MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **75** (2016) 447-462

ISSN 0340 - 6253

3D–PAF Curve: A Novel Graphical Representation of Protein Sequences for Similarity Analysis

Zengchao Mu[1,2], Guojun Li[1], Haiyan Wu[2], Xingqin Qi[2*]

[1] *School of Mathematics, Shandong University, Jinan 250100, China.*

[2] *School of Mathematics and Statistics, Shandong University (Weihai),*
Weihai 264209, China.

(Received August 6, 2015)

Abstract. Based on the physicochemical properties of amino acids, in this paper, we first propose a novel graphical representation called 3D-PAF curve of protein sequence, which incorporates the accumulative frequencies of adjacent amino acids of the protein sequence. Then, we derive a 8-dimensional numerical vector to characterize a 3D-PAF curve. Because a protein sequence corresponds to 12 kinds of 3D-PAF curves, we take a 96-dimensional vector as the feature vector of the protein sequence. The similarity between any two protein sequences can be measured by the standardized Euclidean distance between their feature vectors. Finally we apply this new method on two data sets (nine ND5 proeins, and 35 coronavirus spike proteins) to analysis the similarities of protein sequences. The results both demonstrate the validity of our method.

1 Introduction

With more and more genome sequences being available on-line, biological sequence comparison becomes focus of research in bioinformatics and computational biology. Up to now, lots of methods have been proposed to analyze DNA and protein sequences. These methods can be classified into two categories: alignment-based [1–3] and alignment-free.

[*] Corresponding author: qixingqin@163.com

Alignment–based methods use dynamic programming; it generates a matrix whose elements represent all possible alignments between two sequences. The highest set of sequential scores in the matrix defines an optimal alignment. But the search for optimal solutions encounters difficulties in: (i) computational load with regard to large databases; (ii) choosing the scoring schemes. Therefore, alignment-free methods have been developed to overcome the limitations of alignment-based methods. The graphical representation of biological sequences is one of the most commonly used alignment-free method, which can not only transform biological sequences into visual curves but also offer effective numerical descriptors.

Graphical representation methods were firstly introduced for representation of DNA sequences on the basis of multiple dimension space. In 1983, Hamori and Ruskin firstly proposed a graphical representation to describe DNA sequences [4]. Since then, a large number of graphical representations of DNA sequences have been outlined [5–24]. The graphical representations of proteins emerged only very recently. The increased complexity of biological strings built on a 20-letter alphabet (representing the 20 natural amino acids) delayed the emergence of graphical representations of proteins in comparison with DNA whose strings are built from only four letters. To date, many researchers have put forward various methods of 2D and 3D graphical representation for protein sequences [25–45]. In these representations, 20 amino acids are usually first represented by 20 pre–given vectors. Then, a recurrence formula is given to generate a curve representing proteins based on these vectors, and the numerical characterizations of the curves are used to describe corresponding protein sequences. For example, using indexes of some physicochemical properties of 20 amino acids, He [25] ,Yu [26], Liu [27], Wu [28], Ma [29], Wen [30], Huang [31], Li [32], el Maaty [33] and Gupta [34] proposed a number of different graphical representations of proteins, respectively. Bai [35] presented a method of 3D graphical representation of protein sequences by mapping the 20 amino acids to the 20 vertices of the regular dodecahedron. el Maaty [36] selected a unit sphere to represent any protein sequence on its surface and Abo-Elkhier [37] represented any protein sequence on the surface of a right cone. The Chaos Game Representation for DNA sequences, introduced by Jeffrey [5], was generalized to obtain the graphical representation of protein sequences by He [38] and Randić [39]. They placed the 20 amino acids on the periphery of the unit circle, which is to replace a square with a 20-side polygon.

The preceding graphical representation techniques of protein sequence only consider the information of single amino acids, and do not consider the information between adjacent amino acids. In this article, we propose a novel graphical representation called 3D-PAF curve of protein sequences based on five-letter model of amino acids which converts the 20 amino acids to only five letters. Meanwhile, we first incorporate the accumulative frequencies of adjacent amino acids into 3D-PAF curves of protein sequences. Then we transform the 3D-PAF curve into a numerical characterization that will facilitate quantitative comparisons of protein sequences. Based on the distance between the feature vectors of two protein sequences, the similarity matrices among proteins can be calculated. Finally, we apply this approach for similarity analysis of protein sequences on two data sets. The results all show that our method is effective.

2 Graphical Representation of Protein Sequences

Much effort has been made by considering minimalist models with a few types of amino acid residues to simplify the natural set of residues of 20 types for better physical understanding and practical purposes [40–45]. In these models, the compositions are much simpler than the real ones. In the following, we put forward a novel 3D graphical representation of proteins based on the five-letter model of 20 amino acids.

Based on the method introduced by Li [41], the 20 amino acids can be classified into five groups:

$Group1 = \{C, M, F, I, L, V, W, Y\}$

$Group2 = \{A, T, H\}$

$Group3 = \{G, P\}$

$Group4 = \{D, E\}$

$Group5 = \{S, N, Q, R, K\}$

We choose a representative letter in each group, which are I, A, G, E, and K, respectively. Thus a protein primary sequence can be reduced into a five-letter sequence by substituting each letter with its representative letter. For example, the five-letter sequence of MVHLTPEEKSAVTALWGKVNVDEVGGEALGR, which is the first 31 amino acid residues of the gorilla $\beta-$globin protein, is IIAIAGEEKKAIAAIIGKIKIEEIGGEAIGK.

In the following, we will construct a graphical representation of a protein sequence. By using two mappings φ and ϕ, we map the five representative letters and their pairs to the points on the underside of a right cone, whose coordinates are given as follows

$$\varphi(X_i) = (\cos(2i\pi/5), \sin(2i\pi/5), 1), \qquad i = 1, 2, 3, 4, 5 \tag{1}$$

$$\phi(X_iX_j) = \varphi(X_i) + \frac{1}{4}(\varphi(X_j) - \varphi(X_i)), \qquad i, j = 1, 2, 3, 4, 5 \tag{2}$$

where X_i is one of the five representative letters I, A, G, E, and K and X_iX_j is one of the twenty-five letter pairs II, IA, IG,... and KK.

Given a five-letter sequence $S = S_1, S_2, ..., S_n$, we start at the origin and inspect it by stepping one element at a time. For step i, the letter S_i is mapped to a point $P_i(x_i, y_i, z_i)$ in the 3D space by the following mapping

$$\psi(S_i) = \psi(S_{i-1}) + \varphi(S_i) + \sum_{X,Y \in \{I,A,G,E,K\}} f_{XY} \cdot \phi(XY) \tag{3}$$

where $\psi(S_0) = (0, 0, 0)$, f_{XY} is the cumulative frequency of the letter pair XY in the subsequence from the first letter to the i-th letter in the sequence. When i runs from 1 to n, we obtain points $P_1, P_2, ..., P_n$. Connecting adjacent points, we can obtain a graphical curve in 3D space for each protein sequence. And we call the curve 3D-PAF curve of the protein sequence.

3 Numerical Characterization of Protein Sequence

In this section, we give a numerical characterization of the 3D-PAF curve that will facilitate quantitative comparisons of protein sequences. One of the possibilities to achieve this goal is to characterize the graphical curves by invariants. In order to find some invariants which are sensitive to the form of the graphical curve, the graphical curve of protein can be transformed into another mathematical object, a matrix. One of the matrices which meet this condition is the L/L matrix, in which each off-diagonal element is defined as a quotient of the Euclidean distance between two vertices of the graphical curve and the sum of geometrical lengths of edges between the same pair of vertices measured along the graphical curve and all diagonal elements are equal to zero. Once a real symmetric matrix is given, one often uses some of matrix invariants as descriptors of the sequence. Therefore, the comparison of sequences is converted into a numerical comparison of vectors instead of letters comparison. Here, we use the absolute values of the first eight leading eigenvalues of L/L matrix to characterize the corresponding 3D-PAF curve. In order to eliminate the influence of length of sequence, we normalize the eigenvalue by dividing it by the length of the protein sequence. That is, we take the vector $(|\frac{\lambda_1}{N}|, |\frac{\lambda_2}{N}|, ..., |\frac{\lambda_8}{N}|)$ as the numerical characterization of the 3D-PAF curve, where λ_i is the *i-th* leading eigenvalue($i = 1, 2, ..., 8$) of the L/L matrix and N is the length of the protein sequence.

In our model, the five representative letters are mapped on the circumference of the underside of a right cone. Each arrangment of the five letters on the circumference corresponds to a kind of 3D-PAF curve of protein sequence. The number of circular permutations of the five letters is 4!=24. Among the 24 kinds of 3D-PAF curves, two symmetric curves have the same L/L matrices, so we only use the 12 kinds of intrinsically different 3D-PAF curves to represent each protein sequence. By combining all numerical characterizations of 12 3D-PAF curves, a protein primary sequence can be characterized by a 96-dimensional feature vector.

Given a data set consisting of N protein sequences, we can obtain a $N \times 96$ matrix, each row of which corresponds to a protein sequence. Since the values of different columns are on completely different scales, we take standardized Euclidean distance between row vectors as the similarity measure between the corresponding protein sequences. The smaller the standardized Euclidean distance between the two row vectors is, the more similar are the two corresponding protein sequences.

4 Results and Discussion

4.1 The similarity analysis of nine ND5 proteins

To illustrate our method, we compare the similarities of the ND5 protein sequences across nine species listed in Table 1. As mentioned in Section 3, we compute the feature vectors of the nine ND5 protein sequences. Then the distance matrix for the nine ND5 proteins is constructed by using standardized Euclidean distance and is shown in Table 2.

From Table 2, we find a fact that the ND5 proteins of human, gorilla, pigmy chimpanzee and common chimpanzee are more similar to each other, also the proteins of fin whale and blue whale are very similar to each other, and so do the pair of mouse and rat. On the other hand, the protein of opossum is quite dissimilar to all other species. Also, we can see that the entries of human−pigmy chimpanzee and human−common chimpanzee are smaller than the entry of human−gorilla. That is to say, the ND5 protein of human is more similar to that of common chimpanzee and pigmy chimpanzee than that of gorilla.

Table 1: The information for nine ND5 protein sequences

NO.	Species	ID(NCBI)	length
1	Human (Homo sapiens)	AP_000649	603
2	Gorilla (Gorilla gorilla)	NP_008222	603
3	Common chimpanzee (Pan troglodytes)	NP_008196	603
4	Pigmy chimpanzee (Pan paniscus)	NP_008209	603
5	Fin whale (Balenoptera physalus)	NP_006899	606
6	Blue whale (Balenoptera musculus)	NP_007066	606
7	Rat (Rattus norvegicus)	AP_004902	610
8	Mouse (Mus musculus)	NP_904338	607
9	Opossum (Didelphis virginiana)	NP_007105	602

We believe that the results are not coming by accident since they are consistent with the known fact of evolution. ClustalW is one of the most multiple sequence alignment method. To compare our method with ClustalW, we list the results of multiple sequence alignment among the nine species by using ClustalW under MEGA6.0 software, see the distance matrix in Table 3. Observing Table 2 and Table 3, we can see that the sequence similarity results are almost consistent in both our method and ClustalW. The phylogenetic trees constructed based on our method and CLUSTALW respectively in Fig. 1 show same results.

Table 2: The distance matrix for the nine ND5 protein sequences calculated by our method

	Human	Gorilla	C.Chim.	P.Chim.	F.Whale	B.Whale	Rat	Mouse	Opossum
Human	0	8.2096	7.8061	6.9508	11.9203	13.3220	15.7248	13.5202	16.4476
Gorilla		0	9.2376	8.2985	13.2242	14.0953	17.8533	14.4987	16.7827
C.Chim.			0	6.1237	12.3322	13.8987	16.7306	14.4406	19.2442
P.Chim.				0	11.3038	13.0564	16.1126	13.6451	18.0387
F.Whale					0	7.2549	14.9376	12.9769	16.0624
B.Whale						0	16.3437	13.1873	15.5091
Rat							0	12.9581	17.3784
Mouse								0	14.5673
Opossum									0

In addition, we calculate the correlation coefficients between our results and ClustalW. The correlation coefficient between the first row of Tables 2 and 3 is 0.9286. The first rows in both matrices are relative to human protein, the second ones to gorilla and so on. The correlation coefficients for the rows relative to all nine species are listed in the first column of Table 4. Analogously, the correlation coefficients between the results of

Table 3: The distance matrix for the nine ND5 protein sequences calculated by ClustalW

	Human	Gorilla	C.Chim.	P.Chim.	F.Whale	B.Whale	Rat	Mouse	Opossum
Human	0	0.104	0.067	0.069	0.375	0.377	0.456	0.443	0.464
Gorilla		0	0.096	0.093	0.390	0.387	0.469	0.453	0.494
C.Chim.			0	0.048	0.370	0.370	0.461	0.448	0.472
P.Chim.				0	0.368	0.368	0.453	0.443	0.459
F.Whale					0	0.034	0.410	0.422	0.486
B.Whale						0	0.407	0.415	0.486
Rat							0	0.241	0.494
Mouse								0	0.469
Opossum									0

Table 4: The correlation coefficients for nine ND5 proteins of our method and the methods in Ref. [27–29, 31, 32, 38, 42], as compared with clustalW method

	our method	Ref. [27] (Table4)	Ref. [28] (Table3)	Ref. [29] (Table3)	Ref. [31] (Table1)	Ref. [32] (Table4)	Ref. [38] (Table3)	Ref. [42] (Table4)
Human	0.9286	0.9380	0.9268	0.9620	0.8887	0.9497	0.9612	0.8940
Gorilla	0.9275	0.9276	0.9086	0.9524	0.9293	0.9570	0.9698	0.8461
C.Chim.	0.9273	0.9357	0.9060	0.9692	0.9470	0.9542	0.9681	0.8552
P.Chim.	0.9292	0.9323	0.7647	0.9644	0.9132	0.9421	0.9650	0.7691
F.Whale	0.9299	0.8853	0.5180	0.9653	0.9163	0.9817	0.9583	0.9280
B.Whale	0.9325	0.8862	0.5290	0.9657	0.9154	0.9833	0.9576	0.8749
Rat	0.9635	0.8687	0.6903	0.9542	0.9255	0.9884	0.9539	0.8979
Mouse	0.9269	0.8449	0.6305	0.9559	0.9251	0.7493	0.9296	0.8681
Opossum	0.9651	0.9961	0.6645	0.9986	0.8599	0.6649	0.9985	0.7556

Refs. [27–29,31,32,38,42] and ClustalW are also calculated in order to show the advantages of our method, see Table 4. We find that our method has higher correlation coefficients with ClustalW than other methods except the ones in Refs. [29, 38]. But, we will show the robustness of methods in Refs. [29, 38] are much less than ours. To illustrate this point, we add a sequence to the ND5 data set which is built by subtracting the first amino acid from the ND5 sequence of pigmy chimpanzee and is denoted as P.chim0. The phylogenetic trees of the new data set shown in Fig. 2 are constructed by our method and Ref. [29, 38]'s methods, respectively. P.chim and P.chim0 are clustered together in Fig. 2(a), but P.chim0 is very dissimilar to other species in Fig. 2(b) and Fig. 2(c) which is obviously unreasonable.

Incorporating accumulative frequencies of adjacent amino acids into the graphical representation is one of the characteristics of our method. In order to illustrate its effect on the graphical representation, in Fig. 3, we show the phylogenetic tree of the nine ND5 proteins constructed by our method without considering the accumulative frequencies of

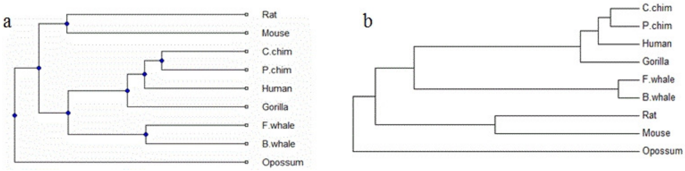

Figure 1: Phylogenetic trees of the nine ND5 proeins constructed by (a) our method and (b) CLUSTALW

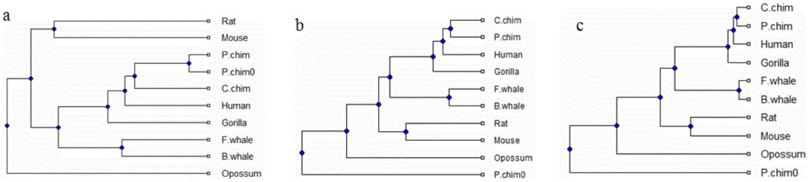

Figure 2: Phylogenetic trees of the modified ND5 proteins constructed by (a) our method, (b) Ref. [29] and (c) Ref. [38]

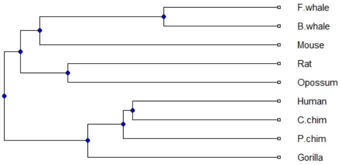

Figure 3: phylogenetic tree of the nine ND5 proteins constructed by our method without considering the accumulative frequencies of adjacent amino acids

adjacent amino acids. That is, let f_{XY} in the equation (3) equal to be zero in the process of construction of graphical curve. Comparing Fig. 3 and Fig. 1, we can easily find that the results in Fig. 3 are inconsistent with the known fact of evolution. Thus, incorporating accumulative frequencies of adjacent amino acids into the graphical representation can reflect more information of the protein and improve its evolutionary study.

4.2 The similarity analysis of 35 coronavirus spike proteins

The coronaviruses (order Nidovirales, family Coronaviridae, genus Coronavirus) are members of a family of large, enveloped, positive-sense single-stranded RNA viruses that replicate in the cytoplasm of animal host cells. Generally, coronaviruses can be divided into three groups: the first group and the second group come from mammalian; the third group comes from poultry (chicken and turkey). A novel coronavirus has been identified as the cause of the outbreak of severe acute respiratory syndrome (SARS). Previous phylogenetic analyses based on sequence alignments show that SARS-CoV belongs to a group distantly related to known group II coronaviruses [49–52]. The spike protein, which is common to all known coronaviruses, is crucial for viral attachment and entry into the host cell. In order to further verify the validity of our method, we perform similarity analysis among the 35 spike protein sequences from coronavirus, which has been studied by different methods [44,45]. Taxonomic information and accession numbers are provided in Table 5.

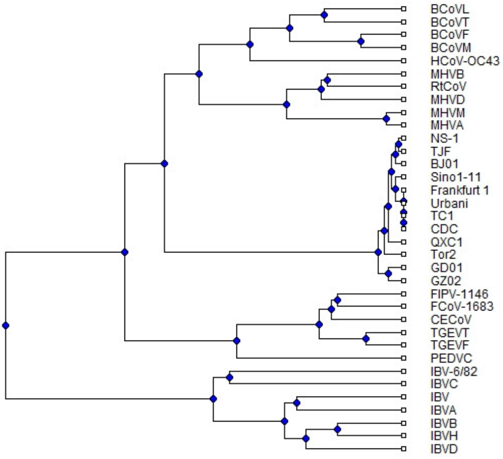

Figure 4: Phylogenetic tree of the 35 spike proteins constructed by our method

The phylogenetic tree of the 35 coronavirus spike proteins, shown in Fig. 4, is built based on our method. Observing Fig. 4, we find that the 35 coronavirus spike proteins can be classified into four groups on the whole. The SARS-CoVs appear to cluster together and form a separate branch, which can be distinguished easily from other three groups of coronaviruses. The coronaviruses belonging to group I (FIPV-1146, FCoV-1683, PEDVC, TGEVT, TGEVF, CECoV), group II (MHVM, MHVB, MHVA, MHVD, RtCoV, BCoVF, BCoVM, BCoVL, BCoVT, HCoV-OC43) and Group III (IBV, IBV-6/82, IBVD, IBVC, IBVA, IBVB, IBVH) can also be clustered together into three different branches, respec-

Table 5: The information of 35 coronavirus spike proteins

No.	ID(NCBI)	Abbreviation	Name	Group
1	P10033	FIPV-1146	Feline infectious peritonitis virus strain 79-1146	I
2	Q66928	FCoV-1683	Feline coronavirus strain 79-1683	I
3	Q91AV1	PEDVC	Porcine epidemic diarrhea virus strain CV777	I
4	Q9DY22	TGEVT	Transmissible gastroenteritis virus strain TO14	I
5	P18450	TGEVF	Porcine transmissible gastroenteritis coronavirus strain FS772/70	I
6	P36300	CECoV	Canine enteric coronavirus strain INSAVC-1	I
7	Q9J3E7	MHVM	Murine hepatitis virus strain ML-10	II
8	Q83331	MHVB	Murine hepatitis virus strain Berkeley	II
9	P11224	MHVA	Murine hepatitis virus strain A59	II
10	O55253	MHVD	Murine hepatitis virus strain DVIM	II
11	Q9IKD1	RtCoV	Rat coronavirus strain 681	II
12	P25190	BCoVF	Bovine coronavirus strain F15	II
13	P15777	BCoVM	Bovine coronavirus strain Mebus	II
14	Q9QAR5	BCoVL	Bovine coronavirus strain LSU-94LSS-051	II
15	Q91A26	BCoVT	Bovine enteric coronavirus 98TXSF-110-ENT	II
16	P36334	HCoV-OC43	Human coronavirus strain OC43	II
17	Q82666	IBV	Infectious bronchitis virus	III
18	P05135	IBV-6/82	Avian infectious bronchitis virus strain 6/82	III
19	P12722	IBVD	Avian infectious bronchitis virus strain D274	III
20	Q64930	IBVC	Infectious bronchitis virus strain CU-T2	III
21	Q82624	IBVA	Infectious bronchitis virus strain Ark99	III
22	P11223	IBVB	Avian infectious bronchitis virus strain Beaudette	III
23	Q98Y27	IBVH	Infectious bronchitis virus strain H52	III
24	AAP41037	Tor2	SARS coronavirus Tor2	IV
25	AAP30030	BJ01	SARS coronavirus BJ01	IV
26	AAR91586	NS-1	SARS coronavirus NS-1	IV
27	AAP51227	GD01	SARS coronavirus GD01	IV
28	AAP33697	Frankfurt 1	SARS coronavirus Frankfurt 1	IV
29	AAP13441	Urbani	SARS coronavirus Urbani	IV
30	AAQ01597	TC1	SARS coronavirus Taiwan TC1	IV
31	AAU81608	CDC	SARS Coronavirus CDC #200301157	IV
32	AAS00003	GZ02	SARS coronavirus GZ02	IV
33	AAR86788	QXC1	SARS coronavirus ShanghaiQXC1	IV
34	AAR23250	Sino1-11	SARS coronavirus Sino1-11	IV
35	AAT76147	TJF	SARS coronavirus TJF	IV

tively. The topology of the phylogenetic tree obtained by our method is quite consistent with the results obtained by other authors [41, 48–52].

Through further observation on the subtrees of the first and third branches, we can see that MHV, BCoV and HCoV-OC43 in the first branch are separated clearly and so do FCoV, CECoV and TGEV in the third branch. However, they are not clearly distinguished by Deng's and Li's methods [44, 45]. In addition, we can find that HCoV-OC43 is most closely related to BCoV. The same result is obtained by other authors based on sequence alignment [49–52].

Table 6: The distance matrix between SARS-CoVs and other three groups of coronavirus

Species	Tor2	BJ01	NS-1	GD01	Frankfurt 1	Urbani	TC1	CDC	GZ02	QXC1	Sino1-11	TJF
FIPV-1146	14.811	14.793	14.760	14.885	14.770	14.770	14.770	14.770	14.997	14.742	14.860	14.785
FCoV-1683	13.630	13.648	13.615	13.708	13.603	13.603	13.603	13.603	13.814	13.627	13.695	13.643
PEDVC	11.858	11.928	11.866	11.967	11.812	11.812	11.812	11.812	11.993	11.836	11.926	11.899
TGEVT	14.337	14.308	14.246	14.426	14.236	14.236	14.236	14.236	14.586	14.216	14.340	14.274
TGEVF	14.863	14.800	14.743	14.943	14.755	14.755	14.755	14.755	15.112	14.705	14.853	14.769
CECoV	14.565	14.546	14.502	14.586	14.494	14.494	14.494	14.494	14.699	14.487	14.592	14.534
MHVM	8.678	8.480	8.418	8.653	8.414	8.414	8.414	8.414	8.749	8.417	8.443	8.425
MHVB	11.584	11.456	11.424	11.532	11.356	11.358	11.358	11.358	11.629	11.416	11.383	11.433
MHVA	8.775	8.540	8.478	8.757	8.480	8.480	8.480	8.480	8.871	8.477	8.505	8.486
MHVD	10.483	10.355	10.285	10.520	10.259	10.259	10.259	10.259	10.628	10.270	10.310	10.298
RtCoV	10.836	10.752	10.705	10.827	10.634	10.634	10.634	10.634	10.925	10.692	10.669	10.722
BCoVF	12.722	12.763	12.688	12.674	12.610	12.610	12.610	12.610	12.753	12.600	12.706	12.749
BCoVM	13.339	13.407	13.324	13.322	13.230	13.230	13.230	13.230	13.392	13.219	13.337	13.387
BCoVL	12.394	12.532	12.486	12.405	12.370	12.370	12.370	12.370	12.421	12.400	12.444	12.536
BCoVT	13.342	13.673	13.634	13.440	13.458	13.458	13.458	13.458	13.402	13.534	13.537	13.674
HCoV-OC43	10.852	10.925	10.874	10.872	10.739	10.739	10.739	10.739	10.871	10.790	10.796	10.926
IBV	14.390	14.595	14.601	14.368	14.447	14.447	14.447	14.447	14.259	14.512	14.468	14.592
IBV-6/82	20.398	20.706	20.732	20.368	20.583	20.583	20.583	20.583	20.263	20.684	20.580	20.718
IBVD	15.451	15.710	15.731	15.444	15.583	15.583	15.583	15.583	15.345	15.662	15.595	15.715
IBVC	18.873	19.172	19.153	18.886	19.030	19.030	19.030	19.030	18.793	19.110	19.058	19.155
IBVA	13.367	13.514	13.502	13.336	13.428	13.428	13.428	13.428	13.225	13.462	13.438	13.503
IBVB	15.292	15.475	15.471	15.253	15.330	15.330	15.330	15.330	15.222	15.358	15.352	15.471
IBVH	15.576	15.774	15.773	15.491	15.645	15.645	15.645	15.645	15.452	15.663	15.684	15.779

The similarity distances between SARS-CoVs and other three groups of coronavirus are listed in Table 6. As we can see from the Table 6, SARS-CoVs are more closely related to group II coronaviruses than to group I and III coronaviruses. This result is consistent with those reported in the literatures [49–52], in which SARS-CoV is considered to be a subgroup of the group II. Furthermore, it is obvious from the Table 6 that SARS-CoV is closely related to MHV and RtCoV, which is consistent with the result reported in the paper [52].

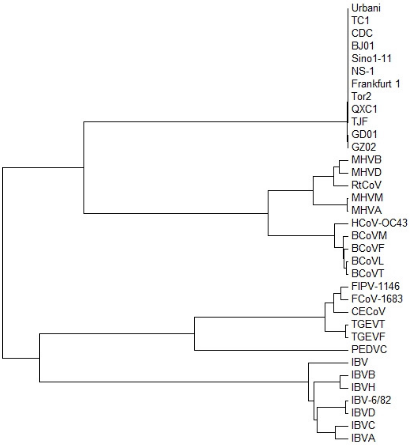

Figure 5: Phylogenetic tree of the 35 spike proteins constructed by CLUSTALW

In Fig. 5, we also construct the phylogenetic tree for the 35 coronavirus spike proteins by CLUSTALW. Observing Fig. 4 and Fig. 5, we can find that our result is very similar to that of CLUSTALW.

5 Conclusion

In this paper, we proposed a novel 3D graphical representation of protein sequences. This approach takes into consideration not only the physicochemical characteristics of amino acids but also the accumulative frequencies of adjacent amino acids that can reflect more information of protein sequence and improve evolutionary study. The method has been applied for similarity analysis of protein sequences on two data sets: nine ND5 proteins and 35 coronavirus spike proteins. The obtained results are quite consistent with the known fact of evolution and show that our method is valid.

Acknowledgments: This work is supported by Natural Science Foundation of China (11401346), Shandong Natural Science Foundation (ZR2015AM017 and ZR2014FM034), and China Postdoctoral Science Foundation (2014M561910).

References

[1] J. D. Thompson, 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.

[2] J. Pei, Multiple protein sequence alignment, *Curr. Opin. Struct. Biol.* **18** (2008) 382–386.

[3] H. Li, N. Homer, A survey of sequence alignment algorithms for next–generation sequencing, *Brief. Bioinf.* **11** (2010) 473–483.

[4] E. Hamori, J. Ruskin, *H* curves, a novel method of representation of nucleotide series especially suited for long DNA sequences, *J. Biol. Chem.* **258** (1983) 1318–1327.

[5] H. L. Jeffrey, Chaos game representation of gene structure, *Nuleic Acids Res.* **18** (1990) 2163–2170.

[6] H. Wang, Y. Zhang, A new approach to molecular phylogeny of H5N1 avian influenza viruses in Asia, *Int. J. Quantum Chem.* **110** (2009) 1964–1971.

[7] Y. Zhang, B. Liao, K. Ding, On 3DD-curves of DNA sequences, *Mol. Simul.* **32** (2006) 29–34.

[8] Y. Zhang, W. Chen, Invariants of DNA sequences based on 2DD-curves, *J. Theor. Biol.* **242** (2006) 382–388.

[9] Y. Zhang, W. Chen, New invariant of DNA sequences, *MATCH Commun. Math. Comput. Chem.* **58** (2007) 197–208.

[10] Y. Zhang, W. Chen, Analysis of similarity/dissimilarity of long DNA sequences based on three 2DD-curves, *Comb. Chem. High Throughput Screen.* **10** (2007) 231–237.

[11] Y. Zhang, M. Tan, Visualization of DNA sequences based on 3DD-Curves, *J. Math. Chem.* **44** (2008) 206–216.

[12] Y. Zhang, W. Chen, A new approach to molecular phylogeny of primate mitochondrial DNA, *MATCH Commun. Math. Comput. Chem.* **59** (2008) 625–634.

[13] C. Yu, M. Deng, S. S.-T. Yau, DNA sequence comparison by a novel probabilistic method, *Inf. Sci.* **181** (2011) 1484–1492.

[14] M. Randić, A. T. Balaban, On a four–dimensional representation of DNA primary sequences, *J. Chem. Inf. Comput. Sci.* **43** (2003) 532–539.

[15] S. Zou, L. Wang, J. Wang, A 2D graphical representation of the sequences of DNA based on triplets and its application, *EURASIP J. Bioinf. Sys. Biol.* **2014** (2014) #1.

[16] Z. H. Qi, T. R. Fan, PN-curve: A 3D graphical representation of DNA sequences and their numerical characterization, *Chem. Phys. Lett.* **442** (2007) 434–440.

[17] J. F. Yu, X. Sun, J. H. Wang, TN curve: A novel 3D graphical representation of DNA sequence based on trinucleotides and its applications, *J. Theor. Biol.* **261** (2009) 459–468.

[18] Y. Wu, A. W. C. Liew, H. Yan, M. Yang, DB-Curve: a novel 2D method of DNA sequence visualization and representation, *Chem. Phys. Lett.* **367** (2003) 170–176.

[19] G. Xie, Z. Mo, Three 3D graphical representations of DNA primary sequences based on the classifications of DNA bases and their applications, *J. Theor. Biol.* **269** (2011) 123–130.

[20] P. Waż, D. Bielińska–Waż, 3D-dynamic representation of DNA sequences, *J. Mol. Model.* **20** (2014) #2141.

[21] M. Randić, J. Zupan, Highly compact 2D graphical representation of DNA sequences, *SAR QSAR Environ. Res.* **15** (2004) 191–205.

[22] N. Jafarzadeh, A. Iranmanesh, C-curve: A novel 3D graphical representation of DNA sequence based on codons, *Math. Biosci.* **241** (2013) 217–224.

[23] B. Liao, T. Wang, 3-D graphical representation of DNA sequences and their numerical characterization, *J. Mol. Struct.* **681** (2004) 209–212.

[24] J. Song, H. Tang, A new 2-D graphical representation of DNA sequences and their numerical characterization, *J. Biochem. Bioph. Meth.* **63** (2005) 228–239.

[25] P. He, X. Li, J. Yang, J. Wang, A novel descriptor for protein similarity analysis, *MATCH Commun. Math. Comput. Chem.* **65** (2011) 445–458.

[26] J. F. Yu, X. Sun, J. H. WANG, A novel 2D graphical representation of protein sequence based on individual amino acid, *Int. J. Quantum Chem.* **111** (2011) 2835–2843.

[27] Y. Liu, D. Li, K. Lu, Y. Jiao, P. He, $P-H$ Curve, a graphical representation of protein sequences for similarities analysis, *MATCH Commun. Math. Comput. Chem.* **70** (2013) 451–466.

[28] Z. C. Wu, X. Xiao, K. C. Chou, $2D - MH$: A web-server for generating graphic representation of protein sequences based on the physicochemical properties of their constituent amino acids, *J. Theor. Biol.* **267** (2010) 29–34.

[29] T. Ma, Y. Liu, Q. Dai, Y. Yao, P. He, A graphical representation of protein based on a novel iterated function system, *Physica A* **403** (2014) 21–28.

[30] J. Wen, Y. Y. Zhang, A 2D graphical representation of protein sequence and its numerical characterization, *Chem. Phys. Lett.* **476** (2009) 281–286.

[31] G. Huang, J. Hu, Similarity/dissimilarity analysis of protein sequences by a new graphical representation, *Curr. Bioinf.* **8** (2013) 539–544.

[32] Z. Li, C. Geng, P. He, Y. Yao, A novel method of 3D graphical representation and similarity analysis for proteins, *MATCH Commun. Math. Comput. Chem.* **71** (2014) 213–226.

[33] M. I. A. el Maaty, M. M. Abo–Elkhier, M. A. A. Elwahaab, 3D graphical representation of protein sequences and their statistical characterization, *Physica A* **389** (2010) 4668–4676.

[34] M. K. Gupta, R. Niyogi, M. Misra, A 2D graphical representation of protein sequence and their similarity analysis with probabilistic method, *MATCH Commun. Math. Comput. Chem.* **72** (2014) 519–532.

[35] F. Bai, T. Wang, On graphical and numerical representation of protein sequences, *J. Biomol. Struct. Dyn.* **23** (2006) 537–545.

[36] M. I. A. el Maaty, M. M. Abo–Elkhier, M. A. A. Elwahaab, Representation of protein sequences on latitude-like circles and longitude–like semi-circles, *Chem. Phys. Lett.* **493** (2010) 386–391.

[37] M. M. Abo–Elkhier, Similarity/dissimilarity analysis of protein sequences using the spatial median as a descriptor, *J. Biophys. Chem.* **3** (2012) 142–148.

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

[39] M. Randić, D. Butina, J. Zupan, Novel 2-D graphical representation of proteins, *Chem. Phys. Lett.* **419** (2006) 528–532.

[40] Y. Liu, Y. Zhang, A new method for analyzing H5N1 avian influenza virus, *J. Math. Chem.* **47** (2010) 1129–1144.

[41] C. Li, L. Xing, X. Wang, 2-D graphical representation of protein sequences and its application to coronavirus phylogeny, *BMB Rep.* **41** (2008) 217–222.

[42] Y. Yao, S. Yan, J. Han, Q. Dai, P. He, A novel descriptor of protein sequences and its application, *J. Theor. Biol.* **347** (2014) 109–117.

[43] B. Liao, B. Liao, X. Lu, Z. Cao, A novel graphical representation of protein sequences and its application, *J. Comput. Chem.* **32** (2011) 2539–2544.

[44] W. Deng, Y. Luan, DV-curve representation of protein sequences and its application, *Comput. Math. Meth. Med.* **2014** (2014) #203871.

[45] D. Li, J. Wang, C. Li, New 3-D graphical representation of protein sequences and its application, *Chin. J. Bioinf.* **7** (2009) 60–63.

[46] M. Li, J. H. Badger, X. Chen, S. Kwong, P. Kearney, H. Zhang, An information–based sequence distance and its application to whole mitochondrial genome phylogeny, *Bioinf.* **17** (2001) 149–154.

[47] H. H. Out, K. Sayood, A new sequence distance measure for phylogenetic tree construction, *Bioinf.* **19** (2003) 2122–2130.

[48] J. Wen, C. Li, Similarity Analysis of DNA Sequences based on the LZ Complexity, *Internet El. J. Mol. Des.* **6** (2007) 1–12.

[49] E. J. Snijder, P. J. Bredenbeek, Unique and conserved features of genome and proteome of SARS–coronavirus, an early split–off from the coronavirus group 2 lineage, *J. Mol. Biol.* **331** (2003) 991–1004.

[50] S. K. P. Lau, P. C. Y. Woo, Severe acute respiratory syndrome coronavirus–like virus in Chinese horseshoe bats, *Proc. Natl. Acad. Sci. U. S. A.* **102** (2005) 14040–14045.

[51] Z. Shi, Z. Hu, A review of studies on animal reservoirs of the SARS coronavirus, *Virus Res.* **133** (2008) 74–87.

[52] P. Liò, N. Goldman, Phylogenomics and bioinformatics of SARS–CoV, *Trends Microbiol.* **12** (2004) 106–111.