Communications in Information and Systems
Volume 18, Number 3, 193–208, 2018

Similarity analysis of protein sequences based on a new graphical representation method

Yuyan Zhang and Jia Wen

Similarity analysis of protein sequences is often utilized to identify the similarities/dissimilarities of protein sequences, which is the key step to predict the structures and functions of the newly identified proteins. Integrating the properties of isoelectric point and hydrophobic factors for amino acids, a new graphical representation of protein sequence is proposed to depict the features of proteins, in which both the local and global information of protein sequence are shown. Our new graphical curve has no degeneracy or arbitrariness, and the relationship between a protein sequence and its corresponding graphical curve is one-to-one. In addition, two numerical characterizations derived from the protein graph are utilized to quantify each protein sequence. The examination of similarity of the DN6 proteins from eight different species shows the utility of our new method.

Introduction

Similarity analysis of protein sequences is often utilized to identify the similarities/dissimilarities of different protein sequences, which is also the key step to predict the structures and functions of newly identified proteins. Mathematical comparison for a large volume of DNA and protein molecular data is still one of the challenges for bio-scientists. The traditional approaches for the similarity analysis of DNA and protein sequences are based on computer-oriented and computer-intensive comparison of sequences by sequence alignment. In recent years, many methods have been reported to analyze the huge amounts of gene data. The graphical representation method is one of those methods, which has such an important advantage over other methods: it provides not only visual qualitative inspection of gene data to

Key words and phrases: protein sequence; relative distance; geometrical center; leading eigenvalue; similarity analysis.

recognize major differences among similar gene sequences, but also mathematical characterizations of graphical curves. Hamori and Ruskin first used a 3-D H curve to represent a gene sequence [1]. Later, Gates published a 2-D graphical representation that is simpler than H curve [2]. However, Gates' graphical representation has high degeneracy. Several authors outlined different graphical representations for proteins and DNA sequences [3–20].

The first graphical representation of protein sequence emerges more than ten years, and a lot of works proposed in graphical representation of proteins have been outlined [16–26]. The team from Prof. Yau has published several papers in this area [9, 27–31]. Following their previous work, they utilized the moment vectors to depict protein sequences and generate a universal protein map [27, 28]. Then motivated by the protein map, a novel method called protein space was developed to realize the nature of protein space. Their proposed methods are applied successfully in the graphical representation for protein sequence. Especially, a vector graph was used depict the distribution of bases in a protein sequence, and the relationship between a vector and its corresponding sequence is one-to-one.

Most existing graphical representations for protein sequence involve some degree of arbitrariness, such as the assignment of amino acids to vertices of graphs of geometrical objects considered. One feasible way to avoid such difficulties is to search for ordering 20 natural amino acids according to some of their physicochemical properties. Randic [22], Yao et al. [23], He et al.[24], Wen and Zhang [25], and Yu et al.[26] outlined different protein maps by constructing these spatial order points based on a selected pair of physicochemical properties of amino acids. However, Randic's protein map is static, and will be invisible for the long sequence. Similarly, protein graphs proposed by Yao et al., He et al., and Yu et al. accompany with the loss of information due to overlapping and crossing of the curve with itself, or some degree of arbitrariness.

From the work of Yao et al. [23], among all physicochemical properties of amino acids, the isoelectric point and hydropathy index have been verified to be the most important indexes for protein molecules, which display the essential features of protein molecule efficiently. Therefore, utilizing the isoelectric point and hydropathy index for 20 amino acids, a new 2-D graphical representation of protein sequences is proposed to reflect the innate information of protein sequences. Significantly, our new graph effectively conquers the degeneracy of graphical curve, and has no arbitrariness. Associating with this new graphical representation, two numerical characterizations of mutations are utilized to compare the similarities among protein sequences belonging to

Amino acid	Symbol	Isoelectric point (pI)	Hydropathy index	x-Coordinate	y-Coordinate
Glycine	G	5.97	−0.4	5.97	−0.4
Alanine	A	6.01	1.8	6.01	1.8
Threonine	T	5.87	−0.7	5.87	−0.7
Serine	S	5.68	−0.8	5.68	−0.8
Proline	P	6.48	−1.6	6.48	−1.6
Valine	V	5.97	4.2	5.97	4.2
Leucine	L	5.98	3.8	5.98	3.8
Isoleucine	I	6.02	4.5	6.02	4.5
Methionine	M	5.74	1.9	5.74	1.9
Phenylalanine	F	5.48	2.8	5.48	2.8
Tyrosine	Y	5.66	−1.3	5.66	−1.3
Tryptophan	W	5.89	−0.9	5.89	−0.9
Aspartic acid	D	2.77	−3.5	2.77	−3.5
Glutamic acid	E	3.22	−3.5	3.22	−3.5
Aspargine	N	5.41	−3.5	5.41	−3.5
Glutamine	Q	5.65	−3.5	5.65	−3.5
Lysine	K	9.74	−3.9	9.74	−3.9
Arginine	R	10.76	−4.5	10.76	−4.5
Histidine	H	7.59	−3.2	7.59	−3.2
Cysteine	C	5.07	2.5	5.07	2.5

Table 1: The selected pair of properties of 20 amino acids and their corresponding coordinates in 2-D Cartesians.

eight ND6 (NADH dehydrogenase subunit 6) proteins: human (Homo sapiens, AP_000650), gorilla (Gorilla gorilla, NP_008223), common chimpanzee (Pan troglodytes, NP_008197), wallaroo (Macropus robustus, NP_007405), harbor seal (Phoca vitulina, NP_006939), gray seal (Halichoerus grypus, NP_007080), rat (Rattus norvegicus, AP_004903), and mouse (Mus musculus, NP_904339).

1. A new graphical representation of protein sequence

Twenty kinds of natural amino acids form the completed protein molecule with diverse functions. Recognizing the properties of amino acids is essential to separate protein molecules. There are several different kinds of physicochemical properties for amino acid, such as the relative molecular mass, solubility limit, specific rotation, isoelectric point (pI), hydropathy index, melting point, and pK_a values for terminal amino acid groups COOH and

NH_{3+}, etc., which are all important factors for the structure and function of protein molecule. After analyzing and comparing six types of physicochemical properties of amino acids, Yao et al. [23] found the isoelectric point and hydropathy index are two of the most important factors in characterizing amino acid, which is consistent with the fact that hydrophobic interaction and charges are the key factors for sustaining the structures and function of protein molecules. Therefore, in this study, the isoelectric point and hydropathy index are selected to characterize the identities of 20 amino acids, which are presented in Table 1.

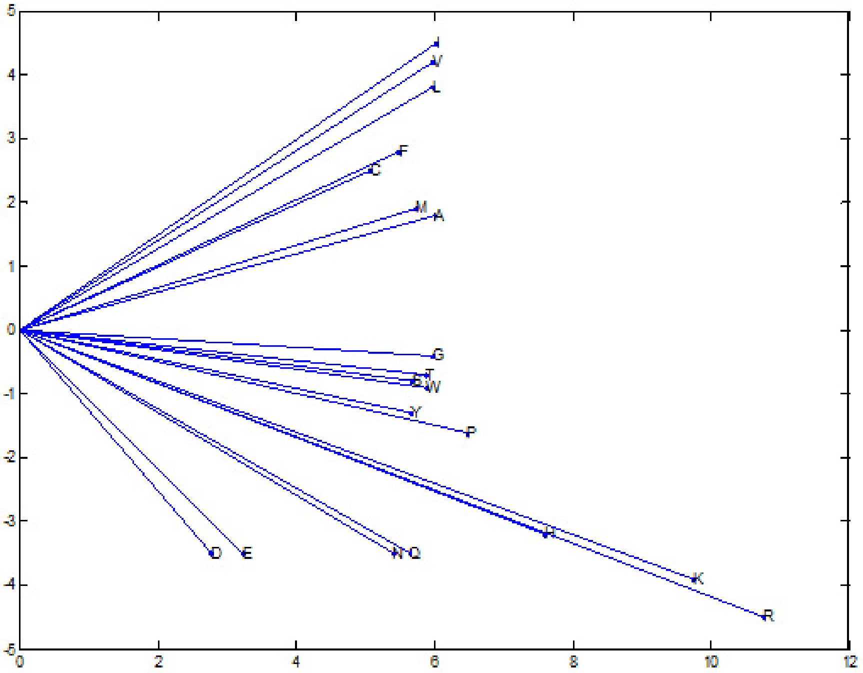

Figure 1: Twenty special vectors in 2-D Cartesian coordinates representing 20 amino acids, respectively.

If point $(0, 0)$ is taken as the origin, Figure 1 illustrates 20 special vectors representing 20 amino acids in a 2-D Cartesian coordinates, respectively. The numerical values in the last two columns of Table 1 are the x and y coordinates for 20 amino acids along the x-axis representing the isoelectric point and along the y-axis representing the hydrpathy index, respectively.

Hence, each amino acid is numerically characterized by a unique special vector. Obviously, the numerical values of x and y coordinates of each amino acid are identical to the numerical values of the isoelecric point and hydropathy index, respectively, which could reflect the innate information of protein molecular, and have no arbitrariness. Similar to our previous work [25], the points in the protein graph are obtained by the sum of vectors representing amino acids in the protein sequence. Given a protein sequence with N amino acids, $S = S_1 S_2 \cdots S_N$, inspect it by stepping one amino acid at a time. For step i ($i = 1, 2, \dots N$), the point $P_i(x_i, y_i)$ can be constructed as follow:

$$x_i = \sum_{k=1}^{i} S_k^1, \quad y_i = \sum_{k=1}^{i} S_k^2,$$

where S_k^j ($j = 1, 2$) represents the jth component of the vector corresponding to S_k. When i from 1 to N, we have points $P_1, P_2, \dots, P_N$. Connecting the adjacent points, we can obtain a protein graph.

In addition, we can easily prove that our new graph has no circuit or degeneracy in our proposed model [9, 25, 27], as follows: Let the number of amino acid forming a circuit be and let the number of G, A, T, S, P, V, L, I, M, F, Y, W, D, E, N, Q, K., R, H, and C in a circuit be g, a, t, s, p, v, l, i, m, f, y, w, d, e, n, q, k, r, h, and c, respectively. Therefore, we have gG, aA, tT, sS, pP, vV, lL, iI, mM, fF, yY, wW, dD, eE, nN, qQ, kK, rR, hH, and cC form a circuit, the following equation will hold:

$$\begin{aligned}
&g(5.97, -0.4) + a(6.01, 1.8) + t(5.87, -0.7) + s(5.68, -0.8) \\
&+ p(6.48, -1.6) + v(5.97, 4.2) + l(5.98, 3.8) + i(6.02, 4.5) \\
&+ m(5.74, 1.9) + f(5.48, 2.8) + y(5.66, -1.3) + w(5.89, -0.9) \\
&+ d(2.77, -3.5) + e(3.22, -3.5) + n(5.41, -3.5) + q(5.65, -3.5) \\
&+ k(9.74, -3.9) + r(10.75, -4.5) + h(7.59, -3.2) + c(5.07, 2.5) = (0, 0).
\end{aligned}$$

The sum of x-coordinates indicates that

$$\begin{aligned}
&5.97g + 6.01a + 5.87t + 5.68s + 6.48p + 5.97v + 5.98l + 6.02i + 5.74m \\
&+ 5.48f + 5.66y + 5.89w + 2.77d + 3.22e + 5.41n + 5.65q + 9.74k \\
&+ 10.75r + 7.59h + 5.07c = 0.
\end{aligned}$$

It follows that

$$\begin{aligned} g = a = t = s = p = v = l = i = m = f = y = w \\ = d = e = n = q = k = r = h = c = 0 \end{aligned}$$

as the number of amino acids is a nonnegative number. Therefore, no circuit exists in the graph in a nontrivial case where $N > 0$.

Moreover, the relationship between a protein graph and its corresponding protein sequence follows one-to-one, and sequence alignment can be done by simply identifying similar segments of protein graph. In addition, protein sequence can be recovered from its protein graph mathematically without loss of any biological information.

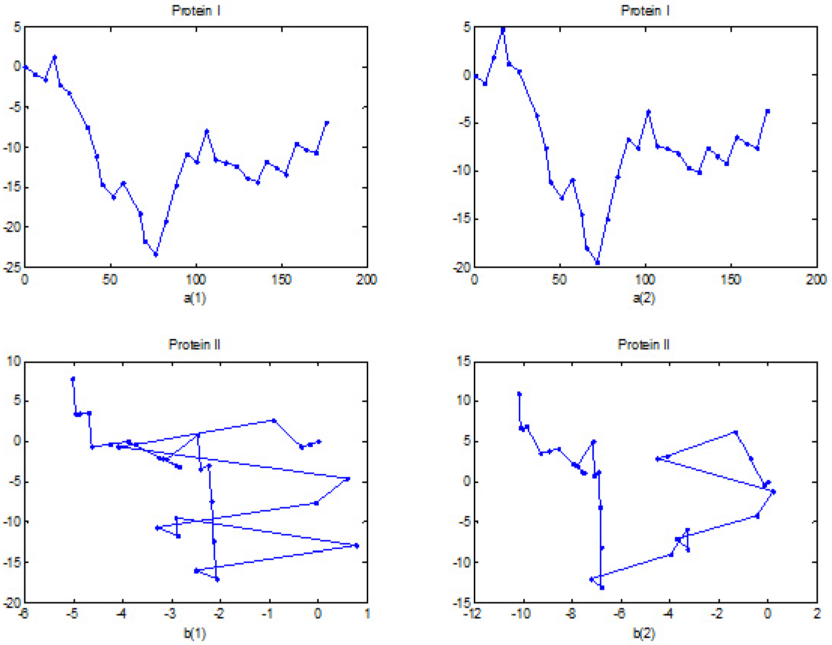

Figure 2: The 2-D graphical representations for segmented proteins I and II from P00729 and P38109, respectively.

We illustrate our graphical representation method on two shorter segments of proteins from yeast (P00729) and Saccharomyces cerevisiae (P38109). These two segments, taken from the Handbook of Chemoniformatics by Randic [22], have been testified in several graphical representations. Figure 2[a(1)] and 2[a(2)] show zigzag curves for these two shorter segments of proteins based on our model and the corresponding protein sequences are

Segmented protein I: WTFESRNDPAKDPVILWLNGGPGCSSLTGL
Segmented protein II: WFFESRNDPANDPIILWLNGGPGCSSFTGL.

Compared with Yu et al.'s protein graphs [26] (Figure 2-[b(1), b(2)]) for the same sequences, our protein graphs are more visible, intuitional, and easily analyzed. But Yu et al.'s protein curves are tedious for having degeneracy, and visual inspection will be difficult for a long sequence.

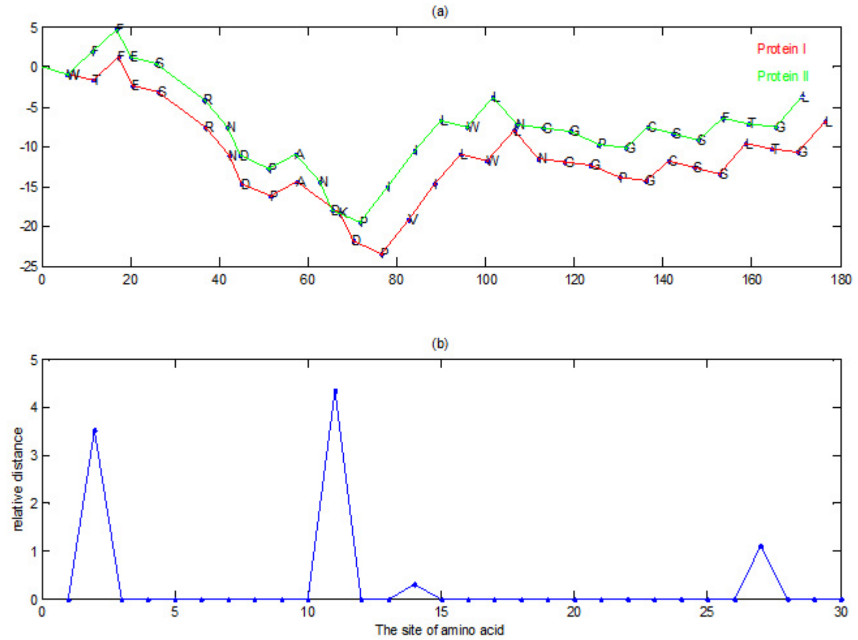

Figure 3: The graphical alignment for segmented proteins I and II from P00729 and P38109, respectively.

Taking a closer look at Figure 2, we also find that two protein graphs are similar on the whole which indicate that they have several local matching segments. In order to quantify the differences between two proteins, we paint two graphical curves in the same palette (see Figure 3(a)). Because each amino acid is numerically characterized by a unique special vector, the same vectors denote with the same amino acid. In Figure 3(a), it is easy to observe the similarities between two protein fragments which have different amino acids only at the sites of 2, 11, 14, and 27. And the relative distance is proposed to depict the diversities of proteins through considering the Euclidean distance between corresponding vectors of amino acids in Figure 3(b). The relative distances at the sites of 2, 11, 14 and 27 are 3.5217, 4.3484, 0.3401 and 1.1180, respectively. Obviously, based on ordering of amino acids with respect to acid dissociation constants of the side chains, the graphical alignments of proteins indicate that not all mismatches are equal. For example, at the site of 14, where Valine in Protein I is substituted by Isoleucine in Protein II, the mismatch is accompanied with a small relative distance. The two amino acids, both classified as 'small hydrophobic', can more easily replace one another than amino acids belonging to different classes. While there are four mismatches between the two proteins, the relative distances associated with small amplitudes are less likely to drastically change the properties of proteins. As said by Randic [22], proteins could be more similar than that would be suggested by the mere count of mismatches.

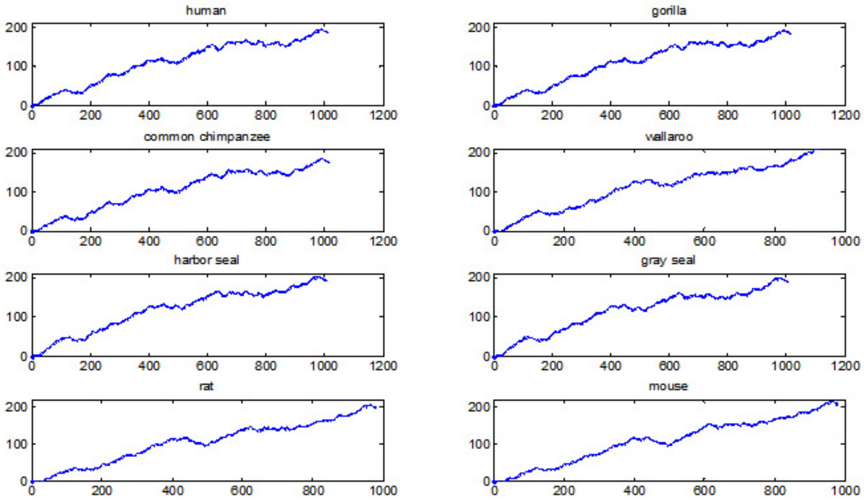

Figure 4: The 2-D graphical representations of eight ND6 protein sequences.

In Figure 4, we illustrate the 2-D graphical representations of ND6 protein sequences from eight different species. Comparing with these curves, we find that the protein graphs of human, gorilla, and common chimpanzee are more similar on the whole, as well as the protein graphs for harbor seal-gray seal and rat-mouse. In addition, the protein graph of wallaroo is obviously different from the other species. These results are consistent with the known fact of evolution and results by other authors [26–34].

2. Numerical characterization of protein graph

In this section, two numerical characterizations of protein graphical curve are utilized to facilitate quantitative comparison of protein sequences. One of the possibilities to achieve this aim is to characterize the graphical curve by mathematical invariants. We first transform the graphical curve into a matrix. Once we have a matrix to represent a sequence, we can use some of matrix invariants as descriptors of protein sequences. Here, we will avoid the difficulties in computational complexity, and introduce two numerical characterizations of such 2-D graphical curve which is easy to obtain.

2.1. Geometrical center of protein graph

According to the graphical curves of proteins, each protein sequence can be represented by a set of vector points in a 2-D Cartesian coordinates. In order to find some invariants that are sensitive to the graphical curve, the geometrical center of the points in corresponding 2-D graphical curve (x^0, y^0) is regarded as a descriptor of characteristic curve [13, 23, 25], which is calculated as follows:

$$x^0 = \frac{1}{N}\sum_{i=1}^{N} x_i, \quad y^0 = \frac{1}{N}\sum_{i=1}^{N} y_i,$$

where x_i and y_i are the coordinates of the ith amino acid in Cartesian coordinates with the point $(0, 0)$ as the origin of all the sequences, and N is the length of the sequence.

2.2. Leading eigenvalue of a graphical curve matrix

The eigenvalues of a graphical curve matrix is a graph invariant representing the characterizations of original protein graph which can better reflect upon the inherent properties of protein sequence. Especially, the leading eigenvalue of matrix has been widely used in the comparison of protein sequences [7, 13, 16, 19, 20, 23, 35–37].

Species	Human	Gorilla	C.Chimpanzee	Wallaroo	H.Seal	G.Seal	Rat	Mouse
(length)	(174)	(174)	(174)	(167)	(175)	(175)	(172)	(172)
x^0	508.5879	509.8840	509.7814	487.3193	509.3739	509.3193	497.1198	497.0613
y^0	113.5316	112.1448	105.8448	117.3629	120.3017	117.0760	106.6884	111.1767
λ_{max} (1.0×10^7)	1.4918	1.5004	1.5016	1.2984	1.4794	1.4739	1.3908	1.3964

Table 2: The geometrical centers (x^0, y^0) and leading eigenvalues λ_{max} of 2-D graphical curves representing eight ND6 protein sequences.

The numerical characterization of the graphical curve is given by matrix M, as follows:

$$M = \begin{pmatrix} M_{xx}, & M_{xy} \\ M_{yx} & M_{yy} \end{pmatrix},$$

whose components are defined as:

$$M_{xy} = M_{yx} = \sum_{i=1}^{N}(x_i - x^0)(y_i - y^0),$$

$$M_{xx} = \sum_{i=1}^{N}(x_i - x^0)^2, \quad M_{yy} = \sum_{i=1}^{N}(y_i - y^0)^2.$$

The above four numbers give a quantitative description of protein graph. The leading eigenvalue λ_{max} of matrix M is used as a descriptor to numerically characterize corresponding protein graph, and further applied to the similarity analysis of protein sequences.

In Table 2, we list geometrical centers (x^0, y^0) and leading eigenvalues λ_{max} gotten from the graphical curves of eight ND6 proteins, respectively.

3. Similarities of eight DN6 protein sequences

In this section, the geometrical center of protein graph and leading eigenvalue of protein graphical curve matrix are utilized to compare similarities of eight DN6 protein sequences, respectively.

If each geometrical center of protein graph represents corresponding protein sequence in a 2-D Cartesian coordinates, the protein map corresponding to eight DN6 proteins is shown in Figure 5. Using the relative distance between two points as an index for comparison, we perform similarity analysis

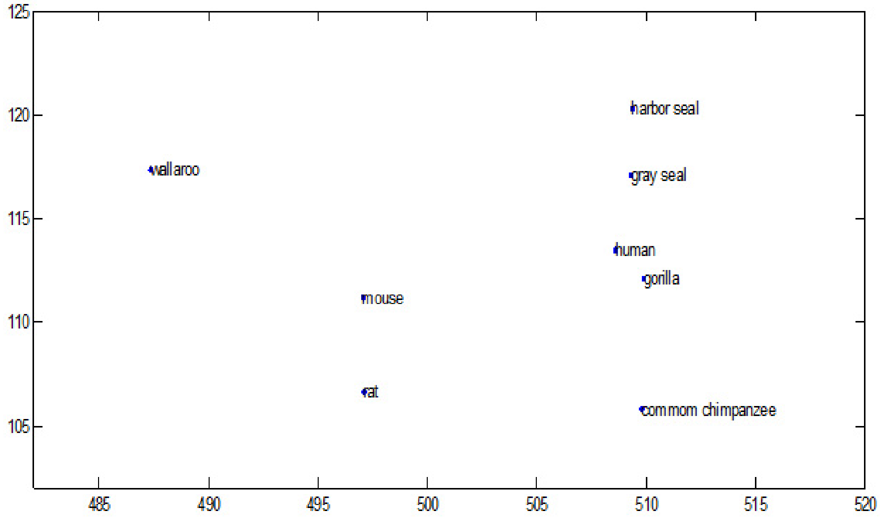

Figure 5: The protein map of eight ND6 protein sequences.

of eight DN6 proteins on protein map. If two protein sequences are more similar, the relative distance of corresponding points should be smaller. Observing Figure 5, we can easily find that the proteins of human, gorilla and common chimpanzee are more similar with each other, although the common chimpanzee seems a little far from human and gorilla. In addition, the groups of (harbor seal, gray seal) and (rat, mouse) are also similar. However, the wallaroo is far from other species. These results coincide with former results in Figure 4. It is obvious that the protein map reflecting the similarities of protein sequences is more visible and intuitional, and the obtaining results are also credible.

Meanwhile, the similar results are also obtained by the leading eigenvalue of protein graphical curve matrix. In Table 3, the similarity matrix for eight DN6 proteins is constructed by computing the absolute value of the difference between any two leading eigenvalues listed in Table 2. The similarity comparison presented by the numerical value is based on the assumption that two proteins are more similar and their corresponding value should be smaller. That is to say, the smaller is the value, the more similar are the two proteins. From Table 3, the proteins of human, gorilla, and common chimpanzee are more similar to each other, as well as DN6 proteins of harbor seal-gray seal and rat- mouse. On the other hand, the wallaroo is dissimilar to others among the eight species. The results of Table 3 are more consistent to the known fact of evolution and former results [26–34].

Species	Human	Gorilla	C.Chimpanzee	Wallaroo	H.Seal	G.Seal	Rat	Mouse
Human	0	0.0858	0.0974	1.9341	0.1248	0.1797	1.0102	0.9546
Gorilla		0	0.0116	2.0199	0.2105	0.2654	1.0960	1.0404
C.Chimpanzee			0	2.0315	0.2222	0.2771	1.1076	1.0520
Wallaroo				0	1.8093	1.7544	0.9239	0.9795
H.Seal					0	0.0549	0.8854	0.8298
G.Seal						0	0.8305	0.7749
Rat							0	0.0556
Mouse								0

Table 3: The similarity matrix (1.0×10^6) for eight ND6 protein sequences.

4. Conclusions

Integrating a selected pair of physicochemical properties of amino acids, the isoelectric point and hydrophy index, we propose a new graphical representation of protein sequences to depict the features of protein sequence, which can precisely reflect the innate information of protein molecular. Comparing with former methods, our new graphical representation cannot only visualize the local and global features of protein sequence, but also effectively avoid the loss of information for graph degeneracy. Importantly, our protein graph is more visible, intuitional, and having no arbitrariness. Associating with our new graphical representation, two numerical characterizations deriving from protein graphical curve are utilized as descriptors to characterize protein sequence which facilitate the similarity comparison of sequences, in that, their computational complexity is very low, and it can be easy to implement. Moreover, the results obtained from the similarity analysis of protein sequences demonstrate that the geometrical center of protein graph and leading eigenvalue of the graphical curve matrix are powerful in precisely reflecting the relationship of protein sequences.

Acknowledgements

The authors thank the references for many valuable comments that have improved this manuscript. This paper is supported by Youth Funding of Suihua University (K1501006), Scientific Research Funding of Suihua University (K1501009, 2017-XGYYWF-017), Scientific Research Funding of Heilongjiang Education Department (2017-KYYWF-0721), and Science and Technology Bureau of Suihua (SHKJ2015-020).

References

[1] E. Hamori and J. Ruskin, *H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences*, Journal of Biological Chemistry **258** (1983), no. 2, 1318–1327.

[2] M. A. Gates, *A simple way to look at DNA*, Journal of Theoretical Biology **119** (1986), no. 3, 319–328.

[3] A. Nandy, *A new graphical representation and analysis of DNA sequence structure: I. Methodology and application to globin genes*, Current Science (1994), 309–314.

[4] P. M. Leong and S. Morgenthaler, *Random walk and gap plots of DNA sequences*, Bioinformatics **11** (1995), no. 5, 503–507.

[5] A. Nandy and P. Nandy, *Graphical analysis of DNA sequence structure: II. Relative abundances of nucleotides in DNAs, gene evolution and duplication*, Current science (1995), 75–85.

[6] M. Randić, M. Vračko, N. Lerš, and D. Plavşić, *Novel 2-D graphical representation of DNA sequences and their numerical characterization*, Chemical Physics Letters **368** (2003), no. 1-2, 1–6.

[7] M. Randić, M. Vracko, A. Nandy, and S. C. Basak, *On 3-D graphical representation of DNA primary sequences and their numerical characterization*, Journal of chemical information and computer sciences **40** (2000), no. 5, 1235–1244.

[8] X. Guo, M. Randic, and S. C. Basak, *A novel 2-D graphical representation of DNA sequences of low degeneracy*, Chemical Physics Letters **350** (2001), no. 1-2, 106–112.

[9] S. S.-T. Yau, J. Wang, A. Niknejad, C. Lu, N. Jin, and Y. K. Ho, *DNA sequence representation without degeneracy*, Nucleic Acids Research **31** (2003), no. 12, 3078–3080.

[10] X. Guo and A. Nandy, *Numerical characterization of DNA sequences in a 2-D graphical representation scheme of low degeneracy*, Chemical physics letters **369** (2003), no. 3-4, 361–366.

[11] B. Liao and T. M. Wang, *New 2D graphical representation of DNA sequences*, Journal of Computational Chemistry **25** (2004), no. 11, 1364–1368.

[12] B. Liao, *A 2D graphical representation of DNA sequence*, Chemical Physics Letters **401** (2005), no. 1-3, 196–199.

[13] B. Liao, M. Tan, and K. Ding, *Application of 2-D graphical representation of DNA sequence*, Chemical Physics Letters **414** (2005), no. 4-6, 296–300.

[14] Y. H. Yao and T. M. Wang, *A class of new 2-D graphical representation of DNA sequences and their application*, Chemical Physics Letters **398** (2004), no. 4-6, 318–323.

[15] Y. H. Yao, X. Y. Nan, and T. M. Wang, *Analysis of similarity / dissimilarity of DNA sequences based on a 3-D graphical representation*, Chemical Physics Letters **411** (2005), no. 1-3, 248–255.

[16] M. Randić, J. Zupan, and A. T. Balaban, *Unique graphical representation of protein sequences based on nucleotide triplet codons* Chemical Physics Letters **397** (2004), no. 1-3, 247–252.

[17] M. Randić, D. Butina, and J. Zupan, *Novel 2-D graphical representation of proteins*, Chemical Physics Letters **419** (2006), no. 4-6, 528–532.

[18] M. Randić, J. Zupan, and D. Vikić-Topić, *On representation of proteins by star-like graphs*, Journal of Molecular Graphics and Modelling **26** (2007), no. 1, 290–305.

[19] M. Randić, M. Novič, D. Vikić-Topić, and D. Plavšić, *Novel numerical and graphical representation of DNA sequences and proteins*, SAR and QSAR in Environmental Research **17** (2006), no. 6, 583–595.

[20] F. Bai and T. Wang, *On graphical and numerical representation of protein sequences*, Journal of Biomolecular Structure and Dynamics **23** (2006), no. 5, 537–545.

[21] C. Li, L. Xing, and X. Wang, *2-D graphical representation of protein sequences and its application to coronavirus phylogeny*, BMB Rep **41** (2008), no. 3, 217–222.

[22] M. Randić, *WITHDRAWN: 2-D graphical representation of proteins based on physico-chemical properties of amino acids*, Chemical Physics Letters **444** (2007), no. 1-3, 176–180.

[23] Y. H. Yao, Q. Dai, C. Li, P. A. He, X. Y. Nan, and Y. Z. Zhang, *Analysis of similarity/dissimilarity of protein sequences*, Proteins: Structure, Function, and Bioinformatics **73** (2008), no. 4, 864–871.

[24] P. A. He, Y. P. Zhang, Y. H. Yao, Y. F. Tang, and X. Y. Nan, *The graphical representation of protein sequences based on the physicochemical properties and its applications*, Journal of Computational Chemistry **31** (2010), no. 11, 2136–2142.

[25] J. Wen and Y. Zhang, *A 2D graphical representation of protein sequence and its numerical characterization*, Chemical Physics Letters **476** (2009), no. 4-6, 281–286.

[26] J. F. Yu, X. Sun, and J. H. Wang, *A novel 2D graphical representation of protein sequence based on individual amino acid*, International Journal of Quantum Chemistry **111** (2011), no. 12, 2835–2843.

[27] C. Yu, S. Y. Cheng, R. L. He, and S. S.-T. Yau, *Protein map: an alignment-free sequence comparison method based on various properties of amino acids*, Gene **486** (2011), no. 1, 110–118.

[28] C. Yu, M. Deng, S. Y. Cheng, S. C. Yau, R. L. He, and S. S.-T. Yau, *Protein space: a natural method for realizing the nature of protein universe*, Journal of Theoretical Biology **318** (2013), 197–204.

[29] S. S.-T. Yau, W. G. Mao, M. Benson, and R. L. He, *Distinguishing proteins from arbitrary amino acid sequences*, Scientific Reports **5** (2015), p. 7972.

[30] C. Yin and S. S.-T. Yau, *A coevolution analysis for identifying protein-protein interactions by Fourier transform*, PloS one **12** (2017), no. 4, p.e0174862.

[31] Y. Li, K. Tian, C. Yin, R. L. He, and S. S.-T. Yau, *Virus classification in 60-dimensional protein space*, Molecular phylogenetics and evolution **99** (2016), 53–62.

[32] M. Li, J. H. Badger, X. Chen, S. Kwong, P. Kearney, and H. Zhang, *An information-based sequence distance and its application to whole mitochondrial genome phylogeny*, Bioinformatics **17** (2001), no. 2, 149–154.

[33] H. H. Otu and K. Sayood, *A new sequence distance measure for phylogenetic tree construction*, Bioinformatics **19** (2003), no. 16, 2122–2130.

[34] V. Makarenkov and F. J. Lapointe, *A weighted least-squares approach for inferring phylogenies from incomplete distance matrices*, Bioinformatics **20** (2004), no. 13, 2113–2121.

[35] M. Randić, *On graphical and numerical characterization of proteomics maps*, Journal of Chemical Information and Computer Sciences **41** (2001), no. 5, 1330–1338.

[36] J. Song and H. Tang, *A new 2-D graphical representation of DNA sequences and their numerical characterization*, Journal of Biochemical and Biophysical Methods **63** (2005), no. 3, 228–239.

[37] Ž. Bajzer, M. Randić, D. Plavšić, and S. C. Basak, *Novel map descriptors for characterization of toxic effects in proteomics maps*, Journal of Molecular Graphics and Modelling **22** (2003), no. 1, 1–9.

School of Agriculture and hydraulic Engineering
Suihua University, Suihua 152061, China
E-mail address: alice_yuyan@163.com

School of Information Engineering
Suihua University, Suihua 152061, China
E-mail address: wenjia198021@126.com