.* **89** (2015) 37–45.

[22] Y. S. Li, T. Song, J. S. Yang, Y. Zhang, J. L. Yang, An alignment-free algorithm in comparing the similarity of protein sequences based on pseudo-markov transition probabilities among amino acids, *PLoS One* **11** (2016) #e0167430.

[23] Y. K. Li, K. Tian, C. C. Yin, R. L. He, S. S. T. Yau, Virus classification in 60-dimensional protein space, *Mol. Phylogen. Evol.* **99** (2016) 53–62.

[24] Z. C. Mu, J. Wu , Y. S. Zhang, A novel method for similarity/dissimilarity analysis of protein sequences, *Physica A* **392** (2013) 6361–6366.

[25] L. Ai, J. Feng, Y. H. Yao, A novel fast approach for protein classification and evolutionary analysis, *MATCH Commun. Math. Comput. Chem.* **90** (2023) 381–398.

[26] S. Shi, J. Qiu, X. Sun, S. B. Suo, S. Y. Huang, R. P. Liang, A method to distinguish between lysine acetylation and lysine methylation from protein sequences, *J. Theor. Biol.* **310** (2012) 223–230.

[27] L. Yu, Y. Zhang, I. Gutman, Y. Shi, M. Dehmer, Protein sequence comparison based on physicochemical properties and the position-feature energy matrix, *Sci. Rep.* **7** (2017) #46237.

[28] G. Chang, T. Wang, Phylogenetic analysis of protein sequences based on distribution of length about common substring, *Protein J.* **30** (2011) 167–172.

[29] M. M. Rajah, A. Bernier, J. Buchrieser, O. Schwartz, The mechanism and consequences of SARS-CoV-2 spike-mediated fusion and syncytia formation, *J Mol. Biol.* **434** (2022) #167280.

[30] Z. Mu, T. Yu, X. Liu, H. Zheng, L. Wei, J. Liu, FEGS: a novel feature extraction model for protein sequences and its applications, *BMC Bioinf.* **22** (2021) #297.

[31] C. Wu, R. Gao, Y. de Marinis, Y. Zhang, A novel model for protein sequence similarity analysis based on spectral radius, *J. Theor. Biol.* **446** (2018) 61–70.

[32] C. L. McIntyre, N. J. Knowles, P. Simmonds, Proposals for the classification of human rhinovirus species A, B and C into genotypically assigned types, *J. Gen. Virol.* **94** (2013) 1791–1806.

[33] A. C. Palmenberg, D. Spiro, R. Kuzmickas, S. L. Wang, A. Djikeng, J. A. Rathe, C. M. Fraser-Liggett, S. B. Liggett, Sequencing and analyses of all known human rhinovirus genomes reveal structure and evolution, *Science* **324** (2009) 55–59.