MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **65** (2011) 445-458

ISSN 0340 - 6253

A Novel Descriptor for Protein Similarity Analysis

Ping-an He [1,*], Xiao-fang Li[1], Jia-liang Yang[2], Jun Wang[3]

[1] Department of mathematics, College of Science, Zhejiang Sci-Tech University,
Hangzhou310018, P.R. China
[2] CAS-MPG Partner Institute of Computational Biology,
Shanghai 200031, P.R. China
[3] Department of Mathematics, Shanghai Normal University,
Shanghai, 200234, P.R. China

(Received June 17, 2010)

Abstract: Based on three kinds of indices of physicochemical properties on amino acids, a 3D graphical representation of protein sequences is introduced. Then, we present a graph matrix representation called the 2-order difference matrix to characterize the 3D graphical representation. Using this 2-order difference matrix, we propose an efficient method to compare protein sequences. Finally, an example is shown to compare the NADH dehydrogenase subunit 5 (ND5) protein sequences of 9 species.

Introduction

There are literally millions of bio-sequences in various sequence databases. Comparison of different sequences remains one of the upmost important tasks in Bioinformatics. It contributes to analyze sequence similarities and thus to infer phylogenetic relationships among extant species. Traditionally, almost all such comparisons are based on alignments, in which a distance function or a score function is used to represent insertion, deletion, and substitution in the compared sequences. In recent years, many alignment-free methods were proposed by incorporating concepts and algorithms from computational statistics, such as stochastic modeling of sequences and other Bayesian theory methods [1]. However, in these approaches, the chemical structures and properties of bio-molecules are ignored, and multiple

* Corresponding author: pinganhe@zstu.edu.cn

sequence alignment is computationally difficult.

An efficient method to compare DNA sequences was proposed by using their graphical representations[2-11], in which a DNA sequence is plotted as a curve on a 2D plane or 3D space. This kind of methods provides a simple way for viewing, sorting, and comparing various structures, and making the analysis of similarity between DNA sequences [2-11].

However, since there are 20 amino acids instead of 4 nucleotides, direct generalization of this method from DNAs to proteins is computationally difficult. As a result, the graphical representations of proteins were introduced only very recently [12-27]. In the representations, the 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 the vectors, and the numerical characterizations of the curves are used to compare the protein sequences. It is worthy of noticing that there are 20 factorial orders of the 20 vectors, among which some special orders can be used to reflect some physicochemical properties of amino acids. For example, Randic [19], Yao [20], Feng [21], and Wen [22] suggested some graphical representations of protein sequences based on the values of $pK1(COOH)$ and $pK2(NH_3^+)$ of the 20 amino acids.

In this paper, we first present a 3D graphical representation of proteins based on 3 kinds of physicochemical properties on the 20 amino acids, namely Hydrophilicity, $pK1(COOH)$, and isoelectric point (pI) at 25^oC . The surface hydrophilicity is a main factor to determine the solubility of proteins. High hydrophilicity will benefit the interaction between proteins and water, and thus increase the solubility of proteins. The proton-donating and proton-accepting ability of $pK1(COOH)$ is essential for the chemical properties of proteins, and the ionization constant of $pK1(COOH)$ determines the catalytic activities of enzymes. The pI is the pH of an aqueous solution of an amino acid (or peptide) at which the molecules on average have no net charge. It could reflect the innate structure of the protein sequence, rather than the apparent legitimate structure.

Then, a graph matrix representation is proposed to characterize the 3D graphical representation of a protein, and a 2-order difference matrix between two matrices is introduced to compare the similarities/dissimilarities of two proteins. As an example, we compare the ND5 protein sequences of nine different species, which are listed in Table 1.

Table 1 The ND5 proteins of nine different species

Species	ID(NCBI)	Sequence length
Human	AP_000649	603
Gorilla	NP_008222	603
Pigmy Chimpanzee	NP_008209	603
Common Chimpanzee	NP_008196	603
Fin Whale	NP_006899	606
Blue Whale	NP_007066	606
Rat	AP_004902	610
Mouse	NP_904338	607
Opossum	NP_007105	602

The graphical representation of protein sequences

Amino acids are the basic building blocks of protein molecules. We have adopted 3 parameters, namely hydrophilicity, $pK1(COOH)$ and pI at 25°C, to construct the 3D Cartesian coordinates of amino acids. The values of the 3 parameters are listed in Table 2 [28] (2nd-4th column).

As shown in Table 2 (2nd-4th column), $pK1(COOH)$ and pI values are positive. So, if we directly use the values as coordinates of the points representing the 20 amino acids in a Cartesian (x, y, z) coordinate system, then the points will all lie in the 1st and 4th quadrants. It is noticing that the average values of hydrophilicity, $pK1(COOH)$ and pI are -0.2150, 2.1870 and 6.3185, respectively. We take the point (-0.2150, 2.1870, 6.3185) as the origin and form a new Cartesian system. As a result, the 20 points are distributed in 7 quadrants in the new system. In Table 2 (5th-7th column), we list the new components of these points.

Table 2 Three parameters of the 20 amino acids and their coordinates in the new Cartesian system

Amino acid	Hydrophilicity	pK1(a-COOH)	pI (at 25°C)	x-coordinate	y-coordinate	z-coordinate
A	-0.5	2.35	6.11	-0.2850	0.1630	-0.2085
C	-1.0	1.71	5.02	-0.7850	-0.4770	-1.2985
D	3.0	1.88	2.98	3.2150	-0.3070	-3.3385
E	3.0	2.19	3.08	3.2150	0.0030	-3.2385
F	-2.5	2.58	5.91	-2.2850	0.3930	-0.4085
G	0.0	2.34	6.06	0.2150	0.1530	-0.2585
H	-0.5	1.78	7.64	-0.2850	-0.4070	1.3215
I	-1.8	2.32	6.04	-1.5850	0.1330	-0.2785
K	3.0	2.20	9.47	3.2150	0.0130	3.1515
L	-1.8	2.36	6.04	-1.5850	0.1730	-0.2785
M	-1.3	2.28	5.74	-1.0850	0.0930	-0.5785
N	0.2	2.18	10.76	0.4150	-0.0070	4.4415
P	0.0	1.99	6.30	0.2150	-0.1970	-0.0185
Q	0.2	2.17	5.65	0.4150	-0.0170	-0.6685
R	3.0	2.18	10.76	3.2150	-0.0070	4.4415
S	0.3	2.21	5.68	0.5150	0.0230	-0.6385
T	-0.4	2.15	5.60	-0.1850	-0.0370	-0.7185
V	-1.5	2.29	6.02	-1.2850	0.1030	-0.2985
W	-3.4	2.38	5.88	-3.1850	0.1930	-0.4385
Y	-2.3	2.20	5.63	-2.0850	0.0130	-0.6885

Given a protein sequence $S = s_1 s_2 \quad s_n$, we inspect it by stepping one amino acid at a time. For step $i\left(i=1,2, \quad n\right)$, a 3D space point $P_i(x_i, y_i, z_i)$ can be constructed as follows:

$$\begin{cases} x_i = \sum_{k=1}^{i} S_k^1 \\ y_i = \sum_{k=1}^{i} S_k^2 \\ z_i = \sum_{k=1}^{i} S_k^3 \end{cases},$$

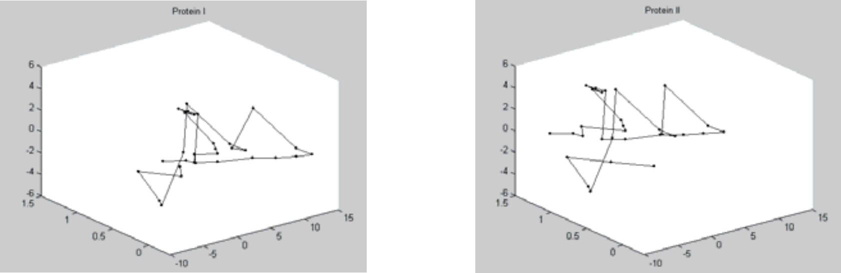
where $S_k^j \left(j=1,2,3\right)$ represents the *j*th component of the vector corresponding to s_k, and we set $p_0 = (0,0,0)$. When i runs from 1 to n, we obtain vertices $P_1, P_2, \quad, P_n$. Connecting the adjacent vertices, we can obtain a protein curve in 3D space.

To illustrate our method, we plot in Fig. 1 the 3D curves of the following two short protein segments of Saccharomyces cerevisiae,

Protein I：WTFESRNKPAKDPVILWLNGGPGCSSLTGL

Protein II：WFFESRNKPANDPIILWLNGGPGCSSFTGL

These two proteins were also plotted by Randic [19] using a recent 2D graphical representation.



a b

Fig. 1 The 3D graphical representations of Protein I and II.

Taking a closer look at Fig. 1, we find that the two protein curves are similar on the whole, which indicates that they have several local matching segments. However, there are 4 hugely different points (the 2nd, 11th, 14th, 27th points), which implies that there are 4 substitutions between the two protein sequences in position 2, 11, 14 and 27. As one can see, our curves are easier to view and analyze and have the potential for long protein sequences.

It is worthy of noticing that circuit or degeneracy can occur in the graphical representation when the number of 20 amino acids occurring in a fragment of protein are all equal. In the protein sequences, however, the probability of the condition is very small, almost approximate 0.

The 2-order difference matrix

In this section, we introduce a novel sequence descriptor called the 2-order difference matrix of two matrices, to numerically characterize each protein.

Let $A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ .. & .. & .. & .. \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{pmatrix}$ and $B = \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1m} \\ b_{21} & b_{22} & \cdots & b_{2m} \\ .. & .. & .. & .. \\ b_{n1} & b_{n2} & \cdots & b_{nm} \end{pmatrix}$ be two $n \times m$ matrices, then the 1-order difference matrix of A and B is defined as

$$N = A - B = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ .. & .. & .. & .. \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix} \qquad \begin{matrix} x_{ij} = a_{ij} - b_{ij} \\ (i = 1,\ldots,n,\ j = 1,\ldots,m) \end{matrix}.$$

Given N, the 2-order difference matrix of A and B is defined as

$$M = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1m} \\ y_{21} & y_{22} & \cdots & y_{2m} \\ .. & .. & .. & .. \\ y_{n,1} & y_{n,2} & \cdots & y_{n,m} \end{pmatrix} \qquad \begin{matrix} y_{ij} = x_{ij} - x_{(i+1)j} \\ i = 0,\ldots,n-1,\ j = 1,\ldots,m \end{matrix},$$

where $x_{0j} = 0$.

Recall that protein sequences are represented by a set of points in the 3D space in the last section. Thus, a protein of length n may be represented by an $n \times 3$ matrix P of the following form:

$$P = \begin{pmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ & & \\ x_n & y_n & z_n \end{pmatrix}$$

where x_i y_i z_i are the components of P_i corresponding to the ith amino acid of the protein.

In the following, we use Protein I and II in the previous section as an example to illustrate the definitions. The corresponding matrix, the 1-order and 2-order difference matrices of the two proteins are given in Table 3. It can be seen from the construction of M that: if two sequences have a same subsequence, then the values of corresponding rows in the 2-order difference matrix are all 0. As shown in Table 3, protein I and II have the same fragment FESRNKPA from site 3 to 10. In correspondence, the entries of the 2-order difference matrix M are all 0 from the 3rd to 10th rows.

Table 3 The corresponding matrix, the 1-order and 2-order difference matrices of two proteins

I	x	y	z	II	x	y	z	1-order difference			2-order difference		
W	-3.185	0.193	-0.4385	W	-3.185	0.193	-0.4385	0	0	0	0	0	0
T	-3.370	0.156	-1.1570	F	-5.470	0.586	-0.8470	2.10	-0.43	-0.31	-2.10	0.43	0.31
F	-5.655	0.549	-1.5655	F	-7.755	0.979	-1.2555	2.10	-0.43	-0.31	0	0	0
E	-2.440	0.552	-4.8040	E	-4.540	0.982	-4.4940	2.10	-0.43	-0.31	0	0	0
S	-1.925	0.575	-5.4425	S	-4.025	1.005	-5.1325	2.10	-0.43	-0.31	0	0	0
R	1.290	0.568	-1.0010	R	-0.810	0.998	-0.6910	2.10	-0.43	-0.31	0	0	0
N	1.705	0.561	3.4405	N	-0.395	0.991	3.7505	2.10	-0.43	-0.31	0	0	0
D	4.920	0.254	0.1020	D	2.820	0.684	0.4120	2.10	-0.43	-0.31	0	0	0
P	5.135	0.057	0.0835	P	3.035	0.487	0.3935	2.10	-0.43	-0.31	0	0	0
A	4.850	0.220	-0.1250	A	2.750	0.650	0.1850	2.10	-0.43	-0.31	0	0	0
K	8.065	0.233	3.0265	N	3.165	0.643	4.6265	4.90	-0.41	-1.60	-2.80	-0.02	1.29
D	11.280	-0.074	-0.3120	D	6.380	0.336	1.2880	4.90	-0.41	-1.60	0	0	0
P	11.495	-0.271	-0.3305	P	6.595	0.139	1.2695	4.90	-0.41	-1.60	0	0	0
V	10.210	-0.168	-0.6290	I	5.010	0.272	0.9910	5.20	-0.44	-1.62	-0.30	0.03	0.02
I	8.625	-0.035	-0.9075	I	3.425	0.405	0.7125	5.20	-0.44	-1.62	0	0	0
L	7.040	0.138	-1.1860	L	1.840	0.578	0.4340	5.20	-0.44	-1.62	0	0	0
W	3.855	0.331	-1.6245	W	-1.345	0.771	-0.0045	5.20	-0.44	-1.62	0	0	0
L	2.270	0.504	-1.9030	L	-2.930	0.944	-0.2830	5.20	-0.44	-1.62	0	0	0
N	2.685	0.497	2.5385	N	-2.515	0.937	4.1585	5.20	-0.44	-1.62	0	0	0
G	2.900	0.650	2.2800	G	-2.300	1.090	3.9000	5.20	-0.44	-1.62	0	0	0
G	3.115	0.803	2.0215	G	-2.085	1.243	3.6415	5.20	-0.44	-1.62	0	0	0
P	3.330	0.606	2.0030	P	-1.870	1.046	3.6230	5.20	-0.44	-1.62	0	0	0
G	3.545	0.759	1.7445	G	-1.655	1.199	3.3645	5.20	-0.44	-1.62	0	0	0
C	2.760	0.282	0.4460	C	-2.440	0.722	2.0660	5.20	-0.44	-1.62	0	0	0
S	3.275	0.305	-0.1925	S	-1.925	0.745	1.4275	5.20	-0.44	-1.62	0	0	0
S	3.790	0.328	-0.8310	S	-1.410	0.768	0.7890	5.20	-0.44	-1.62	0	0	0
L	2.205	0.501	-1.1095	F	-3.695	1.16	0.3805	5.90	-0.66	-1.49	-0.70	0.22	-0.13
T	2.020	0.4640	-1.8280	T	-3.880	1.124	-0.3380	5.90	-0.66	-1.49	0	0	0
G	2.235	0.617	-2.0865	G	-3.665	1.277	-0.5965	5.90	-0.66	-1.49	0	0	0
L	0.650	0.790	-2.3650	L	-5.250	1.450	-0.8750	5.90	-0.66	-1.49	0	0	0

The similarity/dissimilarity analysis of proteins

Using the 3D graphical representation of protein, we plot the curves of nine ND5 proteins in Fig. 2. It is shown from Fig.2 that the ND5 proteins curves of human, gorilla, common chimpanzee, and pigmy chimpanzee are more similar each other than other curves, although gorilla is slightly different. The curves of fin whale and blue whale, and those of rat and mouse are also quite similar. In addition, we find that the ND5 protein curve of opossum is obviously different from other species.

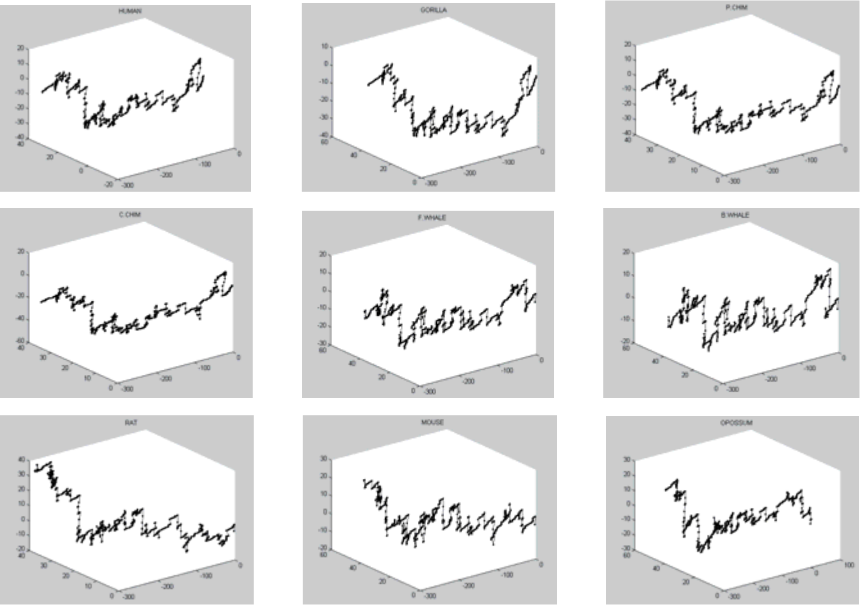


Fig. 2 The 3D graphical representations of nine ND5 proteins.

As we can see from Table 1 that the lengths of the nine ND5 protein sequences are different, so we will obtain nine matrices with different number of rows. In order to use the 2-order difference matrix, we must make the numbers of rows in the matrices equal. A simple strategy is to add several copies of the last row to the small matrices. For example, let B be a $m \times 3$ matrix, we expand it to a $n \times 3$ $(n > m)$ matrix B' as

$$B = \begin{pmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ & & \\ x_m & y_m & z_m \end{pmatrix}_{m \times 3} \rightarrow B' = \begin{pmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ & & \\ x_m & y_m & z_m \\ & & \\ x_m & y_m & z_m \end{pmatrix}_{n \times 3}$$

An alternative way to explain this strategy is: we successively add a new row, which is equal to 0 (the average of the 20 amino acids) plus the last row, to matrix B.

Let A and B be the representing matrices of two protein sequences with dimensions $n\times 3$ and $m\times 3$ ($n>m$), respectively. We calculate the 2-order difference of A and B' and consider it as the 2-order difference of A and B. Then, the distance between A and B is defined as

$$d=\sum_{i=1}^{n}\sum_{j=1}^{m}|y_{ij}|,$$

which is used to represent the distance between the two protein sequences.

In Table 4, we list the pair-wise distances among nine proteins by the above definition. Distances are the quantitative measure of the diversity between a pair of objects. The smaller are the distances, the more similar are the two proteins. As can be seen from Table 4, the smallest distance comes from the fin whale and blue whale pair. In addition, the distances among human, gorilla, common chimpanzee and pigmy chimpanzee, and that between rat and mouse are also quite small. On the other hand, we find that the ND5 protein of opossum (the most remote species form the remaining mammals) is very dissimilar to the other eight species. Another interesting fact is that the distances between human and common chimpanzee, and between human and pigmy chimpanzee are smaller than that of human and gorilla. That is to say, the ND5 protein of human is more similar to common chimpanzee and pigmy chimpanzee than gorilla. We believe that the results are not coming by accident since they are consistent with the known fact of evolution.

Table 4 The 2-order difference matrix distances for the ND5 protein sequences of nine species calculated by the 3D graphical representation

	Gorrilla	P.chim	C.chim	F.whale	B.whale	Rat	Mouse	Opossum
Human	**120.4**	**70.98**	**94.23**	422.35	432.29	580.06	547.15	968.89
Gorrilla		**107.3**	**134.49**	457.39	465.97	577.7	527.15	995.87
P.chim			**58.41**	418.15	426.73	576.98	545.21	967.36
C.chim				415.36	424.38	589.55	554.12	983.88
F.whale					**56.54**	559.59	559.95	1042.9
B.whale						550.65	554.01	1060.6
Rat							**328.18**	1077.1
Mouse								1073.1

As we know, ClustalW is one of the most popular alignment-based methods [29]. To compare our method with ClustalW, we list the multiple sequence alignment of the nine species by ClustalW as a distance matrix in Table 5.

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

Observing Table 5, the sequence similarity results is almost consistent to Table 4. Comparing Table 4 and Table 5, element by element, the results are plotted in the Fig 3. The Fig 3 clearly shows that the two tables agree in pointing that most similar are pairs of species with the 2-order difference matrix less than 120 and cluster W values below 11.

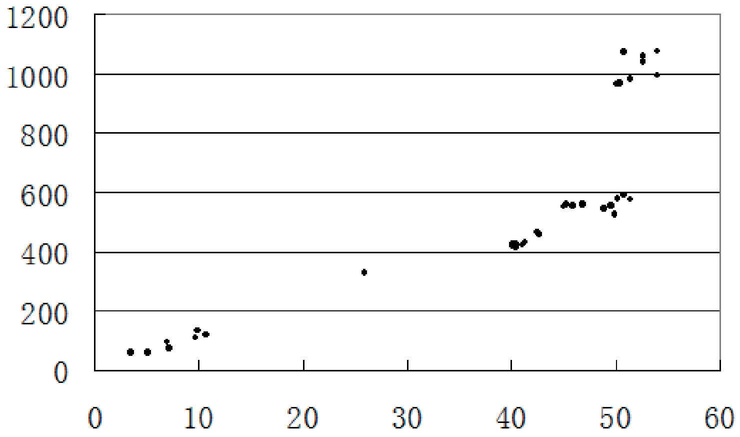


Fig. 3 The correlation analysis between Table 4 and Table 5.

Moreover, using UPGMA method, the phylogenetic tree obtained from Table 4 is shown in Fig 4, which is consistent with the results gotten by Table 5.

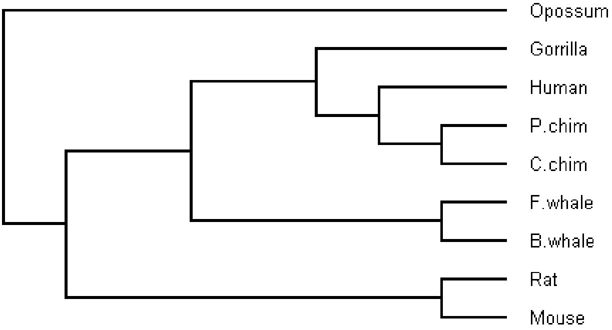


Fig. 4 The phylogenetic tree obtained form Table 4.

In addition, the projections of the 3D graphical representation of proteins on the X-Y, X-Z and Y-Z plane are considered to compare the similarities of the nine ND5 proteins, which might provide a possible way to reveal the biological functions of hydrophilicity-pK1, hydrophilicity-pI and pK1-pI. For example, to compare the nine ND5 proteins in Table 1, we take the values of hydrophilicity and pK1 in Table 2 as a 2D map. The distances among the nine ND5 proteins are calculated based on the method of the 2-order different matrix in Table 6-8, to provide a similarity analysis for the nine proteins.

In table 6-8, like the Table 4, the smallest distance comes form the fin whale and blue whale pair. And meanwhile, the ND5 protein of human is more similar to common chimpanzee and pigmy chimpanzee than gorilla.

Table 6 The 2-order difference matrix distances based on the 2D graphical representation for the ND5 protein sequences of nine species (hydrophilicity and pK1)

	Gorrilla	P.chim	C.chim	F.whale	B.whale	Rat	Mouse	Opossum
Human	**56.4200**	**41.2700**	**44.3800**	226.1080	231.3120	303.8380	280.5760	547.5320
gorrilla		**52.9900**	**60.1800**	236.8680	241.0120	306.5380	279.7760	562.1920
P.chim			**31.3300**	222.8980	227.0420	303.5880	278.0860	552.6880
C.chim				223.9080	228.2920	311.5180	285.1760	562.4120
F.whale					**32.7900**	295.3160	288.7380	590.3700
B.whale						290.6860	279.4080	597.2540
Rat							**159.2180**	608.3500
Mouse								603.7240

Table 7 The 2-order difference matrix distances based on the 2D graphical representation for the ND5 protein sequences of nine species (hydrophilicity and pI)

	Gorrilla	P.chim	C.chim	F.whale	B.whale	Rat	Mouse	Opossum
Human	**110.28**	**64.01**	**86.45**	391.16	400.59	545.04	513.58	911.04
Gorrilla		**98.91**	**125.81**	426.14	434.47	542.88	494.28	936.46
P.chim			**52.78**	388.07	396.4	543.51	513.23	909.85
C.chim				385.97	394.5	556.05	521.95	925.35
F.whale					**52.65**	527.37	528.12	980.6
B.whale						517.56	522.21	997.45
Rat							**307.15**	1017.5
Mouse								1014.1

Table 8 The 2-order difference matrix distances based on the 2D graphical representation for the ND5 protein sequences of nine species (pK1 and pI)

	Gorrilla	P.chim	C.chim	F.whale	B.whale	Rat	Mouse	Opossum
Human	**74.1000**	**36.6800**	**57.6300**	227.4385	232.6725	311.2455	300.1530	481.2055
Gorrilla		**62.7000**	**82.9900**	251.7785	256.4525	305.9855	280.2530	495.5055
P.chim			**32.7100**	225.3385	230.0125	306.8655	299.1130	474.2515
C.chim				220.8485	225.9625	311.5355	301.1230	482.0755
F.whale					**27.6400**	296.4900	303.0315	518.0590
B.whale						293.0500	306.3915	529.5930
Rat							**189.9985**	530.8310
Mouse								531.2165

Observing Table 6-8, we can see that the results about the similarity for nine proteins are almost consistent. Meanwhile, it is very significant that our approach for the first time has shown that pigmy chimpanzee is more similar to human than common chimpanzee, as it is, as can be seen from Table 4, Table 6-8, while this is not shown for ClusterW approach (Table 5) based on traditional methods of computer manipulations with protein sequences.

However, there are some inconsistent results, which might reflect the efficiency of the physicochemical properties of amino acids.

Conclusion

When analyzing similarities of protein sequences and phylogenetic relationships, comparison between different protein sequences is a vital step in bioinformatics. Although alignment is a most popular method used to compare protein sequences, graphical techniques provide us with a novel alignment-free way to compare different protein sequences.

In this work, we first present a 3D graphical representation of proteins. Then the 2-order difference matrix is introduced to numerically characterize the graphical representation to find matching fragment of amino acids between the two proteins. In addition, a new distance is suggested to compare the similarities of proteins. An illustrating example shows that our method is fast, convenient and has the potential for long protein sequences.

Moreover, the similarities of proteins are compared based on the projection of the 3D graphical representation of proteins on the X-Y, X-Z and Y-Z plane, respectively, to reveal the biological functions of hydrophilicity-pK1, hydrophilicity-pI and pK1-pI.

Acknowledgments

We thank the referees for many valuable comments that have improved this manuscript. This research was supported through the Zhejiang Provincial Natural Science Foundation of China (No. Y6090261) and the Zhejiang Provincial Open Foundation of the Most Important Subjects (No. SWYX0813).

Reference

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

[2] E. Hamori, Novel DNA sequence representation, *Nature* **314** (1985) 585.

[3] M. A. Gates, A simple way to look at DNA, *J. Theor. Biol*. **119** (1986) 319-328.

[4] P. M. Leong, S. Mogenthaler, Random walk and gap plots of DNA sequences, *Comput. Appl. Biosci.* **12** (1995) 503-511.

[5] A. Nandy, P. Nandy, Graphical analysis of DNA sequences structure: II. Relative abundances of nucleotides in DNAs, gene evolution and duplication, *Curr. Sci*. **68** (1995) 75-85.

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

[7] M. Randić, M. Novič, D. Vikić-Topić, D. Plavšić, Novel numerical and graphical representation of DNA sequences and proteins, *SAR. QSAR. Environ. Res.* **17** (2006) 583-595.

[8] Y. H. Yao, X. Y. Nan, T. M. Wang, Analysis of similarity/dissimilarity of DNA sequences based on a 3-D graphical representation, *Chem. Phys. Lett.* **411** (2005) 248-255.

[9] H. J. Jeffrey, Chaos game representation of gene structure, *Nucleic. Acids. Res.* **18** (1990) 2163-2170.

[10] C. T. Zhang, R Zhang, H. Y. Ou, The Z curve database: a graphic representation of genome sequences, *Bioinformatics* **19** (2003) 593-599.

[11] S. S. Yau, J. Wang, A. Niknejad, C. Lu, N. Jin, Y. K. Ho, DNA sequence representation without degeneracy, *Nucleic Acids. Res.* **31** (2003) 3078-3080.

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

[13] 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.

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

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

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

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

[18] B. Liao, J. W. Luo, R. F. Li, W. Zhu, Novel method for analyzing proteome, *Int. J. Quant. Chem.* **107** (2007) 1295-1300.

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

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

[21] J. Feng, Y. M. Wang, Characterization of protein primary sequences based on partial ordering, *J. Theoretical. Biol.* **254** (2008) 752-755.

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

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

[25] 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.

[26] W. Zhu, B. Liao, J. W. Luo, R. F. Li, Numerical characterization and similarity analysis of neurocan gene, *MATCH Commun. Math. Comput. Chem.* **57** (2007) 143-155.

[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] H. B. Shen, K. C. Chou, PseAAC. A flexible web-server for generating various kinds of protein pseudo amino acid composition, *Analytical. Biochem.* **373** (2008) 386-388.

[29] C. Z. Guo, M. Q. Sun, Clustal W- A software for multiple sequence alignment of protein and nucleic acid sequence, *Lett. Biotech*. **11** (2000) 146-149.