MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **70** (2013) 451-466

ISSN 0340 - 6253

P-H Curve, a Graphical Representation of Protein Sequences for Similarities Analysis

Yuxin Liu [1, a], Dan Li[1, a], Kebo Lu[2], Yandong Jiao[3, 4], Ping-An He [1, 4*]

1 College of Science, Zhejiang Sci-Tech University, Hangzhou 310018, P. R. China
2 Department of Mathematics, Ocean University of China, Qingdao 266100, P. R. China
3 School of Sciences, Hebei University of Technology, Tianjin 300401, P. R. China
4 State Key Laboratory of Scientific and Engineering Computing, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, P. R. China

(Received September 28, 2012)

Abstract: Based on two kinds of physicochemical properties of amino acids, a novel 2D graphical representation of protein sequences was constructed. This graphical representation has no circuit or degeneracy, therefore, the correspondence between the sequences and the graphical curves is one-to-one. Then，a mathematical descriptor was introduced to compare the similarities between protein sequences. The NADH dehydrogenase subunit 5 (ND5) protein sequences of 9 species was used as an example to illustrate our approach. By the correlation and significance analysis, the comparisons between our results and some other graphical representation results with the ClustalW results were given to show the usefulness and efficiency of our approach.

Introduction

With more and more genome sequences being available online, analysis of the biological sequences becomes focus of research. Many methods have been proposed to analyze biological sequences, such as alignment, alignment-free methods [1-3]. In the alignment-free methods, the graphical representation of biological sequences can transform biological

* Correspondence to: P. –A. He; E-mail: pinganhe@zstu.edu.cn.

[a] These authors contributed equally to the work.

sequences into visual zigzag curves and offer efficient numerical characterization. Thus it has emerged as a very powerful tool for the visualization and analysis of the biological sequences [1-3]. Furthermore, methods based on graphical representations have been extensively applied in many relevant realms of bioinformatics [1-5].

The graphical representation of protein sequence was delayed from DNA graphical representation. The problem of graphical representations of protein stems from the combinatorial complexity associated with 20 factorial ways in which 20 amino acids (AAs) can be ordered [3]. Recently, many graphical representations of protein have been suggested to describe and analyze protein sequences [3-34]. 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, Randic [6] ,Yao [7-8], Yau [9], Wen [10], Li [11], He [12-13], Wu [14], Maaty [15] and Yu [16] have proposed a number of different graphical representations of proteins, respectively. According to the genetic code, Randic [17-20], and Bai [21] have given some graphical representations of proteins and the sequence descriptors. In addition, applying reduced protein models, Li [22-23] and Randic *et al.* [24], proposed several graphical representations of proteins which is similar to DNA representation. Based on the idea of cyclic order of 20 amino acids, the graphical approach of Jeffrey [3] for DNA representations was generalized to obtain the graphical representations of proteins by Randic [25-26] and He [27-29] *et al.* Instead of a square, they used a twenty-side polygon and placed the twenty amino acids on the periphery of the unit circle. Similar to existing graphical representation of DNA, some modified graphical representations of proteins were constructed to compare the similarities/dissimilarities of proteins [30-34].

In the paper, two indices of physicochemical properties of 20 amino acids, hydrophobicity value and isoelectric point, were considered to graphically represent protein sequences. The graphical representation has no circuit or degeneracy, so that the correspondence between protein sequences and protein graphs is one to one. Based on the ratio between the distance and the cosine of correlation angle of two vectors corresponding to two curves, the numerical

description was introduced to compare the similarities of proteins. Finally, an example of similarities/dissimilarities among the NADH dehydrogenase subunit 5 (ND5) proteins of nine different species was given to illustrates the efficiency of our approach. By the correlation and significance analysis, the comparisons between our results and results of other graphical representation with the ClustalW results were given. The results show that our approach has higher correlations to ClustalW results for all nine species than other approaches.

The graphical representation of protein sequences

Proteins consist of twenty kinds of natural amino acids. Since the earliest protein sequences and structures were determined, it has been clear that the positioning and properties of amino acids are key to understand many biological processes. In this paper, based on two kinds of physicochemical properties indexes of 20 amino acids, hydrophobicity and isoelectric point (pI) at 25^{o}C, the graphical representation of protein was constructed as follows:

Firstly, the 2D Cartesian coordinates of each amino acids were considered according to their values of hydrophobicity and isoelectric point (pI) at 25^{o}C. The values of the two parameters were listed in Table 1. Observing Table 1, we can find that all the values of pI are positive. In order to save the space, we can deal with by the formula (1),

$$\begin{cases} x_i = \dfrac{p_i - \min}{\max - \min} \\ y_i = h_i \end{cases} \tag{1}$$

where p_i and h_i are the values of pI and hydrophobicity for each amino acid, max and min are the maximum value and minimum value of pI among 20 amino acids.

Thus, through above-mentioned transformation, the new coordinates of 20 amino acids were listed at the last two columns in Table 1.

Fig.1 illustrates 20 special vectors respecting the 20 amino acids in 2D Cartesian coordinates, respectively. Thus, each amino acid can be numerically characterized by a unique vector. Observing Fig1, we can easily find the hydrophilicity for each amino acid, in which the I ,V， L , F, C, M, A are hydrophobic, the G,S,T,W,Y are neutral and the P, K, R, H, Q, N,

E are hydrophilic.

Table 1 Two parameters of the 20 amino acids and their new coordinates

Number	Amino acids	pI(at 25℃)	Hydrophobicity	x-coordinate	y- coordinate
1	G	5.97	-0.4	0.4005	-0.4
2	A	6	1.8	0.4043	1.8
3	T	5.6	-0.7	0.3542	-0.7
4	S	5.68	-0.8	0.3642	-0.8
5	P	6.3	-1.6	0.4418	-1.6
6	V	5.96	-4.2	0.3992	-4.2
7	L	5.98	3.8	0.4018	3.8
8	I	6.02	4.5	0.4068	4.5
9	M	5.74	1.9	0.3717	1.9
10	F	5.48	2.8	0.3392	2.8
11	Y	5.66	-1.3	0.3617	-1.3
12	W	5.89	-0.9	0.3095	-0.9
13	D	2.77	-3.5	0	-3.5
14	E	3.22	-3.5	0.0563	-3.5
15	N	5.41	-3.5	0.3304	-3.5
16	Q	5.65	-3.5	0.3605	-3.5
17	K	9.74	-3.9	0.8723	-3.9
18	R	10.76	-4.5	1	-4.5
19	H	7.59	-3.2	0.6033	-3.2
20	C	5.07	2.5	0.2879	2.5

Then, given a protein sequence with N amino acids $S = s_1, s_2, \cdots, s_N$, we inspect it by stepping one amino acid at a time. For the step $i (i = 1,2,\cdots,N)$, the point $P_i(x_i, y_i)$ can be constructed as follows:

$$\begin{cases} x_i = \sum_{k=1}^{i} S_{\mathrm{k}}^1 \\ y_i = \sum_{k=1}^{i} S_k^2 \end{cases} \tag{2}$$

where $S_k^j (j = 1,2)$ represent the jth component of the vector corresponding to S_k. The points in the 2D graphical representation are obtained by the sum of vectors representing amino acids in the sequence. When i is from 1 to N, we have points $P_1, P_2, \cdots, P_N$. Connecting the

adjacent points, we can obtain a curve for each protein. And we call the curve P-H curve of the protein. With our method, the P-H curves of the following two short protein segments of Saccharomyces cerevisiae are plotted in Fig. 2.

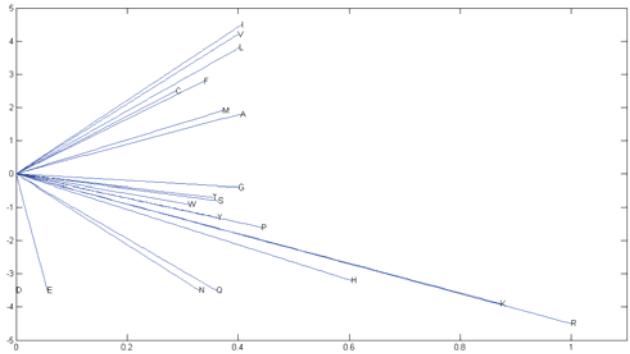

Fig. 1 The plot for the vectors of 20 amino acids in 2D Cartesian coordinates.

Protein Ⅰ:WTFESRNKPAKDPVILWLNGGPGCSSLTGL

Protein Ⅱ:WFFESRNKPANDPIILWLNGGPGCSSFTGL

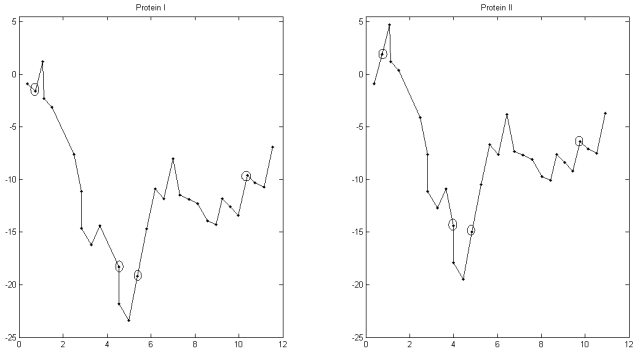

Fig. 2 The 2D graphical representations of Protein Ⅰ and Ⅱ

Observing Fig.2, we can find that two protein curves are very similar to each other and several local matching segments occur in two curves. By observing these similarities, we can find the different amino acids only at the sites 2, 11, 14, and 27.

The values of isoelectric point (*pI*) are all positive. By formula (2), we can easily know

that the graphical representation has no circuit or degeneracy which can be proved in mathematics. Therefore, the correspondence between protein sequences and graphical curves is one-to-one.

Furthermore, we get the graphical representation of the ND5 proteins of nine different species (Fig. 3) by our approach.

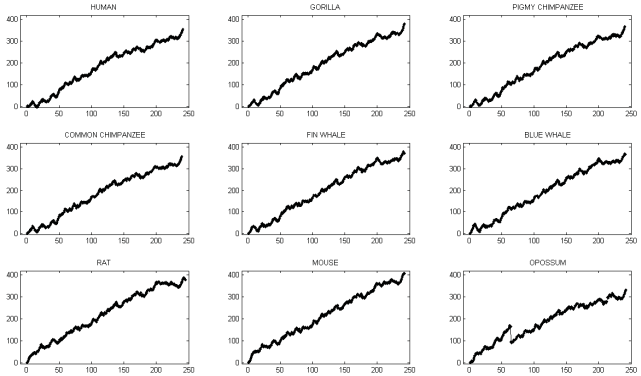

Fig. 3 The graphical representation of the ND5 proteins of nine different species

It is shown from Fig.3 that the ND5 proteins curves of human, gorilla, common chimpanzee, and pygmy chimpanzee are more similar to each other than other curves. The curves of fin whale and blue whale, and those of rat and mouse are also quite similar. In addition, we can find that the ND5 protein of opossum (the most remote species form the remaining mammals) is very dissimilar to the other eight species.

The numerical characterization of protein sequences

In order to numerically characterize each protein, a novel sequence descriptor called the difference vector of protein sequence is introduced as follows.

Given two protein sequences with the same length n : $S^1 = s_1^1, s_2^1, \cdots, s_n^1$ and $S^2 = s_1^2, s_2^2, \cdots, s_n^2$, P_i^1 and P_i^2 are the ith points of the graphical curves of two protein sequences. Let $\vec{p}_i^1$ and $\vec{p}_i^2$ $(i = 1,2,\cdots,n)$ be the vectors corresponding to P_i^1 and P_i^2. We

define the difference vectors $\overrightarrow{\Delta}_i^1 = \vec{p}_i^{\,1} - \vec{p}_{i-1}^{\,1}$ and $\overrightarrow{\Delta}_i^2 = \vec{p}_i^{\,2} - \vec{p}_{i-1}^{\,2}$ for two P-H curves, in which let $\vec{p}_0^{\,j} = (0, 0), \quad (j = 1, 2)$. Then, the distance between the P-H curves of two protein sequences, S^1 and S^1, is defined to characterize the similarity of two protein sequences, as follows:

Table 2 the value of d_i for two protein sequences

S^1	x	y	Δ_i^1		S^2	x	y	Δ_i^2		d_i
W	0.3905	-0.9000	0.3905	-0.9000	W	0.3905	-0.9000	0.3905	-0.9000	0
T	0.7447	-1.6000	0.3542	-0.7000	F	0.7297	1.9000	0.3392	2.8000	4.2092
F	1.0839	1.2000	0.3392	2.8000	F	1.0688	4.7000	0.3392	2.8000	0
E	1.1402	-2.3000	0.0563	-3.5000	E	1.1252	1.2000	0.0563	-3.5000	0
S	1.5044	-3.1000	0.3642	-0.8000	S	1.4894	0.4000	0.3642	-0.8000	0
R	2.5044	-7.6000	1.0000	-4.5000	R	2.4894	-4.1000	1.0000	-4.5000	0
N	2.8348	-11.1000	0.3304	-3.5000	N	2.8198	-7.6000	0.3304	-3.5000	0
D	2.8348	-14.6000	0	-3.5000	D	2.8198	-11.1000	0	-3.5000	0
P	3.2766	-16.2000	0.4418	-1.6000	P	3.2616	-12.7000	0.4418	-1.6000	0
A	3.6809	-14.4000	0.4043	1.8000	A	3.6658	-10.9000	0.4043	1.8000	0
K	4.5532	-18.3000	0.8723	-3.9000	N	3.9962	-14.4000	0.3304	-3.5000	0.6789
D	4.5532	-21.8000	0	-3.5000	D	3.9962	-17.9000	0	-3.5000	0
P	4.9950	-23.4000	0.4418	-1.6000	P	4.4380	-19.5000	0.4418	-1.6000	0
V	5.3942	-19.2000	0.3992	4.2000	I	4.8848	-15.0000	0.4068	4.5000	0.3001
I	5.8010	-14.7000	0.4068	4.5000	I	5.2516	-10.5000	0.4068	4.5000	0
L	6.2028	-10.9000	0.4018	3.8000	L	5.6533	-6.7000	0.4018	3.8000	0
W	6.5932	-11.8000	0.3905	-0.9000	W	6.0438	-7.6000	0.3905	-0.9000	0
L	6.9950	-8.0000	0.4018	3.8000	L	6.4456	-3.8000	0.4018	3.8000	0
N	7.3254	-11.5000	0.3304	-3.5000	N	6.7760	-7.3000	0.3304	-3.5000	0
G	7.7259	-11.9000	0.4005	-0.4000	G	7.1765	-7.7000	0.4005	-0.4000	0
G	8.1264	-12.3000	0.4005	-0.4000	G	7.5770	-8.1000	0.4005	-0.4000	0
P	8.5682	-13.9000	0.4418	-1.6000	P	8.0188	-9.7000	0.4418	-1.6000	0
G	8.9687	-14.3000	0.4005	-0.4000	G	8.4193	-10.1000	0.4005	-0.4000	0
C	9.2566	-11.8000	0.2879	2.5000	C	8.7071	-7.6000	0.2879	2.5000	0
S	9.6028	-12.6000	0.3642	-0.8000	S	9.0713	-8.4000	0.3642	-0.8000	0
S	9.9580	-13.4000	0.3642	-0.8000	S	9.4355	-9.2000	0.3642	-0.8000	0
L	10.3867	-9.6000	0.4018	3.8000	F	9.7747	-6.4000	0.3392	2.8000	1.0021
T	10.7409	-10.3000	0.3542	-0.7000	T	10.1289	-7.1000	0.3542	-0.7000	0
G	11.1414	-10.7000	0.4005	-0.4000	G	10.5294	-7.5000	0.4005	-0.4000	0
L	11.5432	-6.9000	0.4018	3.8000	L	10.9312	-3.7000	0.4018	3.8000	0

$$D(S^1,S^2)=\sum_{i=1}^{N}\frac{d(\vec{\Delta}_i^{\,1},\vec{\Delta}_i^{\,2})}{\left|\cos<\vec{\Delta}_i^{\,1},\vec{\Delta}_i^{\,2}>\right|} \tag{3}$$

As an example, we compute the x-coordinate and y- coordinate of Protein Ⅰ and Protein Ⅱ,and their difference vectors and the Euclidean distance d_i of two difference vector listed in Table 2.

In Table 2, we can easily find out where the substitutions between the two protein sequences occur according to the value of d_i. That is to say, if the two protein sequences have the same amino acids, the value of d_i is 0, otherwise the value of d_i is larger than 0.

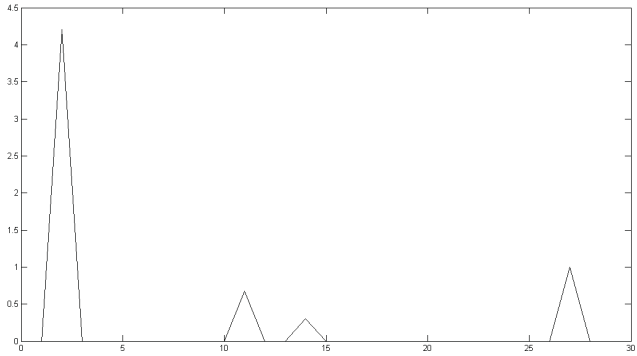

Fig. 4 The value of every d_i between Protein Ⅰ and Ⅱ

Then we plotted the points (i,d_i) for two short protein segments of Saccharomyces cerevisiae, and connected the adjacent points in Fig. 4. Form Fig. 4, we can easily find some differences between two proteins and their different location. And from Table 2, we can compute the value of D is 6.1903 and use this value as distance between two protein sequences.

From Fig.4 we can easily find out that there are four different points (the 2nd, 11th, 14th, 27th points) which values of d_i are not 0, implying that there are 4 substitutions between the two protein sequences in position 2, 11, 14 and 27, and this is consistent with the fact that we obtained from two protein sequences.

The similarities/dissimilarities analysis of proteins

Given two protein sequences S_1 and S_2 with different length n and m $(n < m)$, we define a generalized distance form the formula (3) to compare their similarities.

First, the sum $\vec{\mu}$ of the weighted mean is defined as follows:

$$\vec{\mu} = \sum_{i=1}^{n} \frac{count(i)}{n} \vec{\alpha}(i) \tag{4}$$

where n is the length of short protein sequence, $count(i)$ is the number of occurrences of the ith amino acid in the shorter protein sequence, $\vec{\alpha}(i)$ is the vector of the ith amino acid in Table 2. Then, the sum $\vec{\mu}$ of the weighted mean of the shorter protein sequence was computed by using formula (4), and inserted in the end of the last point until the shorter protein sequence has the same length as the longer protein sequence. Thus, we obtained a new protein sequence $S_1^{'}$ with length m for the shorter sequence S_1, and it is like this:

$$\vec{S}_1^{'} = \vec{S}_1 + (m-n)\vec{\mu} \tag{5}$$

Finally, the distance between two protein sequences $S_1^{'}$ and S_2 based on the formula (3) was defined as the distance between two protein sequences S_1 and S_2

To illustrate our method, we compared the similarities among sequences belonging to nine ND5 proteins(Table 3) by using the above method.

Table 3 The ND5 proteins of nine different species

i	Species	ID(NCBI)	Length
1	Human	AP_000649	603
2	Gorilla	NP_008222	603
3	Pygmy Chimpanzee	NP_008209	603
4	Common Chimpanzee	NP_008196	603
5	Fin Whale	NP_006899	606
6	Blue whale	NP_007066	606
7	Rat	AP_004902	610
8	Mouse	NP_904338	607
9	Opossum	NP_007105	602

The values of D between every pair of species are listed in Table 4. Observing Table 4, we

can see that the values of D among human, gorilla, common chimpanzee and pygmy chimpanzee are all quite small, and the value of D between the fin whale and blue whale and the value between rat and mouse are also quite small. On the other hand, we find that the ND5 protein of opossum is very dissimilar to the other eight species. Another interesting fact is that the value of D between human and common chimpanzee, and the value between human and pygmy chimpanzee are smaller than that of human and gorilla. That is to say, the ND5 protein of human is more similar to that of common chimpanzee and pygmy chimpanzee than that of gorilla. We believe that the results are not coming by accident since they are consistent with the known fact of evolution.

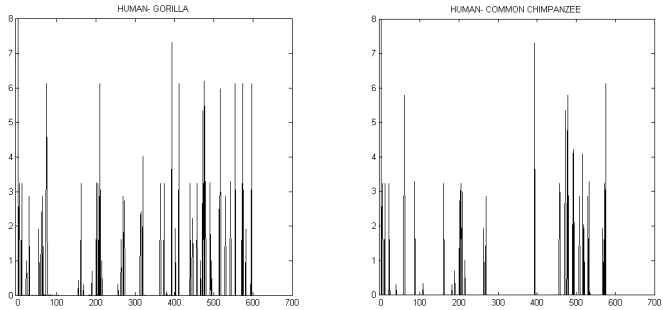

Fig. 5 The value of d_i between the amino acids in the same sequential positions in protein sequences. Human-gorilla(left) and human-common chimpanzee(right).

As we know that ClustalW is one of the most popular multiple sequence alignment programs for DNA or proteins [35]. To compare our method with ClustalW, we listed the results of multiple sequence alignment among the nine species by using ClustalW, which was shown in a distance matrix in Table 5. Observing Table 4 and Table 5, we can see that the sequence similarity results are almost consistent in both our method and ClustalW.

The correlation coefficient r is a measure of the strength of the linear relationship between two variables. It is defined in terms of the covariance of the variables divided by their standard deviations. In the paper, the correlation coefficient between the results from our

method and those from ClustalW are calculated to show the relationship between these two methods.

Table 4 the values of D between the ND5 proteins of nine different species

	Gorilla	P.Chim	C.Chim	F.Whale	B.whale	Rat	Mouse	Opossum
Human	174	111.4	109.7	531.1	549.2	736.2	703.5	1129.4
Gorilla		104.5	139.5	561.3	572.6	688	654.9	1187.5
P.Chim			67.3	525.7	536.8	720.9	691.4	1153.7
C.Chim				533.8	544.8	726.1	697.7	1169
F.Whale					69.2	631.5	607.7	1232.6
B.whale						622.3	585	1240.6
Rat							388	1288.3
Mouse								1233.7

The correlation coefficient r between the first row of the similarities/dissimilarities matrix based on the values of D in Table 4 and the first row of the ClustalW distance matrix in Table 5 was calculated, and result was 0.9351. The first row in both matrices is relative to human protein, the second to gorilla and so on. In the first column of Table 6, the correlation coefficients for the rows relative to all nine species were listed. Observing the values of first column of Table 6, we can see that all correlation coefficients r are more than 0.8. These imply that the results of two methods are consistent to each other.

Table 5 The distances for the ND5 protein sequences of nine species calculated by ClusterW

	Gorilla	P.Chim	C.Chim	F.Whale	B.whale	Rat	Mouse	Opossum
Human	10.7	7.1	6.9	41.0	41.3	50.2	48.9	50.4
Gorilla		9.7	9.9	42.7	42.4	51.4	49.9	54.0
P.Chim			5.1	40.1	40.1	50.2	48.9	50.1
C.Chim				40.4	40.4	50.8	49.6	51.4
F.Whale					3.5	45.3	46.8	52.7
B.whale						45.0	45.9	52.7
Rat							25.9	54.0
Mouse								50.8

Similarly, the correlation coefficients among the results of Refs.[6, 9, 13, 14, 21], and the distance matrix Table 5 are calculated to compare with our approach. All these results were

listed in Table 6. Observing the results in Table 6, we can see that our approach possesses higher correlation coefficients with ClutalW for all species than other approaches.

Table 6 The correlation coefficient results for the nine ND5 proteins of our approach and the approaches in Refs. [6, 9, 13, 14, 21], as compared with ClustalW

	Our approach & ClustalW	Ref. [6] (Table 3) & ClustalW	Ref. [6] (Table 4) & ClustalW	Ref. [9] (Table 3) & ClustalW	Ref. [9] (Table 4) & ClustalW	Ref. [13] (Table 3) & ClustalW	Ref. [14] (Table 3) & ClustalW	Ref. [14] (Table 4) & ClustalW	Ref. [21] (Table 4) & ClustalW
Human	0.9351	0.8849	0.9236	0.9143	0.4566	0.9059	0.9306	0.7177	0.9729
Gorilla	0.9263	0.7398	0.9317	0.6969	0.7850	0.8800	0.9293	0.7748	0.9763
P.chim	0.9309	0.8889	0.9542	0.9222	0.7861	0.6822	0.8403	0.7661	0.9819
C.chim	0.9334	0.8921	0.9607	0.9257	0.7676	0.8819	0.9344	0.7845	0.9756
F.whale	0.8801	0.6839	0.7388	0.6026	0.2839	0.3287	0.3508	0.5318	0.9485
B.whale	0.8808	0.7297	0.8148	0.6981	-0.0731	0.3380	0.6486	0.5512	0.9450
Rat	0.8647	0.8085	0.5882	0.7167	0.3693	0.6696	0.4453	0.8376	0.8709
Mouse	0.8343	0.7612	0.5221	0.6711	0.4881	0.5914	0.4192	0.4559	0.7296
Opossum	0.9957	-0.4344	-0.2992	-0.4746	-0.2044	-0.1342	-0.2975	-0.4326	0.5447

Table 7 The t-values computed for the correlation coefficients $|r| > 0.8$; on the basis of these, the significance is determined

	Our approach & ClustalW	Ref. [6] (Table 3) & ClustalW	Ref. [6] (Table 4) & ClustalW	Ref. [9] (Table 3) & ClustalW	Ref. [9] (Table 4) & ClustalW	Ref. [13] (Table 3) & ClustalW	Ref. [14] (Table 3) & ClustalW	Ref. [14] (Table 4) & ClustalW	Ref. [21] (Table 4) & ClustalW
Human	6.9783	5.0273	6.3742	5.9713		5.6601	6.7254		11.1322
Gorilla	6.5034		6.7839			4.9017	6.6582		11.9256
P.chim	6.7401	5.1334	8.4366	6.3088			4.1012		13.7249
C.chim	6.8837	5.2218	9.1533	6.4773		4.9489	6.9417		11.7615
F.whale	4.9041								7.9198
B.whale	4.9229		3.7180						7.6437
Rat	4.5551	3.6343							4.6891
Mouse	4.0045							4.0567	
Opossum	28.5348								

Because the set of data in this paper is small ($n = 9$), it is easy to produce a high correlation. Therefore, the significance of the correlation was used to check whether the correlation of the

two sets of data was sufficiently strong or likely to have occurred by chance. We tested the statistical significance for correlation coefficient values that are more than 0.8 by using the *t*-test. Our sample size equals 9, so the degree of freedom is 7. A *t*-value of 3.499 or more than 3.499 indicates a significance of less than 0.01 chance of having occurred by coincidence. In Table 7, we listed the *t*-values corresponding to the *r*-values which are more than 0.8. All computed *t*-values are more than 3.499. From Table 7, we can easily find that all our results are more than 3.499, which indicates that the *r*-values in Table 6 have not occurred by chance.

Finally, using the UPGMA method, the phylogenetic tree was obtained from Table 4, this was shown in Fig. 6, and is consistent with the results obtained with Clustal W methods.

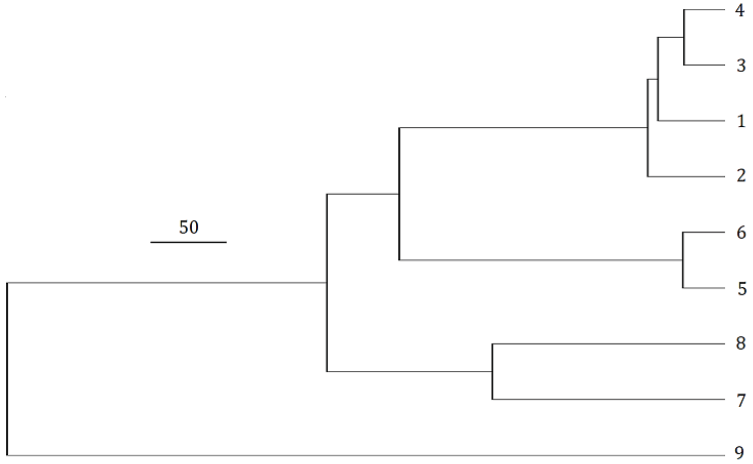

Fig. 6 The phylogenetic tree of nine ND5 proteins based on the values of D in Table 4. (1: human, 2: gorilla, 3: pygmy chimpanzee, 4: common chimpanzee, 5: fin whale, 6: blue whale, 7: rat, 8: mouse, 9: opossum)

Conclusions

In general, proteins with similar sequences frequently carry out similar structures. A marked similarity between two protein sequences may reflect the fact that they are derived by evolution from the same ancestral sequence Therefore, the similarity/dissimilarity analysis of protein sequences is an important and interesting topic in bioinformatics. Many mathematical approaches have been proposed to describe protein sequences. The graphical representations

of protein transform protein sequences from letters string into graph which accompanied with mathematical objects such as vectors or matrices to use as sequence descriptors and compare these mathematical objects. For example, if the mathematical objects are vectors, we can calculate the Euclidean distance or correlation angle between each two vectors. Based on their values, the similarity /dissimilarity matrix can be obtained.

In this paper, we proposed a novel 2D graphical representation of protein sequences, called P-H curve. Then mathematical descriptor is used to characterize the differences between two P-H curves. Based on the distance between the P-H curves of two protein sequences, the similarities/dissimilarities matrix among proteins can be calculated to compare their similarities/dissimilarities.

Finally, an example shows that our method is fast, convenient and has the potential for long protein sequences.

Acknowledgments: We thank the referees for many valuable comments that have improved this manuscript. We appreciate the financial support of this work that was provided by National Natural Science Foundation of China (61170110, 11171042, 11001072 and 11101381). This work was also partially supported by the National Center for Mathematics and Interdisciplinary Sciences, CAS.

References

[1] B. W. Dorota, Graphical and numerical representations of DNA sequences: statistical aspects of similarity, *J. Math. Chem.* **49** (2011) 2345-2407.

[2] A. Nandy, M. Harle, S. C. Basak, Mathematical descriptors of DNA sequences: development and applications, *ARKIVOC* **9** (2006) 211-238.

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

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

[5] B. Liao, B. Y. Liao, X. M. Sun, Q. G. Zeng, A Novel method for similarity analysis and protein sub-cellular localization prediction, *Bioinformatics* **26** (2010) 2678-2683.

[6] M. Randić, 2-D graphical representation of proteins based on physico-chemical properties of amino acids, *Chem. Phys. Lett*. **440** (2007) 291-295.

[7] Y. H. Yao, Q. Dai, C. Li, P. A. He, X. Y. Nan, Y. Z. Zhang, Analysis of similarity/ dissimilarity of protein sequences, *Proteins* **73** (2008) 864-871.

[8] Y. H. Yao, Q. Dai, L. Li, X. Y. Nan, P. A. He, Y. Z. Zhang, Similarity/dissimilarity studies of protein sequences based on a new 2D graphical representation, *J. Comput. Chem.* **31** (2010) 1045-1052.

[9] S. S. T. Yau, C. L. Yu, R. He, A protein map and its application, *DNA Cell Bio.* **l27** (2008) 241-250.

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

[11] F. Q. Li, G. H. Huang, B. Liao, Z. B. Liu, H-L Curve: A novel 2-D graphical representation of protein sequences, *MATCH Commun. Math. Comput. Chem.* **61** (2009) 519-532.

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

[13] P. A. He, J. Z. Wei, Y. H. Yao, Z. X. Tie, A novel graphical representation of proteins and its application, *Physica A* **391** (2012) 93–99.

[14] 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.

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

[16] J. F. Yu, X. Sun, J. H. Wang, a novel 2D graphical representation of protein sequence based on individual amino acid, *Int. J. Quant. Chem.* **111** (2011) 2835-2843.

[17] M. Randić, J. Zupan, A. T. Balaban, Unique graphical representation of protein sequences based on nucleotide triplet codons, *Chem. Phys. Lett.* **397** (2004) 247-252.

[18] M. Randić, 2-D graphical representation of proteins based on virtual genetic code, *SAR QSAR Environ. Res.* **15** (2004) 147-157.

[19] M. Randić, A. T. Balaban, M. Novič, A. Založnik, T. Pisanski, A novel graphical representstion of proteins, *Period. Biol.***107** (2005) 403-414.

[20] M. Randić, K. Mehulić, D. Vukicević, T. Pisanski, D. Vikić-Topić, D. Plavsić, Graphical representation of proteins as four-color maps and their numerical characterization, *J. Mol. Graph. Model.* **27** (2009) 637-641.

[21] F. L. Bai, T. M. Wang, A 2-D graphical representation of protein sequences based on nucleotide triplet codons, *Chem. Phys. Lett.* **413** (2005) 458-462.

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

[23] C. Li , X. Q. Yu, L. Yang, X. Q. Zheng, Z. F. Wang, 3-D maps and coupling numbers for protein sequences, *Physica A* **388** (2009) 1967-1972.

[24] M. Randić, M. Vračko, M. Novič, D. Plavsić, Spectral representation of reduced protein models, *SAR QSAR Environ. Res.* **20** (2009) 415-427.

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

[26] M. Randić, J. Zupan, D. Vikić-Topić, On representation of proteins by star-like graphs, *J. Mol. Graph. Model.* **26** (2007) 290-305.

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

[28] P. A. He, A new graphical representation of similarity/dissimilarity studies of protein sequences, *SAR QSAR Environ. Res.* **21** (2010) 571-580.

[29] P. A. He, D. Li, Y. P. Zhang, X. Wang, Y. H. Yao, A 3D graphical representation of protein sequences based on the Gray code, *J. Theor. Biol.* **304** (2012) 81-87.

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

[31] M. Randić, On a geometry-based approach to protein sequence alignment, *J. Math. Chem.* **43** (2008) 756-772.

[32] M. Randić, M. Novič, M. Vračko, On novel representation of proteins based on amino acid adjacency matrix, *SAR QSAR Environ. Res.* **19** (2008) 339-349.

[33] M. Randić, M. Novič, Representation of proteins as walks in 20-D space, *SAR QSAR Environ. Res.* **19** (2008) 317-337.

[34] A. Nandy, A. Ghosh, P. Nandy, Numerical characterization of protein sequences and application to voltage-gated sodium channel alpha subunit phylogeny, *In Silico Biol.* **9** (2009) 77-87.

[35] M. A. Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A . Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, D.G. Higgins, Clustal W and Clustal X version 2.0, *Bioinformatics* **23** (2007) 2947-2948.