Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Genomics 109 (2017) 123–130

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Methods Paper

A new method to analyze protein sequence similarity using Dynamic Time Warping

Wenbing Hou [a], Qiuhui Pan [b,a], Qianying Peng [c], Mingfeng He [a,*]

[a] School of Mathematical Sciences, Dalian University of Technology, Dalian 116024, PR China
[b] School of Innovation and Entrepreneurship, Dalian University of Technology, Dalian 116024, PR China
[c] Department of Academics, Dalian Naval Academy, Dalian 116001, PR China

ARTICLE INFO

Article history:
Received 12 August 2016
Received in revised form 6 December 2016
Accepted 10 December 2016
Available online 11 December 2016

Keywords:
Protein sequences similarity analysis
Discrete Fourier Transform
Dynamic Time Warping
Phylogenetic tree

ABSTRACT

Sequences similarity analysis is one of the major topics in bioinformatics. It helps researchers to reveal evolution relationships of different species. In this paper, we outline a new method to analyze the similarity of proteins by Discrete Fourier Transform (DFT) and Dynamic Time Warping (DTW). The original symbol sequences are converted to numerical sequences according to their physico-chemical properties. We obtain the power spectra of sequences from DFT and extend the spectra to the same length to calculate the distance between different sequences by DTW. Our method is tested in different datasets and the results are compared with that of other software algorithms. In the comparison we find our scheme could amend some wrong classifications appear in other software. The comparison shows our approach is reasonable and effective.

1. Introduction

With the advance of sequencing techniques, the database of DNA, RNA and protein has been enlarged rapidly, promoting the development of bioinformatics effectively. It has been increasingly important to develop efficient ways to obtain the information hidden in the gene data. In the last few decades, several methods to classify the genes have been proposed. In 1983, Hamori and Ruskin proposed a visible 3-D curve with the name of H-curve to tell the relations between different DNAs [1]. As the first graphical representation, it motivates other researchers in the following years to develop more graphical representations of DNA sequences including 2D, 3D and even multidimensional representations [2–14]. Besides the graphical representations, researchers try to combine some techniques from other disciplines into the study of genes and have proposed novel methods. For example, the Discrete Fourier Transform, which is broadly applied in signal process, has been introduced into the process of genes [15,16]. It is proved effective in the analysis of DNA sequences.

Methods for similarity analysis of proteins also have been proposed recently. Considering a protein sequence consists of 20 kinds of different amino acids while a DNA sequence only consists of four bases, it is much more complex to express a protein than a DNA sequence. However, there are some methods which are generalized from the ways of analyzing the DNA sequences [17–21]. Yau et al. propose a method with the name of protein map [22] following their previous work. They use the moment vectors to represent proteins and generate a universal protein map [23]. Motivated by the protein map, they also develop a novel method, with the name of protein space, to realize the nature of protein universe [24]. Their method is applied successfully in their following papers and proved effective [25,26]. He et al. present a new way of generalized Chaos Game Representation (CGR) method to outline a dynamic 3D graphical representation [27] which is analogous to the original CGR method proposed by Jeffrey for graphical representation of DNA [3]. El-Lakkani and Mahran introduce a two dimensional graphical representation of protein sequences. They propose a new mathematical descriptor in their paper to measure the similarity of two protein sequences [28]. Li et al. present a graphical representation with the name of UC-Curve [29]. The amino acids are assigned to the circumference of a unit circle with a cyclic order. Geometric center vectors of UC-Curves and Euclidean distances are extracted to analyze pairwise similarities. Moreover, techniques from other disciplines have been applied in the analysis of proteins successfully. Wąż and Bielińska-Wąż introduce the moments of inertia as new descriptors in the calculation of similarities [30,31]. Based on their works, Czerniecka et al. propose a 20-D dynamic representation of protein sequences [32] and the scheme is proved reasonable.

In this paper, we outline a new method based on Discrete Fourier Transform (DFT) and Dynamic Time Warping (DTW) to calculate the similarities of proteins. The original symbol sequences are converted to numerical sequences according to their physico-chemical properties and the similarities are calculated based on DFT and DTW. We test our

* Corresponding author.
E-mail address: mfhe@dlut.edu.cn (M. He).

http://dx.doi.org/10.1016/j.ygeno.2016.12.002
0888-7543/

scheme with different datasets and compare our results with some existing softwares. It is demonstrated that the consequences from our test are in agreement with evolutionary relation satisfactorily.

2. Models and methods

2.1. Numerical representation of protein sequence

Amino acids are considered as the basic component of proteins. The study of proteins always starts with the study of amino acids. The physico-chemical properties of amino acids are considered to have immense effects on the properties of proteins [33]. It is an effective way to study the similarity of proteins by the properties of amino acids. In our work, we choose two main amino acids properties, namely hydropathy and isoelectric point, to construct a new way to represent the protein sequences. The detail values are declared in Table 1. All values are cited from reference [23].

According to the hydropathy value of amino acids, we could list their ranks: I > V > L > F > C > M > A > G > T > S > W > Y > P > H > D > N > E > Q > K > R. A radian θ_i will be assigned to each of the amino acid according to the hydropathy value rank. Based on the ranks, the value of θ_i will change from 0 to 2π at the interval of $\frac{1}{20}\pi$. For example, the radian 0 will be assigned to the amino acid I and $\frac{6}{20}\pi$ will be assigned to the amino acid A. Similarly, we also list another rank based on the isoelectric point values: R > K > H > P > T > I > A > L > G > V > W > M > S > Y > Q > F > N > C > E > D. Another radian φ_i, ranging from 0 to 2π at the interval of $\frac{1}{20}\pi$, will be assigned to amino acids according to the isoelectric point values. Then we build a three-dimensional representation of the amino acids. The coordinates of amino acids are calculated as follows:

$$x_i = \sin(\theta_i)\cos(\varphi_i),\quad y_i = \sin(\theta_i)\sin(\varphi_i),\quad z_i = \cos(\theta_i),\quad i = 1,2,\ldots 20 \quad (1)$$

Now, an amino acids sequence $S = s_1 s_2 s_3 \ldots s_N$ with the length of N could be represented by a new sequence $F = \{c_1, c_2, c_3, \ldots c_N\}$, where $c_i = (x_i, y_i, z_i)$. The coordinates of different axis are extracted respectively, forming new sequences

$$u_1(n) = \{u_1(0), u_1(1), \ldots u_1(N-1)\} = \{x_1, x_2, \ldots x_N\},$$
$$u_2(n) = \{u_2(0), u_2(1), \ldots u_2(N-1)\} = \{y_1, y_2, \ldots y_N\},$$
$$u_3(n) = \{u_3(0), u_3(1), \ldots u_3(N-1)\} = \{z_1, z_2, \ldots z_N\},$$

Table 1
Hydropathy and isoelectric point values of 20 amino acids.

Amino acid	Abbreviation	Hydropathy	Isoelectric point
Isoleucine	I	4.5	6.02
Valine	V	4.2	5.96
Leucine	L	3.8	5.98
Phenylalanine	F	2.8	5.48
Cysteine	C	2.5	5.07
Methionine	M	1.9	5.74
Alanine	A	1.8	6.00
Glycine	G	−0.4	5.97
Threonine	T	−0.7	6.16
Serine	S	−0.8	5.68
Tryptophan	W	−0.9	5.89
Tyrosine	Y	−1.3	5.66
Proline	P	−1.6	6.30
Histidine	H	−3.2	7.59
Aspartic acid	D	−3.5	2.77
Asparagine	N	−3.5	5.41
Glutamic acid	E	−3.5	3.22
Glutamine	Q	−3.5	5.65
Lysine	K	−3.9	9.74
Arginine	R	−4.5	10.76

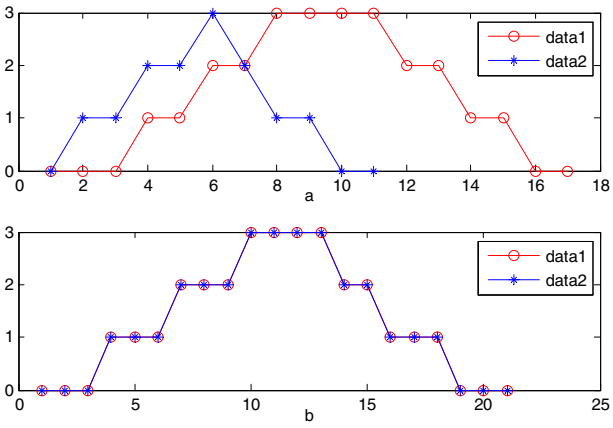

Fig. 1. Original sequences and sequences after warping.

It is obvious that every symbol sequence will be represented by three numerical sequences according to our method. The representation will be unique because every amino acid has a unique coordinate, which means our approach could avoid the confusion from similar proteins.

2.2. Discrete Fourier Transform

The DFT is a common way in signal processing which is used to transform the signals in time domain into frequency domain. The latent information hidden in the signal in time domain could be discovered in this transformation without any loss. In recent years, the DFT has also been used in DNA sequences analysis. The classical application of DFT including prediction the location of exons in DNA sequences, genomic signature and periodicity analysis [34–37].

Considering the signal sequences $u_1(n)$, $u_2(n)$ and $u_3(n)$ defined in Section 2.1, the DFT of signal at frequency $k$is calculated by

$$U_i(k) = DFT[u_i(n)] = \sum_{n=0}^{N-1} u_i(n) e^{-j\frac{2\pi}{N}nk},\quad k = 0,1,\ldots,N-1; i = 1,2,3 \quad (2)$$

where $j = \sqrt{-1}$.

Table 2
Information of sequences used in our test.

Sequence name	NCBI accession number
Blue whale	NP_007066
Bornean orangutan	NP_008235
Cat	NP_008261
Common chimpanzee	NP_008196
Fin whale	NP_006899
gibbon	NP_007832
gorilla	NP_008222
Gray seal	NP_007079
Harbor seal	NP_006938
Human	AP_000649
Horse	ADQ55101
Mouse	NP_904338
Opossum	NP_007105
Pigmy chimpanzee	NP_008209
Platypus	NP_008053
Rat	AP_004902
Rhino	YP_002520019
Sumatran orangutan	NP_007845
Wallaroo	NP_007404
Tiger	ADK73290
Korean bovine	YP_209215
Spain bovine	AKK32014

In our method, every protein sequences will be represented by three numerical sequences. The DFT power spectrum of the signal at frequency k will be defined as

$$PS(k) = \sum_{i=1}^{3} |U_i(k)|^2, \quad k = 0, 1 \ldots, N-1 \tag{3}$$

2.3. Dynamic Time Warping

The Dynamic Time Warping has been widely used in the analysis of speech signals. It is first proposed by Sakoe and Chiba in 1978 [38], aiming to eliminate the nonlinear fluctuation in speech pattern time axis. This property could be used in the analysis of genes if we consider the protein sequences as genomic signal inputs. Recently, researchers have applied the DTW algorithm in the analysis of genetic signals. Skutkova et al. used DTW to classify DNA signals and they have obtained some excellent results [39,40].

In Fig. 1, we give an example to illustrate the function of DTW. Assume data1 and data2, which have similar wave shapes, are same spoken word from different speakers. The subfigure a shows the two original signals has similar shapes, but obviously they are in different time scales. It is hard to tell whether they are from the same word. However, in subfigure b, the two signals have the same wave shapes after the DTW, which means the two signals are from the same word.

In this paper, DTW is applied to calculate the distance of different power spectra. We assume there are two power spectra

$$PS_1(k_1) \quad (k_1 = 0, 1, \ldots, M-1), \quad PS_2(k_2) \quad (k_2 = 0, 1, \ldots, N-1)$$

Table 3
similarity/dissimilarity of 22 kinds of animals.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
1	0																					
2	0.773	0																				
3	0.6073	0.7402	0																			
4	0.7356	0.5935	0.7373	0																		
5	0.2693	0.7698	0.573	0.7089	0																	
6	0.8056	0.6492	0.7325	0.6371	0.796	0																
7	0.8362	0.7991	0.7825	0.6393	0.8809	0.6817	0															
8	0.6922	0.787	0.6168	0.8027	0.7342	0.7899	0.9324	0														
9	0.7234	0.7823	0.6484	0.7766	0.7437	0.7727	0.8938	0.2107	0													
10	0.7832	0.6711	0.722	0.4971	0.7885	0.5414	0.6569	0.7941	0.8339	0												
11	0.7148	0.7234	0.6448	0.7555	0.7189	0.7683	0.8991	0.683	0.6984	0.7577	0											
12	0.7184	0.8341	0.7096	0.8027	0.7216	0.8231	0.9665	0.7239	0.7332	0.8207	0.7348	0										
13	0.7284	0.8274	0.7539	0.8621	0.7458	0.8686	1	0.7839	0.7394	0.8784	0.7287	0.7241	0									
14	0.7525	0.6402	0.7168	0.4056	0.7466	0.6185	0.6591	0.8118	0.7849	0.5381	0.7685	0.8589	0.9143	0								
15	0.7398	0.7428	0.7173	0.7593	0.7239	0.7997	0.8807	0.7663	0.7657	0.7718	0.7532	0.7776	0.8091	0.7927	0							
16	0.7544	0.9225	0.7435	0.8505	0.7864	0.7991	0.8177	0.8257	0.8221	0.7799	0.7279	0.8625	0.8292	0.8865	0.8853	0						
17	0.7875	0.7944	0.8335	0.8411	0.8157	0.8174	0.8964	0.8312	0.8217	0.852	0.7951	0.8116	0.792	0.8712	0.8102	0.8054	0					
18	0.6592	0.5813	0.6342	0.658	0.6543	0.7318	0.8514	0.7249	0.7178	0.7364	0.6682	0.7195	0.7576	0.7077	0.6915	0.7887	0.7581	0				
19	0.6858	0.7034	0.7492	0.7788	0.6866	0.7099	0.8833	0.7057	0.7166	0.7887	0.6786	0.7561	0.7009	0.7752	0.6889	0.8266	0.8192	0.6458	0			
20	0.6882	0.8015	0.5009	0.803	0.6389	0.809	0.8869	0.6557	0.6988	0.7584	0.7406	0.7652	0.7902	0.8204	0.8072	0.7709	0.8307	0.699	0.7963	0		
21	0.6547	0.708	0.648	0.7473	0.6951	0.6789	0.793	0.6331	0.6644	0.7088	0.6631	0.7599	0.7063	0.7526	0.7416	0.7576	0.78	0.6697	0.6959	0.6222	0	
22	0.6455	0.7223	0.5916	0.747	0.6751	0.6874	0.7846	0.6316	0.6914	0.6847	0.652	0.7448	0.7061	0.7319	0.7017	0.726	0.768	0.6829	0.7218	0.5976	0.05929	0

Fig. 2. The phylogenetic tree of 22 species based on our algorithm.

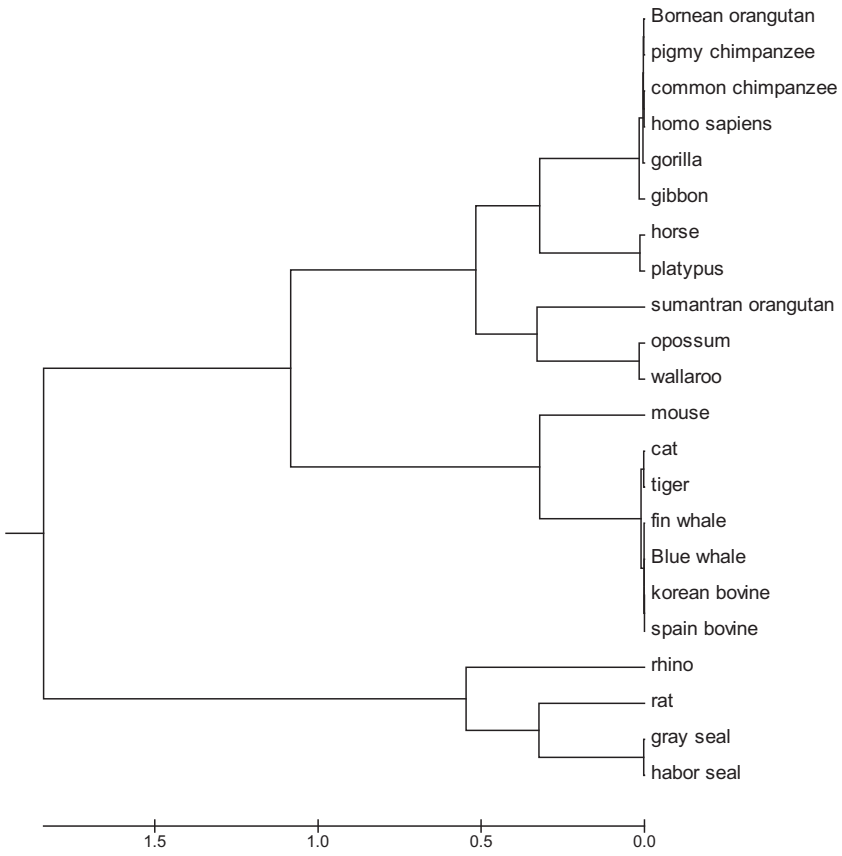

Fig. 3. The phylogenetic tree of 22 species based on Yau's protein map.

To simplify the symbol, we use two sequences to represent the two power spectra

$$a_1, a_2, \ldots a_p \ldots, a_M \quad b_1, b_2, \ldots b_q \ldots, b_N$$

where $a_p = PS_1(p-1)$, $b_q = PS_2(q-1)$ $(p = 1,2,\ldots M; q = 1,2,\ldots N)$.

Define a distance

$$d(p,q) = \|a_p - b_q\|_2 \quad (p = 1,2,\ldots M; q = 1,2,\ldots N)$$

as a metric of the difference between feature vectors a_p and b_q. The accumulated distance is calculated by formula (4).

$$D(p,q) = \begin{cases} 2d(1,1) & p = 1; q = 1 \\ d(1,q) + D(1,q-1) & p = 1; 2 \le q \le N \\ d(p,1) + D(p-1,1) & 2 \le p \le M; q = 1 \\ \min[D(p-1,q) + d(p,q), D(p,q-1) + d(p,q), D(p-1,q-1) + 2d(p,q)] & 2 \le p \le M; 2 \le q \le N \end{cases} \tag{4}$$

Apparently, the accumulated distance depends on the pairwise distance $d(p,q)$ and the minimum from the previous values. The values of $D(p,q)$, which will be used as the metric of similarity of two sequences, will form a table. The sequence warping path is derived on the basis of minimization of the backward way from the right upper corner to the left lower corner [39]. For two sequences, the minor $D(p,q)$ is, the more similar they are.

3. Results and discussion

To verify the approach we proposed, we choose different datasets of various species and take several experiments. We construct the

Table 4
Information of protein sequences used in this paper.

Sequence name	NCBI accession number
A/Adachi/2/1957(H2N2)	BAD16637.1
A/bar-headed_goose/Qinghai/1/2005(H5N1)	BAM85828.1
A/Beijing/4/2009(H1N1)	ACR67256.1
A/Berkeley/1/1968(H2N2)	BAD16641.1
A/blue-winged_teal/Ohio/566/2006(H7N9)	ABS89412.1
A/California/1/1966(H2N2)	AAO46235.1
A/California/04/2009(H1N1)	AEE69012.1
A/cat/Germany/R606/06(H5N1)	ABF61763.1
A/chicken/Dongguan/1096/2014(H7N9)	AJJ96855.1
A/Cygnus_olor/Italy/742/2006(H5N1)	ABF50822.1
A/chicken/Quzhou/2/2015(H7N9)	AKI82227.1
A/Duck/Ohio/118C/93(H1N1)	AAF77041.1
A/blue_winged_teal/Louisiana/A00557206/2009(H7N7)	ALT67567.1
A/canine/Guangxi/1/2011(H9N2)	AEK07935.1
A/chicken/China/AH-10-01/2010(H9N2)	AEE73586.1
A/chicken/Hubei/01-MA01/1999(H9N2)	AEO92432.1
A/chicken/Iran/B263/2004(H9N2)	ACD47112.1
A/England/1/1961(H2N2)	AAO46220.1
A/equine/Prague/1/1956(H7N7)	AAC57418.1
A/equine/Santiago/77(H7N7)	AAQ90293.1
A/fowl/Weybridge(H7N7)	AAA43425.1
A/Georgia/1/1967(H2N2)	AAO46244.1
A/goose/Czech_Republic/1848-K9/2009(H7N9)	ACX53685.1
A/GuangzhouSB/01/2009(H1N1)	ACR49238.1
A/Nagasaki/07N020/2008(H1N1)	ADC45738.1
A/lesser_white-fronted_goose/HuNan/412-3Y/2010(H7N7)	AIW60686.1
A/muscovy_duck/Vietnam/LBM66/2011(H5N1)	BAM36161.1
A/tree_sparrow/Shanghai/01/2013(H7N9)	AGW82590.1

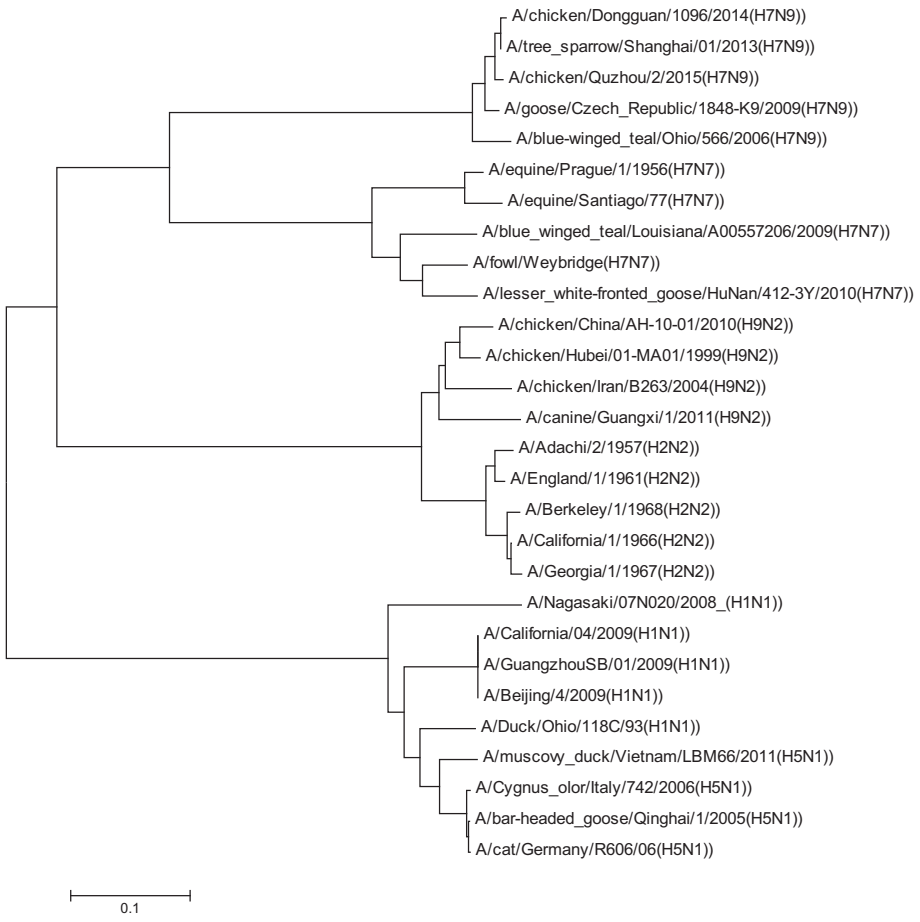

Fig. 4. The phylogenetic tree of 28 influenza A virus by Mega 6.06.

phylogenetic trees to get the cluster results and illustrate the distance between species in the evolution.

3.1. ND5 protein sequences of 22 species

Our scheme is applied to test 22 kinds of animal first. We choose the NADH dehydrogenase subunit 5 (ND5) sequences from NCBI database as our inputs. All the information of sequences we used is listed in Table 2. Table 3 reveals our results. Corresponding to every species, a number is assigned in the table: 1-blue whale, 2-bornean orangutan, 3-cat, 4-common chimpanzee, 5-fin whale, 6-gibbon, 7-gorilla, 8-gray seal, 9-habor seal, 10-human, 11-horse, 12-mouse, 13-opossum, 14-pigmy chimpanzee, 15-platypus, 16-rat, 17-rhino, 18-sumantran orangutan, 19-wallaroo, 20-tiger, 21-korean bovine, 22-spain bovine. It is noticed in Table 3 the pairs (blue whale, fin whale) (common chimpanzee, pigmy chimpanzee) and (Korean bovine, Spanish bovine) have a shorter distance in our analysis. The homologies revealed in the table are in agreement with evolutionary relation satisfactorily. Moreover, we also construct the phylogenetic tree of the 22 species in Fig. 2.

In Fig. 2, some reasonable cluster results are revealed. We find that the primates, such as common chimpanzee, pigmy chimpanzee, human, gorilla, orangutan and gibbon are much closer than other species in the evolutionary distance. Besides, different kinds of whale, bovine and seal are also located in the same branch respectively. All the classifications we've obtained are in agreement with the classical evolution theory. As a comparison, we apply the method in Ref. [21] to analyze the dataset in Table 2. The results are shown in Fig. 3. The results obtained in Figs. 2 and 3 have similar clusters. However, there also exists some difference. The Bornean orangutan and Sumatran orangutan should have a closer relation than other species, but obviously they are located in different branches in Fig. 3. Besides, the results in Fig. 3 also indicate the two kinds of whales and two kinds of bovines are much closer than others in the phylogeny. In Fig. 2, all the improper classifications are corrected. This experiment indicates our scheme is effective in the similarity analysis of animal ND5 proteins.

3.2. Neuraminidase proteins of influenza A virus

The influenza A virus has been a major threat to human and animals [41]. The viruses could be identified to different subtypes according to the different viral surface proteins hemagglutinin and neuraminidase. Until now 18 H serotypes (H1 to H18) and 11 N serotypes (N1 to N11) of influenza A viruses have been identified. The influenza A viruses have caused epidemic among human and animals. Some of the most lethal viruses are H1N1, H2N2, H5N1 and H7N9. We take 28 kinds of influenza A virus as samples in our test. All the sequences are picked from NCBI database. The sequences information is listed in Table 4.

We use the Mega software (version 6.06) to calculate the distance between 28 kinds of influenza A virus, drawing the phylogenetic tree in Fig. 4. In Fig. 4, we notice that most of the virus are classified correctly except the virus (A/Duck/Ohio/118C/93(H1N1)), which belongs to

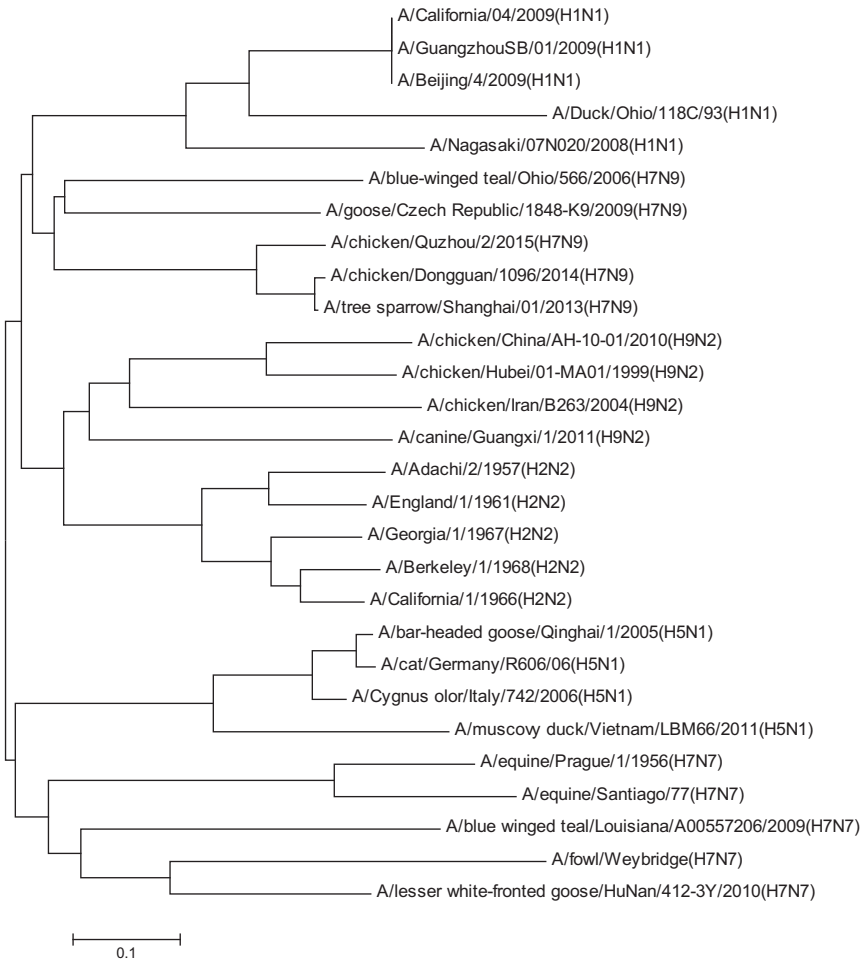

Fig. 5. The phylogenetic tree of 28 influenza A virus calculated by our method.

H1N1 in virology. Clearly, this is an improper classification. Using the same virus data, we apply our method to calculate the similarity of 28 influenza A virus, getting the results shown in Fig. 5. The cluster results from our method matches the classification in virology correctly. The viruses from same type are clustered in the same branch respectively. We notice the wrong classification in Mega software has been corrected in our method. Furthermore, it is also noticed that the viruses appeared in adjacent years are much closer in the phylogeny. For example, the virus (A/blue-winged_teal/Ohio/566/2006(H7N9)) is much closer to the virus (A/goose/Czech_Republic/1848-K9/2009(H7N9)) than (A/chicken/Quzhou/2/2015(H7N9)).

As a comparison, another software is also applied in our test. The cluster results from Clustal X software is illustrated in Fig. 6. The results in Fig. 6 are similar with ours. However, as illustrated in the phylogenetic tree, the viruses which belong to H2N2 are clustered in different branches. We conclude from the figures that the results obtained from different methods have an overall agreement even though there exists some variation between different methods. The phylogenetic trees in different figures reveal similar classification of influenza A virus. Among the three methods, our approach is more accurate in the test.

3.3. Coronavirus spike proteins

As a further comparison, we construct a phylogenetic tree for 50 coronavirus spike proteins. The coronavirus could cause some severe epidemics, for example, SARS. We use some coronavirus spike proteins as inputs to test our method. All the data comes from the Table 3 in reference [21]. The relations revealed in Fig. 7 are similar to the phylogeny reported in reference [21]. All the SARS coronavirus gather in the same branch. The coronavirus from same species has a much closer relationships. Due to the discussions above, our method is proved reasonable and effective.

4. Conclusion

In this work, techniques from signal process have been applied in the analysis of protein sequences. The approach in this paper provides an intuitive solution to analyze the protein sequences. We establish a novel measure based on Discrete Fourier Transform and Dynamic Time Warping to analyze the similarity of protein sequences. Based on the values of hydropathy and isoelectric point, we assign different radians to the amino acids according to their ranks. A three dimensional representation is constructed to represent all the amino acids. With the help of DFT and DTW, we get the power spectra and scale the spectra to the same length. The distances between species are evaluated by constructing phylogenetic trees. We use different datasets including animals and viruses to test our method. Compared to the existing methods and softwares, the computational time of our algorithm is large. However, there still exists ways to improve our method. For example, in the DTW process, a proper filter or sampling method could be considered

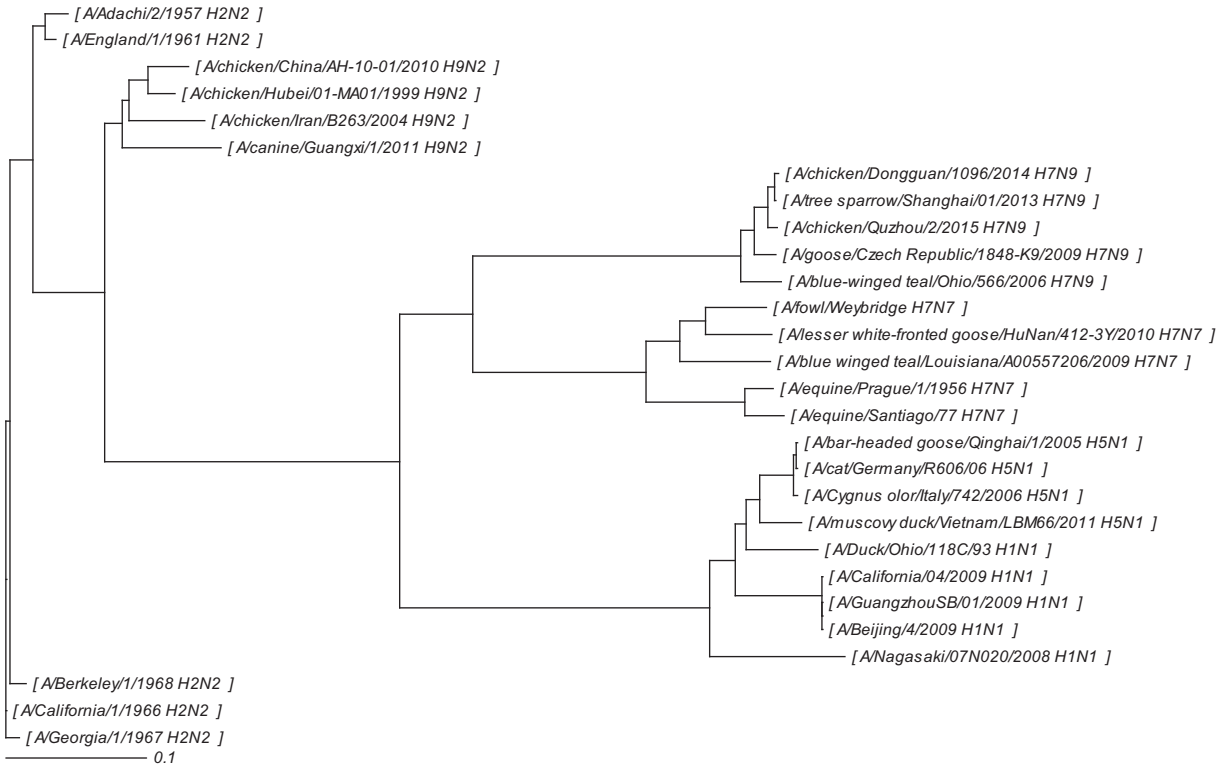

Fig. 6. The phylogenetic tree of 28 influenza A virus calculated by Clustal X.

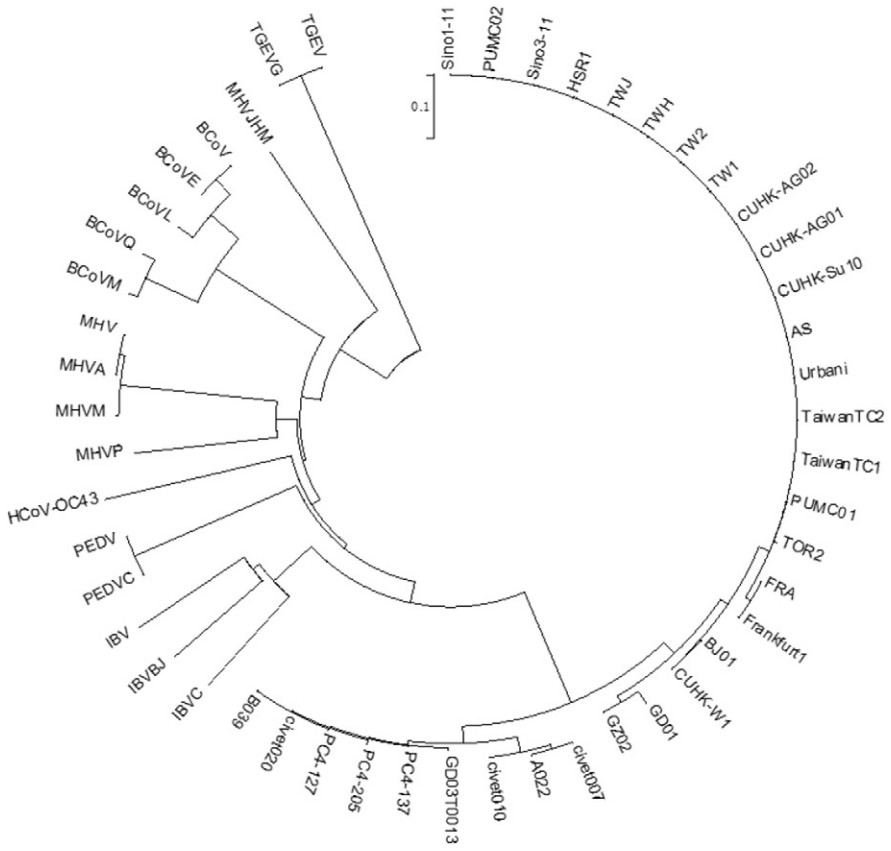

Fig. 7. The phylogenetic tree of 50 coronavirus spike proteins.

to pick some important information from results of the DFT instead of keeping all the values of spectra. Also, the DTW algorithm could be improved to reduce the running time of our method. In the test, we find the method in our paper provides accurate classification of different species. An improved DFT-DTW method will be our goal in the future works.

References

[1] 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.
[2] A. Nandy, A new graphical representation and analysis of DNA sequence structure. 1. Methodology and application to globin genes, Curr. Sci. 66 (1994) 309–314.
[3] H.J. Jeffrey, Chaos game representation of gene structure, Nucleic Acids Res. 18 (1990) 2163–2170.
[4] M. Randić, M. Vračko, N. Lerš, D. Plavšić, Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation, Chem. Phys. Lett. 371 (2003) 202–207.
[5] S.S.T. Yau, J.S. Wang, A. Niknejad, C. Lu, N. Jin, Y.K. Ho, DNA sequence representation without degeneracy, Nucleic Acids Res. 31 (2003) 3078–3080.
[6] X.Q. Liu, Q. Dai, Z.L. Xiu, T.M. Wang, PNN-curve: a new 2D graphical representation of DNA sequences and its application, J. Theor. Biol. 243 (2006) 555–561.
[7] B. Liao, Y. Zhang, K.Q. Ding, T.M. Wang, Analysis of similarity/dissimilarity of DNA sequences based on a condensed curve representation, Theochem. J. Mol. Struct. 717 (2005) 199–203.
[8] Z. Cao, B. Liao, R. Li, A group of 3D graphical representation of DNA sequences based on dual nucleotides, Int. J. Quantum Chem. 108 (2008) 1485–1490.
[9] N. Jafarzadeh, A. Iranmanesh, A novel graphical and numerical representation for analyzing DNA sequences based on codons, MATCH-Commun, Math. Comput. Chem. 68 (2012) 611–620.
[10] A. El-Lakkani, S. El-Sherif, Similarity analysis of protein sequences based on 2D and 3D amino acid adjacency matrices, Chem. Phys. Lett. 590 (2013) 192–195.
[11] N. Jafarzadeh, A. Iranmanesh, C-curve: a novel 3D graphical representation of DNA sequence based on codons, Math. Biosci. 241 (2013) 217–224.
[12] Y.H. Yao, X.Y. Nan, T.M. Wang, A new 2D graphical representation - classification curve and the analysis of similarity/dissimilarity of DNA sequences, Theochem. J. Mol. Struct. 764 (2006) 101–108.
[13] W. Hou, Q. Pan, M. He, A. Novel, 2D representation of genome sequence and its application, J. Comput. Theor. Nanosci. 11 (2014) 1745–1749.
[14] L. Bo, W. Tian-Ming, New 2D graphical representation of DNA sequences, J. Comput. Chem. 25 (2004) 1364–1368.
[15] C. Yin, S.S. Yau, An improved model for whole genome phylogenetic analysis by Fourier transform, J. Theor. Biol. 382 (2015) 99–110.
[16] T. Hoang, C. Yin, H. Zheng, C. Yu, R. Lucy He, S.S. Yau, A new method to cluster DNA sequences using Fourier power spectrum, J. Theor. Biol. 372 (2015) 135–145.
[17] C. Yu, R.L. He, S.S. Yau, Protein sequence comparison based on K-string dictionary, Gene 529 (2013) 250–256.
[18] T. Ma, Y. Liu, Q. Dai, Y. Yao, P.-A. He, A graphical representation of protein based on a novel iterated function system, Phys. A 403 (2014) 21–28.
[19] P.A. He, D. Li, Y. Zhang, X. Wang, Y. Yao, A 3D graphical representation of protein sequences based on the Gray code, J. Theor. Biol. 304 (2012) 81–87.
[20] L. Ling, K. Fen, H. Jilin, N. Xuying, Y. Yuhua, A 3-D graphical method applied to the similarities of protein sequences, 2012 Spring Congress on Engineering and Technology (S-CET 2012), 2012 (4 pp.-4 pp.).
[21] M.K. Gupta, R. Niyogi, M. Misra, An alignment-free method to find similarity among protein sequences via the general form of Chou's pseudo amino acid composition, SAR QSAR Environ. Res. 24 (2013) 597–609.
[22] S.S.T. Yau, C.L. Yu, R. He, A protein map and its application, DNA Cell Biol. 27 (2008) 241–250.
[23] C. Yu, S.Y. Cheng, R.L. He, S.S. Yau, Protein map: an alignment-free sequence comparison method based on various properties of amino acids, Gene 486 (2011) 110–118.
[24] C. Yu, M. Deng, S.Y. Cheng, S.C. Yau, R.L. He, S.S. Yau, Protein space: a natural method for realizing the nature of protein universe, J. Theor. Biol. 318 (2013) 197–204.
[25] S.S. Yau, W.G. Mao, M. Benson, R.L. He, Distinguishing proteins from arbitrary amino acid sequences, Sci. Rep. 5 (2015) 7972.
[26] Y. Li, K. Tian, C. Yin, R.L. He, S.S. Yau, Virus classification in 60-dimensional protein space, Mol. Phylogenet. Evol. 99 (2016) 53–62.
[27] 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.
[28] A. El-Lakkani, H. Mahran, An efficient numerical method for protein sequences similarity analysis based on a new two-dimensional graphical representation, SAR QSAR Environ. Res. 26 (2015) 125–137.
[29] Y. Li, Q. Liu, X. Zheng, P.-A. He, UC-Curve: a highly compact 2D graphical representation of protein sequences, Int. J. Quantum Chem. 114 (2014) 409–415.
[30] P. Wąż, D. Bielińska-Wąż, 3D-dynamic representation of DNA sequences, J. Mol. Model. 20 (2014) 2141.
[31] P. Wąż, D. Bielińska-Wąż, A. Nandy, Descriptors of 2D-dynamic graphs as a classification tool of DNA sequences, J. Math. Chem. 52 (2014) 132–140.
[32] A. Czerniecka, D. Bielińska-Wąż, P. Wąż, T. Clark, 20D-dynamic representation of protein sequences, Genomics 107 (2016) 16–23.
[33] X.H. Xia, W.H. Li, What amino acid properties affect protein evolution? J. Mol. Evol. 47 (1998) 557–564.
[34] C. Yin, S.S. Yau, Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence, J. Theor. Biol. 247 (2007) 687–694.
[35] D. Anastassiou, Frequency-domain analysis of biomolecular sequences, Bioinformatics 16 (2000) 1073–1081.
[36] S.A. Marhon, S.C. Kremer, Gene prediction based on DNA spectral analysis: a literature review, J. Comput. Biol. 18 (2011) 639–676.
[37] M. Akhtar, J. Epps, E. Ambikairajah, Signal processing in sequence analysis: advances in eukaryotic gene prediction, IEEE J. Sel. Top. Sign. Proces. 2 (2008) 310–321.
[38] H. Sakoe, S. Chiba, Dynamic-programming algorithm optimization for spoken word recognition, IEEE Trans. Acoust. Speech Signal Process. 26 (1978) 43–49.
[39] H. Skutkova, M. Vitek, P. Babula, R. Kizek, I. Provaznik, Classification of genomic signals using dynamic time warping, BMC Bioinf. 14 (2013) 7.
[40] H. Skutkova, M. Vitek, K. Sedlar, I. Provaznik, Progressive alignment of genomic signals by multiple dynamic time warping, J. Theor. Biol. 385 (2015) 20–30.
[41] D.J. Alexander, A review of avian influenza in different bird species, Vet. Microbiol. 74 (2000) 3–13.