MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **90** (2023) 381–398
ISSN: 0340–6253
doi: 10.46793/match.90-2.381A

A Novel Fast Approach for Protein Classification and Evolutionary Analysis

Liang Ai[a,], **Jie Feng**[a,*], **Yu Hua Yao**[b,*]

[a] *School of Science, Minzu University of China, Beijing 100081, China*
[b] *School of Mathematics and Statistics, Hainan Normal University, Haikou 571158, China*
ailiang@muc.edu.cn, fengjie0536@163.com, yaoyuhua2288@163.com

(Received November 19, 2022)

Abstract

In this paper, we propose a new fast alignment-free method for protein sequence similarity and evolutionary analysis. First 20 natural amino acids are clustered into 6 groups based on their physicochemical properties, then a 12-dimensional vector is constructed based on the frequency and the average position of occurrence of amino acids in each reduced amino acid sequences. Finally, the Euclidean distance is used to measure the similarity and evolutionary distance between protein sequences. The test on three datasets shows that our method can cluster each protein sequence accurately, which illustrates the effective of our method.

1 Introduction

Similarity analysis of biological sequences is one of the important research directions in bioinformatics. In early research, multiple sequence alignment is usually used to compare and analyze sequences. Many algorithms are very mature [1–3], such as ClustalW algorithm. However, multiple sequence alignment is based on the assumption that homologous

[*] Corresponding author: fengjie0536@163.com; yaoyuhua2288@163.com

sequence fragments are adjacent and conservative, which conflicts with genetic recombination. Moreover, when the sample size is large or the sequence length is long, the time cost of the alignment algorithm is high. Therefore, as soon as the alignment-free method [4] was introduced, it received extensive attention from researchers immediately. The alignment-free method doesn't compare base pair, it takes the sequence as a whole and converts it into a numerical vector for analysis and comparison. Its advantage is that the calculation is fast on the computer and the results are accurate.

The comparison of protein sequences can be roughly divided into two categories: graphical representation methods and numerical vector characterization methods. The basic idea of graphical representation method is mapping amino acids into points in planar or spatial, and then connect the points to obtain spatial curves. Furthermore, we can extract the numerical features of biological sequences from these graphical representations, and use these numerical features for sequence analysis [5–15]. The numerical vector characterization method mainly used to convert the protein sequence into multi-dimensional numerical vector. For example, Chou [16] and Chen et al. [17] combine the 20-dimensional frequency vector of amino acids with the physicochemical properties or interactions between amino acids to construct a $20 + \lambda$ dimensional vector to represent the protein sequence, in which λ refers to the number of physicochemical properties or indicators of interactions between amino acids. Xie et al. [18] used the relative deviation between the random and independently placed sequence distribution maps of amino acids to define the differences among sequences. Li et al. [19] combined the probability of amino acids, the average occurrence location probability and the Markov transfer probability distribution of two adjacent amino acids to construct the protein numerical vector representation. Li et al. [20] used the number of amino acids, the average position and the secondary central moment of normalization of position of 20 amino acids in the protein sequence to form a 60-dimensional numerical vector to measure the similarity between viruses. He et al. [21] selected three biochemical properties of amino acids: the hydropathy index, polar requirement and chemical composition of the side chain, and proposed a 24 dimensional feature vector to compare protein sequences. Mu et al. [22] introduce the concept of distance frequency of amino acid pairs and propose a new numerical characterization of protein sequences, which converts any protein sequence into a distance frequency matrix.

Proteins are composed of amino acids. Previous studies have shown that the physicochemical properties of amino acids are important for protein sequence classification and evolution [23, 24]. In this paper, we cluster 20 natural amino acids into 6 groups based on their physicochemical properties, then a 12-dimensional vector is constructed based on the frequency and the average position of occurrence of amino acids in each reduced amino acid 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 Materials and methods

2.1 Reduced amino acid sequences

The physicochemical properties of amino acids play an important role in protein sequence classification and evolution [23, 24]. In this paper, 20 natural amino acids are sorted into six groups based on their four physicochemical properties, then a 20-letter protein primary sequence can be converted into a 6-letter reduced protein sequence. The four physicochemical properties are dissociation constant value (pK_a), hydropathy index (Hy), polar requirement (Pr) and chemical composition of the side chain (Cc), the values of these properties are listed in Table 1.

In order to eliminate the impact of inconsistency in magnitude of physicochemical properties, we normalize them by equation (1).

$$x'_{ik} = \frac{x_{ik} - \bar{x}_k}{\sigma_k} \quad i = 1, 2, ..., 20; k = 1, 2, 3, 4 \tag{1}$$

where x'_{ik} is the normalized value, x_{ik} is the component of the ith row and the kth column in Table 1,

Table 1. Four physicochemical properties of 20 amino acids

	pK_a	Hy	Pr	Cc
A(Ala)	0	1.8	7	0
C(Cys)	1	2.5	9.1	0.65
D(Asp)	0	-3.5	10	1.33
E(Glu)	1	-3.5	13	1.38
F(Phe)	1	2.8	4.8	2.75
G(Gly)	0	-0.4	8.6	0.89
H(His)	1	-3.2	12.5	0.92
I(Ile)	0	4.5	7.9	0.74
K(Lys)	1	-3.9	8.4	0.58
L(Leu)	0	3.8	4.9	0
M(Met)	0	1.9	4.9	0
N(Asn)	1	-3.5	10.1	0.33
P(Pro)	0	-1.6	5.3	0
Q(Gln)	0	-3.5	5	0
R(Arg)	0	-4.5	6.6	0.39
S(Ser)	0	-0.8	7.5	1.42
T(Thr)	0	-0.7	6.6	0.71
V(Val)	0	4.2	5.2	0.13
W(Trp)	1	-0.9	5.4	0.2
Y(Tyr)	0	-1.3	5.6	0

$$\bar{x}_k = \frac{1}{20}\sum_{i=1}^{20} x_{ik} \tag{2}$$

and

$$\sigma_k = \sqrt{\frac{1}{20-1}\sum_{i=1}^{20}(x_{ik} - \bar{x}_k)^2} \tag{3}$$

are the mean value and the standard deviation of the corresponding property, respectively.

We then cluster 20 amino acids into groups based on the normalized physicochemical property values. The similarities between each two amino acids are calculated by Euclidean distance:

$$d_{ij} = \sqrt{\sum_{k=1}^{4}(x'_{ik} - x'_{jk})^2} \quad i, j = 1, 2, ..., 20 \tag{4}$$

The average linkage method is a good systematic clustering method in many cases, it makes full use of the information between all samples [25]. Here we use the average linkage method to measure the distance between two groups, suppose n_K and n_L are the number of samples in groups G_K and G_L, respectively, and d_{ij} is the distance between sample i in group G_K and sample j in G_L, then the distance between groups G_K and G_L is:

$$D_{KL} = \frac{1}{n_K \times n_L} \sum_{i \in G_K, j \in G_L} d_{ij} \tag{5}$$

For example, as shown in Figure 1, the distance between two groups is $(d_{13} + d_{14} + d_{15} + d_{23} + d_{24} + d_{25})/(2 \times 3)$.

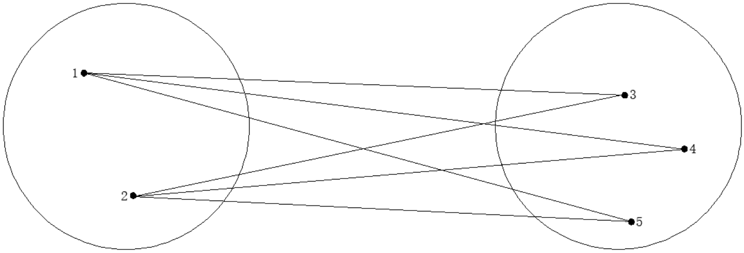

Figure 1. Schematic diagram of the average linkage method.

Based on the normalized physicochemical property values, we calculate the similarities of 20 amino acids and obtain a 20 × 20 distance matrix, then we use the distance matrix to conduct cluster analysis, the average linkage method is used to measure the distance between two groups. In Figure 2, we list the cluster results of 20 natural amino acids based on four physicochemical properties. In order to determine the appropriate groups of amino acids, we divide 20 natural amino acids into 3 to 10 groups according to Figure 2, and then construct phylogenetic trees for three data sets in this paper under different groups. Finally, we find that the best result occurs when 20 amino acids are divided into six groups. Therefore, we cluster 20 natural amino acids into six groups, they are {C}, {A,V,I,L,F,M}, {W,P,G,T,S,N,Q}, {Y}, {D,E} and {K,H,R}. We use C,A,W,Y,D and K to denote these six categories respectively, as shown in Table 2.

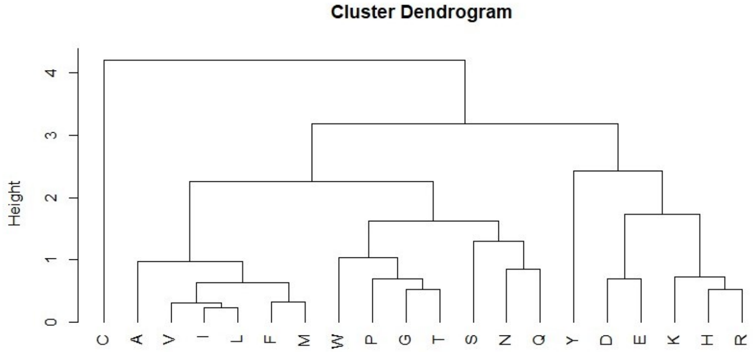

Figure 2. Cluster results of 20 natural amino acids.

Table 2. Classification of the 20 natural amino acids

Amino acids	Denote
C	C
A,V,I,L,F,M	A
W,P,G,T,S,N,Q	W
Y	Y
D,E	D
K,H,R	K

According to Table 2, a 20-letter protein primary sequence can be converted into a 6-letter reduced amino acid sequence. For example, the first 20 characters of human rhinovirus (A hrv-02) are MGAQVSRQN-VGTHSTQNSVS, its reduced alphabet representation is AWAWAWK-WWAWWKWWWWWAW.

2.2 Feature vectors of protein sequences

A reduced amino acid sequence $S = s_1 s_2 s_3 ... s_N$ can be viewed as a linear sequence of N symbols from a finite alphabet $\Omega =$\{C,A,W,Y,D,K\}, that is $s_i \in \Omega, i = 1, 2, ..., N$. We consider two features for each reduced amino sequence, one is the frequency of occurrence of amino acid α, $\alpha \in \Omega$, which is denoted as f_α, and another is the average position of the occurrence of α [21], which is denoted as μ_α. f_α and μ_α are defined as follows:

$$f_\alpha = \frac{\sum_{i=1}^{N} I_{\{s_i=\alpha\}}}{N} \qquad \mu_\alpha = \frac{\sum_{i=1}^{N} i \times I_{\{s_i=\alpha\}}}{\sum_{i=1}^{N} I_{\{s_i=\alpha\}}} \tag{6}$$

in which

$$I_{\{s_i=\alpha\}} = \begin{cases} 0 & , s_i \neq \alpha \\ 1 & , s_i = \alpha \end{cases} \tag{7}$$

Thus for a reduced amino acid sequence, we can obtain a 12-dimensional vector $V = (f_C, f_A, f_W, f_Y, f_D, f_K, \mu_C, \mu_A, \mu_W, \mu_Y, \mu_D, \mu_K)$, which can represent the original protein sequence.

2.3 Comparison of proteins

To illustrate the utility of the above feature vectors of protein sequences, we will apply it to the comparison of protein primary sequences. Due to the different magnitudes of the two features, the 12-dimensional vector needs to be normalized. The similarities between two protein sequences S_i and S_j are computed by using the Euclidean distance:

$$d_{ij} = \sqrt{\sum_{k=1}^{12} (V'_{ik} - V'_{jk})^2} \quad i, j = 1, 2, ..., N \tag{8}$$

The smaller the Euclidean distance is, the more similar the sequences are.

3 Results and discussion

In this section, the method is tested against three data sets. Given N protein primary sequences that are under research, we first convert them into corresponding reduced amino acid sequences. Then the distance of each two sequences are calculated according to Formula (8). At last, we arrange all these values into a matrix, a pair-wise distance matrix is derived. The distance matrix contains the similarity information on the N protein primary sequences, and it can be input to the UPGMA program in the MEGA package (`https://www.megasoftware.net/`) for phylogenetic analysis.

3.1 Phylogenetic analysis of influenza A viruses proteins

Influenza A virus has caused many pandemics around the world, its several subtypes are labeled according to H numbers (hemagglutinin type) and N numbers (neuraminidase type), in which the most lethal subtypes are H1N1, H2N2, H5N1, H7N3, and H7N9. In this section, we consider inferring the phylogenetic relationships of 35 influenza A virus protein sequences [21]. A phylogenetic tree is constructed using our method for this protein sequence dataset and the result is shown in Figure 3.

As we can see from Figure 3, the five influenza A virus subtypes H1N1, H2N2, H5N1, H7N3 and H7N9 are clustered accurately. In contrast, the phylogenetic tree constructed by conventional ClustalW has three subtypes clustered incorrectly, ACZ36780.1 (H5N1), ADI52832.1 (H1N1) and AIK26325.1 (H1N1) are misplaced as shown in Figure 4.

3.2 Phylogenetic analysis of human rhinovirus proteins

Next, we applied our method to analyze 115 human rhinoviruses (HRV) with three sequences of HEV-C as an outgroup for analysis. In past studies, researchers have found that HRV-A and HRV-C share a common ancestor which is a sister group to the HRV-B [26, 27], the phylogenetic

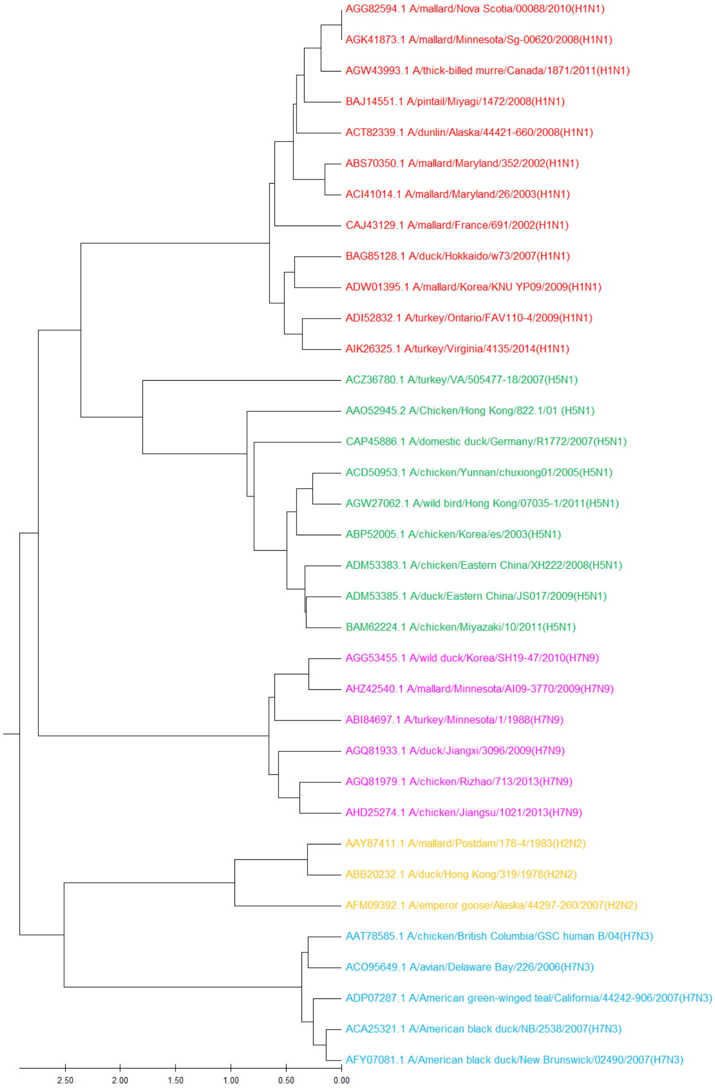
AGG82594.1 A/mallard/Nova Scotia/00088/2010(H1N1)
AGK41873.1 A/mallard/Minnesota/Sg-00620/2008(H1N1)
AGW43993.1 A/thick-billed murre/Canada/1871/2011(H1N1)
BAJ14551.1 A/pintail/Miyagi/1472/2008(H1N1)
ACT82339.1 A/dunlin/Alaska/44421-660/2008(H1N1)
ABS70350.1 A/mallard/Maryland/352/2002(H1N1)
ACI41014.1 A/mallard/Maryland/26/2003(H1N1)
CAJ43129.1 A/mallard/France/691/2002(H1N1)
BAG85128.1 A/duck/Hokkaido/w73/2007(H1N1)
ADW01395.1 A/mallard/Korea/KNU YP09/2009(H1N1)
ADI52832.1 A/turkey/Ontario/FAV110-4/2009(H1N1)
AIK26325.1 A/turkey/Virginia/4135/2014(H1N1)
ACZ36780.1 A/turkey/VA/505477-18/2007(H5N1)
AAO52945.2 A/Chicken/Hong Kong/822.1/01 (H5N1)
CAP45886.1 A/domestic duck/Germany/R1772/2007(H5N1)
ACD50953.1 A/chicken/Yunnan/chuxiong01/2005(H5N1)
AGW27062.1 A/wild bird/Hong Kong/07035-1/2011(H5N1)
ABP52005.1 A/chicken/Korea/es/2003(H5N1)
ADM53383.1 A/chicken/Eastern China/XH222/2008(H5N1)
ADM53385.1 A/duck/Eastern China/JS017/2009(H5N1)
BAM62224.1 A/chicken/Miyazaki/10/2011(H5N1)
AGG53455.1 A/wild duck/Korea/SH19-47/2010(H7N9)
AHZ42540.1 A/mallard/Minnesota/AI09-3770/2009(H7N9)
ABI84697.1 A/turkey/Minnesota/1/1988(H7N9)
AGQ81933.1 A/duck/Jiangxi/3096/2009(H7N9)
AGQ81979.1 A/chicken/Rizhao/713/2013(H7N9)
AHD25274.1 A/chicken/Jiangsu/1021/2013(H7N9)
AAY87411.1 A/mallard/Postdam/178-4/1983(H2N2)
ABB20232.1 A/duck/Hong Kong/319/1978(H2N2)
AFM09392.1 A/emperor goose/Alaska/44297-260/2007(H2N2)
AAT78585.1 A/chicken/British Columbia/GSC human B/04(H7N3)
ACO95649.1 A/avian/Delaware Bay/226/2006(H7N3)
ADP07287.1 A/American green-winged teal/California/44242-906/2007(H7N3)
ACA25321.1 A/American black duck/NB/2538/2007(H7N3)
AFY07081.1 A/American black duck/New Brunswick/02490/2007(H7N3)
2.50
2.00
1.50
1.00
0.50
0.00

Figure 3. Phylogenetic tree of 35 influenza A virus protein sequences constructed by our method. The dataset includes 5 groups: H1N1 (red), H5N1 (green), H7N9 (pink), H2N2 (orange), H7N3 (blue).

tree constructed using our method (Figure 5) is consistent with theirs.

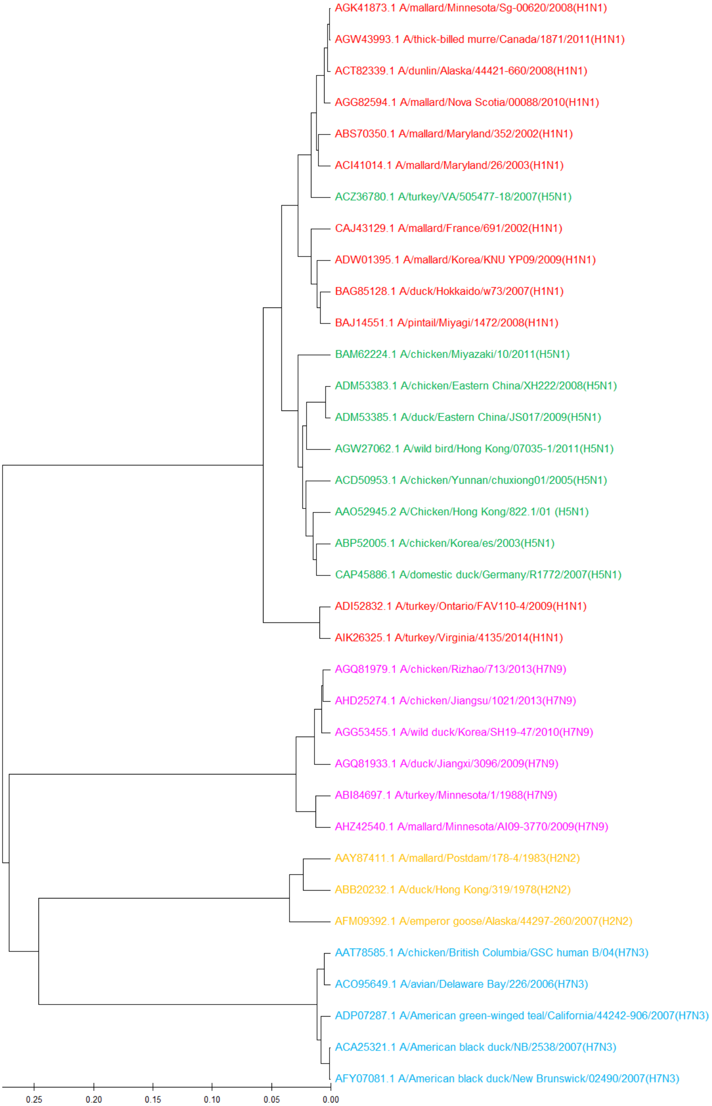
AGK41873.1 A/mallard/Minnesota/Sg-00620/2008(H1N1)
AGW43993.1 A/thick-billed murre/Canada/1871/2011(H1N1)
ACT82339.1 A/dunlin/Alaska/44421-660/2008(H1N1)
AGG82594.1 A/mallard/Nova Scotia/00088/2010(H1N1)
ABS70350.1 A/mallard/Maryland/352/2002(H1N1)
ACI41014.1 A/mallard/Maryland/26/2003(H1N1)
ACZ36780.1 A/turkey/VA/505477-18/2007(H5N1)
CAJ43129.1 A/mallard/France/691/2002(H1N1)
ADW01395.1 A/mallard/Korea/KNU YP09/2009(H1N1)
BAG85128.1 A/duck/Hokkaido/w73/2007(H1N1)
BAJ14551.1 A/pintail/Miyagi/1472/2008(H1N1)
BAM62224.1 A/chicken/Miyazaki/10/2011(H5N1)
ADM53383.1 A/chicken/Eastern China/XH222/2008(H5N1)
ADM53385.1 A/duck/Eastern China/JS017/2009(H5N1)
AGW27062.1 A/wild bird/Hong Kong/07035-1/2011(H5N1)
ACD50953.1 A/chicken/Yunnan/chuxiong01/2005(H5N1)
AAO52945.2 A/Chicken/Hong Kong/822.1/01 (H5N1)
ABP52005.1 A/chicken/Korea/es/2003(H5N1)
CAP45886.1 A/domestic duck/Germany/R1772/2007(H5N1)
ADI52832.1 A/turkey/Ontario/FAV110-4/2009(H1N1)
AIK26325.1 A/turkey/Virginia/4135/2014(H1N1)
AGQ81979.1 A/chicken/Rizhao/713/2013(H7N9)
AHD25274.1 A/chicken/Jiangsu/1021/2013(H7N9)
AGG53455.1 A/wild duck/Korea/SH19-47/2010(H7N9)
AGQ81933.1 A/duck/Jiangxi/3096/2009(H7N9)
ABI84697.1 A/turkey/Minnesota/1/1988(H7N9)
AHZ42540.1 A/mallard/Minnesota/AI09-3770/2009(H7N9)
AAY87411.1 A/mallard/Postdam/178-4/1983(H2N2)
ABB20232.1 A/duck/Hong Kong/319/1978(H2N2)
AFM09392.1 A/emperor goose/Alaska/44297-260/2007(H2N2)
AAT78585.1 A/chicken/British Columbia/GSC human B/04(H7N3)
ACO95649.1 A/avian/Delaware Bay/226/2006(H7N3)
ADP07287.1 A/American green-winged teal/California/44242-906/2007(H7N3)
ACA25321.1 A/American black duck/NB/2538/2007(H7N3)
AFY07081.1 A/American black duck/New Brunswick/02490/2007(H7N3)
0.25
0.20
0.15
0.10
0.05
0.00

Figure 4. Phylogenetic tree of 35 influenza A virus protein sequences constructed by ClustalW. The dataset includes 5 groups: H1N1 (red), H5N1 (green), H7N9 (pink), H2N2 (orange), H7N3 (blue).

As shown in Figure 5, all 3 HEV-C outgroup viruses, 26 HRV-B viruses, 6 HRV-C viruses and 83 HRV-A viruses are clustered correctly. Besides, the phylogenetic tree obtained by our method using protein sequences is in accordance with that obtained by Palmenberg [27] using whole genome sequences. Palmenberg suggested the clade of three viruses HRV-A 08, HRV-A 95 and HRV-A 45 to be a fourth class named HRV-D because clade D has RNA elements—such as the cis-acting replication element, the $3^{'}UTR$ terminal loop feature, and local insertions/deletions and sequence motifs—that are somewhat atypical of other HRV-A strains. Our results support Palmenberg's opinion.

In addition, the phylogenetic tree constructed by ClustalW method is shown in Figure 6. As we can see from Figure 6, three outgroup HEV-C viruses are clustered into a large branch with HRV-B virus instead of to be the outermost one.

3.3 Phylogenetic analysis of coronavirus spike proteins

The third data set is thirty-five coronavirus spike proteins which has been studied by different methods [28, 29]. Coronaviruses are species of virus which are associated with respiratory, intestinal, liver and neurological diseases. By comparing the homology of spike protein sequence among different years, different regions and different hosts, people can analyze the genetic variation and epidemic characteristics of spike protein. Our proposed method is utilized to analyze the homology of coronavirus spike proteins. A phylogenetic tree is constructed for this data set and the result is shown in Figure 7.

From Figure 7, we can see that the four groups of coronavirus spike proteins are clustered accurately. The phylogenetic tree obtained by our method is consistent with the results obtained by other authors [28, 29]. Furthermore, a phylogenetic tree is also produced by the multiple alignment algorithm ClustalW and the topology of the tree is totally same as that by our new method.

In this section, we applied our method to infer the phylogenetic rela-

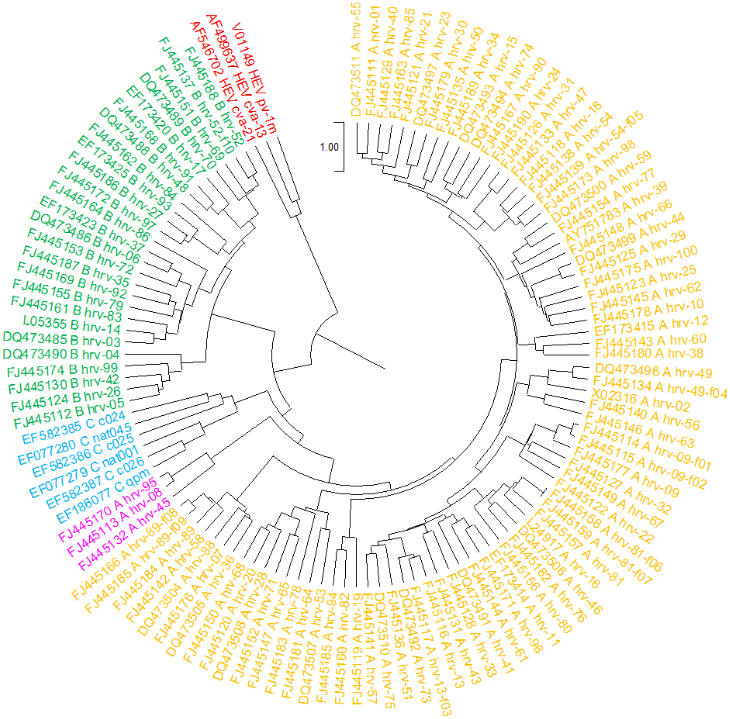

Figure 5. Phylogenetic tree of 115 human rhinoviruses and 3 control viruses constructed by our method. The HEV-C sequences (poliovirus 1M, coxsackievirus a13, and coxsackievirus a21) are used as outgroup. The dataset includes 5 groups: HEV-C (red), HRV-B (green), HRV-C (blue), HRV-D (pink), HRV-A (orange).

tionships of three data sets. The new approach does not require sequence alignment, it is fully automatic. In addition to the traditional sequence alignment method, we also used two alignment-free methods [20,21] to test three sets of data. Based on Li's method [20], 6 of 35 influenza A viruses, 12 of 118 human rhinovirus and 2 of 35 coronavirus spike proteins are not clustered correctly; Based on He's method [21], 35 influenza A viruses are clustered correctly, while 13 of 118 human rhinovirus and 1 of 35 coronavirus spike proteins are not clustered correctly. Moreover, our method has the advantage of less time, the running time of our method on three

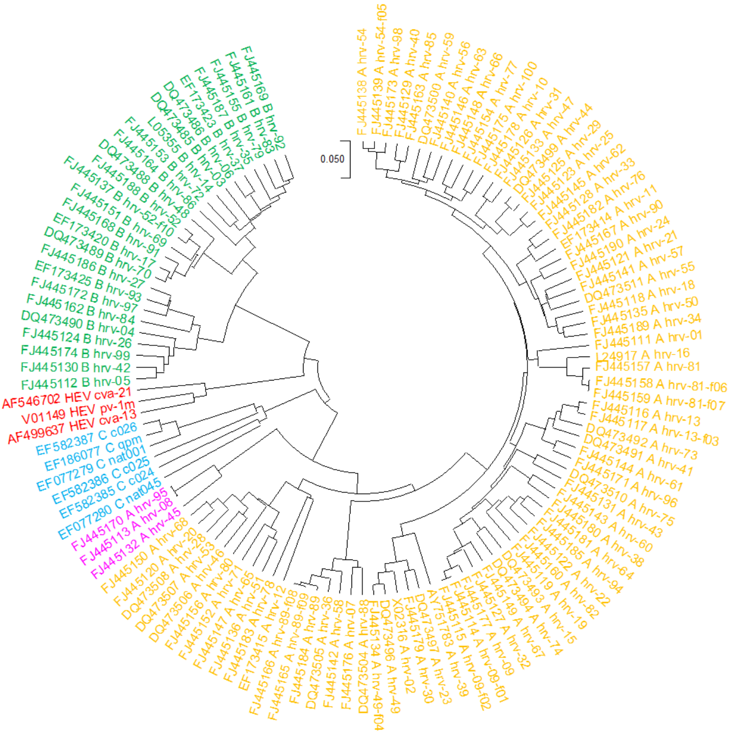

Figure 6. Phylogenetic tree of 115 human rhinoviruses and 3 control viruses constructed by ClustalW. The HEV-C sequences (poliovirus 1M, coxsackievirus a13, and coxsackievirus a21) are used as outgroup. The dataset includes 5 groups: HRV-B (green), HEV-C (red), HRV-C (blue), HRV-D (pink), HRV-A (orange).

data sets are 0.12 seconds, 0.48 seconds and 0.11 seconds separately, which are almost the least time-consuming method among these methods. The running time of different methods are detailed in Table 3.

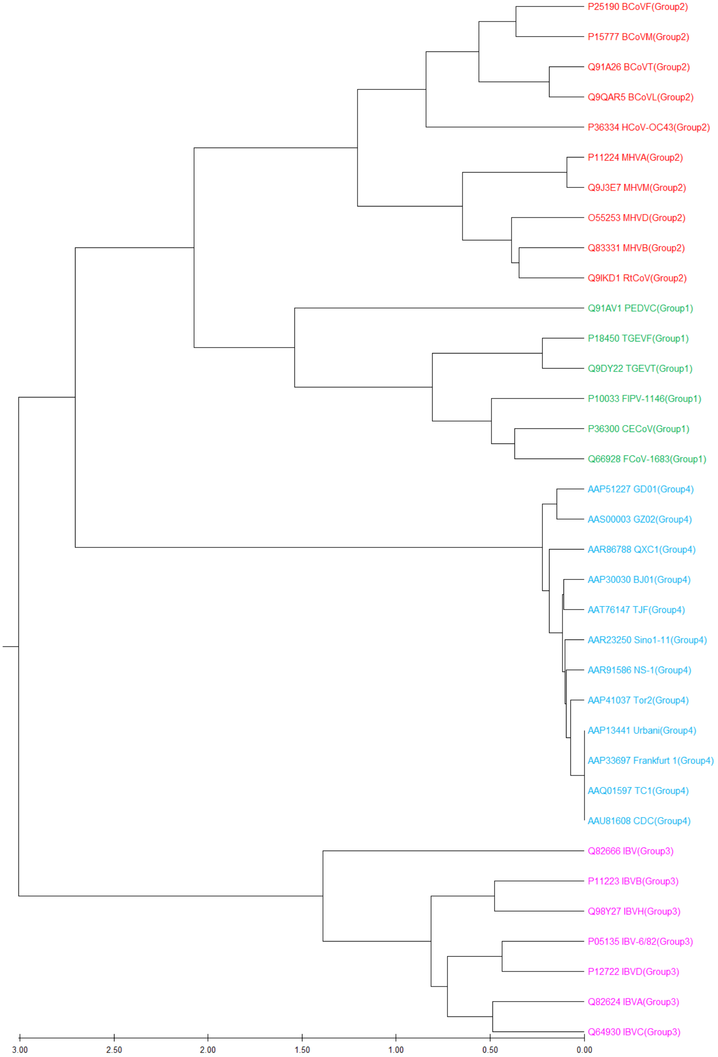

Figure 7. Phylogenetic tree of 35 coronavirus spike proteins constructed by our method. The dataset includes 4 groups: Group2 (red), Group1 (green), Group4 (blue), Group3 (pink).

Table 3. Running time of different methods

Method	Influenza A virus (35)	Human rhinoviruses (118)	Coronavirus spike proteins (35)
Our method	0.12sec	0.48sec	0.11sec
Li's method	0.09sec	1.60sec	0.26sec
He's method	0.36sec	5.56sec	0.96sec
ClustalW	7sec	30min	49sec

4 Conclusion

In this paper, we propose a novel alignment-free method for protein sequence comparison. The reduced amino acid alphabet of 6 types of amino acids based on four physicochemical properties is introduced, then a 12-dimensional vector is constructed based on the frequency and the average position of occurrence of amino acids in each reduced amino acid sequences for the comparison of protein primary sequences. The similarity between two protein sequences is expressed by Euclidean distance, which reflects the degree that one sequence distinguishes from another sequence. Three applications have demonstrated that the proposed approach in this work is a powerful and useful tool for protein comparison. Meanwhile, our approach does not require complicated calculation. The method is more simple, convenient and fast. It can be used for protein-protein interaction network (PPI) analysis or protein function prediction.

***Acknowledgment*:** This work is supported by National Natural Science Foundation of China (grant no.62162025) and Hainan Provincial Natural Science Foundation of China (grant no.122RC653).

References

[1] K. Katoh, K. Misawa, K. Kuma, T. Miyata, MAFFT: a novel method for rapid multiple sequence alignment based on fast fourier transform, *Nucleic Acids Res.* **30** (2002) 3059–3066.

[2] J. D. Thompson, T. J. Gibson, D. G. Higgins, Multiple sequence alignment using ClustalW and ClustalX, *Curr. Protoc. Bioinf.* **00** (2003) #2.3.

[3] R. C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput, *Nucleic Acids Res.* **32** (2004) 1792–1797.

[4] S. Vinga, J. Almeida, Alignment-free sequence comparison – a review, *Bioinformatics* **19** (2003) 513–523.

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

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.

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

[14] 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.

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

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

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

[18] 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.

[19] 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.

[20] 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.

[21] 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.

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

[23] L. Salichos, A. Rokas, Inferring ancient divergences requires genes with strong phylogenetic signals, *Nature* **497** (2013) 327–331.

[24] W. C. Wimley, S. H. White, Experimentally determined hydrophobicity scale for proteins at membrane interfaces, *Nat. Struct. Biol.* **3** (1996) 842–848.

[25] G. R. Li, M. X. Wu, *Multivariate Statistical Analysis*, Sci. Press, Beijing, 2021.

[26] S. E. Jacobs, D. M. Lamson, K. S. George, T. J. Walsh, Human rhinoviruses, *Clin. Microbiol. Rev.* **26** (2013) 135–162.

[27] 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.

[28] Z. C. Mu, G. J. Li, H. Y. Wu, X. Q. Qi, 3D-PAF curve: a novel graphical representation of protein sequences for similarity analysis, *MATCH Commun. Math. Comput. Chem.* **75** (2016) 447–462.

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