Research Article

Similarity/ Dissimilarity and Phylogenetic Analysis of Protein Sequences

Sanjay Sharma*, Tazid Ali

Department of Mathematics, Dibrugarh University, Dibrugarh, Assam, INDIA.

Submission Date: 22-11-2021; Revision Date: 10-12-2021; Accepted Date: 02-01-2022.

ABSTRACT

In this paper, we first arrange the twenty essential amino acids in descending order according to their degeneracy numbers and following the arrangement we denote each as twenty 2D component vectors confined only to the first quadrant. We illustrated the protein sequences as a curve in 2D space by linking together the vectors representing the amino acids in the protein sequence. The proposed representation is then tested on the ND6 (NADH dehydrogenase subunit 6) protein sequences taken from eight different species for analyzing their similarity using a mathematical descriptor called a similar factor and similar matrix. We have seen that our technique produces a better phylogeny that is quite compatible with previously published results on the same data set. The statistical analysis shows that our approach has better correlations with the multiple sequence alignments.

Key words: Degeneracy, Characteristics vector, Phylogeny, Common logarithm, Correlation coefficient.

Correspondence:
Mr. Sanjay Sharma,
Dibrugarh University,
Dibrugarh-786004,
Assam, INDIA.

Email: snjyshrma90@gmail.com
ORCID: 0000-0002-0689-2244

INTRODUCTION

One of the challenges confronting bio-scientists is the mathematical analysis of huge volumes of genomic DNA sequence data. As a result, more mathematical approaches are being used in gene study.[1] A two-dimensional graphical representation of DNA sequences reveals local and global features, as well as the occurrences, variances, and repetition of nucleotides throughout a sequence that are difficult to see directly from DNA sequences.[2] Graphical methods, pioneered by Hamori[3] have proven to be an effective tool for visualizing and analyzing lengthy DNA sequences.[4-7] Proteins are made up of a linear array of amino acids linked by covalent peptide bonds. The primary structure refers to the amino acid sequence that makes up a protein. The amino acid sequence determines the protein's three-dimensional, functional structure i.e. conformation. The amount of known protein sequences in various databases has exploded due to advances in sequencing methods. Protein sequences are recorded in a computer database system as lengthy character strings, with one character representing each amino acid. By reading these sequences directly, it is difficult to extract any characteristics. As a result, a variety of techniques have been developed to analyze the protein sequences. Protein sequence comparison is used to detect the similarities and differences between various protein sequences, as well as to establish connections between proteins that haven't shared an ancestor in billions of years. It also allows comparing the structure and function of newly discovered proteins, as similar sequences are expected to have similar structures.[8] Alignment methods have typically been used to compare protein sequences. A scoring formula (PAM or BLOSUM matrix) is used to quantify the likelihood of amino acid addition, deletion, and replacement in the compared protein sequences.[9] The alignment of protein sequences may be determined with the aid of this scoring function. The alignment approaches, however, entail a significant computational cost, alignment-free graphical

SCAN QR CODE TO VIEW ONLINE

www.ajbls.com

DOI: 10.5530/ajbls.2022.11.6

representation contributes equally to outcomes and has a very low processing cost. Gupta *et al*. presented a novel two-dimensional graphical representation of protein sequences. Their graphical representation is utilized to create a probabilistic distribution of protein sequences and assess sequence similarity using relative entropy (Kullaback-Leibler divergence) and the suggested approach is tested on ND6 protein sequences from eight different species.[10] Based on the hydrophobicity scale of amino acids, Yao *et al*. provided a 2D spectrum-like graphical depiction of protein sequences. To define a spectrum-like graph, the frequencies of the amplitude of four subsequences are utilized, and a 17D vector is used as the descriptor of the protein sequences.[11] Based on six physical properties of amino acids, Yao *et al*. propose a two-dimensional graphical depiction of protein sequences. Protein sequence descriptors are used to quantitatively describe protein graphs. It's useful for storing intrinsic information about protein structure as well as comparative protein research. The coefficient of determination is provided as a new similarity/dissimilarity measure as an unique simplification.[12] Based on the conditional probability of the protein sequence, [13] presented a novel approach for analyzing the similarity/dissimilarity of protein sequences. The protein sequences of eight species' ND6 (NADH dehydrogenase subunit 6) proteins were used to demonstrate the novel method. The amino acid degeneracy number derived from the standard genetic code is used to create a new 2D representation of the protein sequence in this paper. All the twenty essential amino acids are assigned to the first quadrant only. This allows us for the construction of novel 2D vector representation for the protein sequences. Following that, the mathematical descriptor known as the similar factor is utilized to make a comparison of the NADH dehydrogenase subunit 6 (ND6) proteins from eight distinct species. The results found are in line with prior research. The methods and results are discussed below in sections 3 and 4.

MATERIALS AND METHODS

The numerical characterization of the graphical curve is done by the mathematical descriptor called a similar factor.[14] The numerical characterization opens a novel domain in the comparative study of the protein sequences with the prospective of enhancement on the study of evolution, structure, and functions.

A novel 2D graphical representation of Protein Sequence

The genetic code is stored in the two complementary strands of the DNA molecule as linear and non-overlapping sequences. The genetic code is made up of four nucleotide bases called Adenine (A), Guanine (G), Cytosine (C), and Thymine (T). The three nucleotides together in the row of the genetic code is called a triplet. Each triplet codes for the amino acids. There are 20 essential amino acids and 64 triplets, i.e. more than one triplets code for the same amino acid. So amino acids are degenerate. These triplets are arranged in such a way that they can be read by the cellular machinery, ribosome's which convert them into proteins. The construction of the novel method is described below.

First, the amino acids are arranged in descending order according to their degeneracy number taken from the universal genetic code table.[15] The amino acids that have the same degeneracy numbers are arranged accordingly to alphabetical order which is shown in Table 1. Following the arrangement using the amino acids degeneracy number, we treated the amino acids as novels 2D vector representing different kinds of amino acids. These vectors are equal in their first component and the angle between them and the x-axis are specified as respectively -

3°, 6°, 9°, 12°, 15°, 18°, 21°, 24°, 27°, 30°, 33°, 36°, 39°, 42°, 45°, 48°, 51°, 54°, 57°, 60°

To get a new mathematical descriptor by using these specified angles, it needs to be normalized. The specified angle between amino acids and x-axis is normalized by using the following formula:

$$Y = a + \frac{\left|(X - a1)(b - a)\right|}{(a1 - a2)} \quad (1)$$

where X is the specified angle of each amino acid as stated in Table 1, a1 is the highest specified angle of amino acids Tryptophan (W), which is 60º, and a2 is the minimum specified angle of amino acids Leucine (L), which is 3º. Also and are considered respectively so that the normalized value obtained is between 0 and 1. The normalized value in equation (1) is taken as the Y-coordinates of amino acids shown in Table 1. The X-coordinates are 1 for all the 20 vectors. As the X-coordinates are distinct there is no problem of overlapping.

Figure 1 depicts the vectors corresponding to 20 amino acids. The coordinates in the graphical representation

Table 1: Amino acids with their degeneracy value, assign angle and normalized values (Y-coordinates).

Amino Acids	Degeneracy Value	Assign Angle	Normalized value (Y-coordinate)	Amino Acids	Degeneracy Value	Assign Angle	Normalized value (Y-coordinate)
Leucine (L)	6	3°	0.9	Aspartate (D)	2	33°	0.4789
Arginine (R)	6	6°	0.8578	Glutamate (E)	2	36°	0.4368
Serine (S)	6	9°	0.8157	Phenylalanine (F)	2	39°	0.3947
Alanine (A)	4	12°	0.7736	Histidine (H)	2	42°	0.3526
Glycine (G)	4	15°	0.7315	Lysine (K)	2	45°	0.3105
Proline (P)	4	18°	0.6894	Asparagine (N)	2	48°	0.2684
Valine (V)	4	21°	0.6473	Glutamine (Q)	2	51°	0.2263
Threonine (T)	4	24°	0.6052	Tyrosine (Y)	2	54°	0.1842
Isoleucine (I)	3	27°	0.5631	Methionine (M)	1	57°	0.1421
Cysteine (C)	2	30°	0.5210	Tryptophan (W)	1	60°	0.1

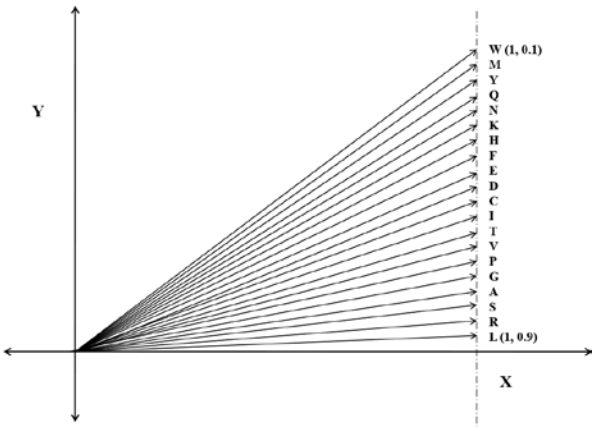

Figure 1: Twenty Vector assign to each of the twenty amino acids.

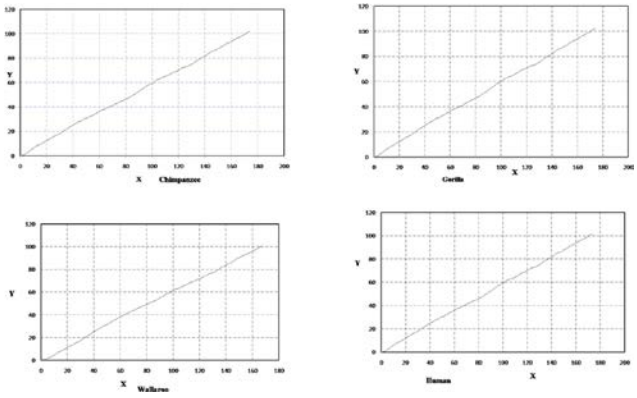

Figure 2: ND6 protein sequences from four species are shown graphically (Common Chimpanzee, Gorilla, Human, and Wallaroo).

of the protein sequence are determined by the sum of vectors representing amino acids. The graph displaying a protein sequence does not create a circuit as it advances along the positive x-axis in an increasing manner. As a result, there is no such thing as overlapping. Otherwise, the protein sequence will not be recovered in the event of overlapping.

The ND6 proteins of the common chimpanzee, gorilla, human, and wallaroo are presented in Figure 2. based on the vector system described in Figure 1. The Y-coordinates are the cumulative y-values in Table 1's fourth column, and the X-coordinates are the number of amino acids in the protein sequence. From Figure 2 we can visualize that the change in the amino acids results in sharp changes in graphs. Since humans, chimpanzees, and gorillas are all members of the same ape family, their graphs are more identical than wallaroo. From Figure 2 we can notice that X-coordinates at 100^{th}, 120^{th} and 160^{th} in the graph of chimpanzee, gorilla, and human their corresponding y-coordinates values are less than 60, 71, and 95 respectively, whereas y-coordinate of wallaroo is greater than 60, 71, and 95 respectively. The graphical representation has been shown to give numerous visual hints to understand the underlying characteristic in protein sequences and can be useful in highlighting the genetic similarity between various protein sequences. However, unless the genetic similarity is determined numerically, the graphical representation is ineffective. As a result, for a comparative analysis of genetic similarities, numerical characterization of the protein sequence is essential, which has been discussed in the next section.

Quantitative measurement of Protein Sequences

Traditional distance equations may be used to calculate the quantitative measurement of protein sequences such as the standard Euclidean distance, Kullback-Leibler distance.[16] Mahalanobis' distance,[17] a geometric measure such as the cosine of the angle between the count vectors[18] and statistical measure

such as the correlation coefficient.[19] Here we have taken *Similar Factor*[14] as a mathematical descriptor to numerically characterize the protein sequences. Let *P*1 and *P*2 be two arbitrary protein sequences, and $\{\vec{V}_1^{\,1}, \vec{V}_2^{\,1}, \vec{V}_3^{\,1} \ldots \vec{V}_n^{\,1}\}$ and $\{\vec{V}_1^{\,2}, \vec{V}_2^{\,2}, \vec{V}_3^{\,2} \ldots \vec{V}_n^{\,2}\}$ were respectively the corresponding sets of characteristics vectors, where was the length for the protein sequences. For example, for an arbitrary protein sequence *WTFESRNDPAKDPVILWLNGGPGCSSLTGL,* the corresponding set of characteristic vectors is

$\{\vec{V}_W\ \vec{V}_T\ \vec{V}_F\ \vec{V}_E\ \vec{V}_S\ \vec{V}_R\ \vec{V}_N\ \vec{V}_D\ \vec{V}_P\ \vec{V}_A\ \vec{V}_K$

$\vec{V}_D\ \vec{V}_P\ \vec{V}_V\ \vec{V}_I\ \vec{V}_L\ \vec{V}_W\ \vec{V}_L\ \vec{V}_N\ \vec{V}_G\ \vec{V}_G\ \vec{V}_P$

$\vec{V}_G\ \vec{V}_C\ \vec{V}_S\ \vec{V}_L\ \vec{V}_T\ \vec{V}_G\ \vec{V}_L\}$

The *Similar Factor* between any pair of sequences is defined as

$$SF = \sum_{i=1}^{n}\left(1 - \frac{\left|f(\vec{V}_i^{\,1}) - f(\vec{V}_i^{\,2})\right|}{90°}\right) \quad (2)$$

where the function $f(\vec{V}_i^{\,-k})(k = 12)$ represents the angles between the x-axis and the th characteristics vector defined in Table 1. The larger the *Similar Factor* is, the more similar the two protein sequences are. For any two different and identical sequences, *SF* is not equal to zero forever as all the species arose from a single molecule and are interrelated with some or more evolutionary relationships. The *SF* measure is relatively dependent on the assigned values of angles chosen to represent each of the twenty amino acids.

We define a similarity matrix *M* for comparing the genetic similarity and difference of *N* different protein sequences denoted as *S*1, *S*2, ... *SN* whose elements m_{ij} are calculated as

$$m_{ij} = \sum_{i=1}^{n}\left(1 - \frac{\left|f(\vec{V}_k^{\,i}) - f(\vec{V}_k^{\,j})\right|}{90°}\right) \quad (3)$$

where $\vec{V}_k^{\,i}$ and $\vec{V}_k^{\,j}$ represent the k^{th} characteristics vector of the i^{th} and k^{th} proteins sequences, m_{ij} is the similar factor between the i^{th} and k^{th} protein sequences. The main diagonal elements are different from each other and $m_{ij} = m_{ji}$, so the similarity matrix *M* is a symmetric matrix. Distance matrix associated with the similarity matrix *M* denoted by *D* whose elements d_{ij} are calculated as

$$d_{ij} = \log_{10}(m_{ij}) \quad (4)$$

The distance matrix is also symmetrical. The larger the distance is, the more similar two protein sequences are and have a close evolutionary or structural relationship. From the distance matrix, we construct a dendrogram using the PAST software.[20]

Dataset

The following eight ND6 protein sequences were used to test the efficiency of our method: Table 2.

All these protein sequences are downloaded from GenBank (http://www.ncbi.nlm.nih.gov).

RESULTS

The phylogeny of eight ND6 proteins was constructed using a new graphical representation of the protein sequence and its similarity factor. The distance matrix for the protein sequences is presented in Table 3 and is calculated by taking the common logarithm of the similarity factor value of the two sequences. The greater the value in the matrix, the more closely the species are related. The identical protein sequences in the main diagonal of the matrix have higher value showing higher similarity. We can observe that species like humans, gorilla and chimpanzees are more closely related to each other than other species. The dendrogram obtained from the distance matrix using PAST software is depicted in Figure 3. From the dendrogram, we can observe that humans, gorilla and chimpanzees fall in same cluster and are closely related. Similarly, mouse –rat and H.seal-G.seal fall in two different clusters predicting that they have a recent close ancestor. As illustrated in Figure 3, a single cluster of wallaroos indicates that it is the most different among the given species. The dendrogram obtained is in agreement with prior research[10,13,12,21] on ND6 protein sequences. In Table 4, we compare our findings to previously published findings on the degree of similarity between humans and other animals. All of the authors utilized the same

Table 2: The information of ND6 protein sequences of eight species.

Species	Acession Number
Human	AP_000650
Gorilla	NP_008223
Commom Chimpanzee	NP_008197
Harbor seal	NP_006939
Gray seal	NP_007080
Rat	AP_004903
Mouse	NP_904339
Wallaroo	NP_007405

The same data set has been used by.[10,13,12,21]

Table 3: The distance matrix for eight ND6 protein sequences.

Species	Human	Gorilla	Chimpanzee	H.seal	G.seal	Rat	Mouse	Wallaroo
Human	2.24054	2.23720	2.23695	2.14134	2.14061	2.16613	2.17240	2.12428
Gorilla		2.24054	2.23594	2.14176	2.14124	2.16554	2.17026	2.12472
Chimpanzee			2.24054	2.13998	2.13924	2.16682	2.17016	2.12199
H.seal				2.24303	2.24146	2.12990	2.12990	2.16385
G.seal					2.24303	2.12893	2.12893	2.16375
Rat						2.23552	2.21642	2.12276
Mouse							2.23552	2.11859
Wallaroo								2.22271

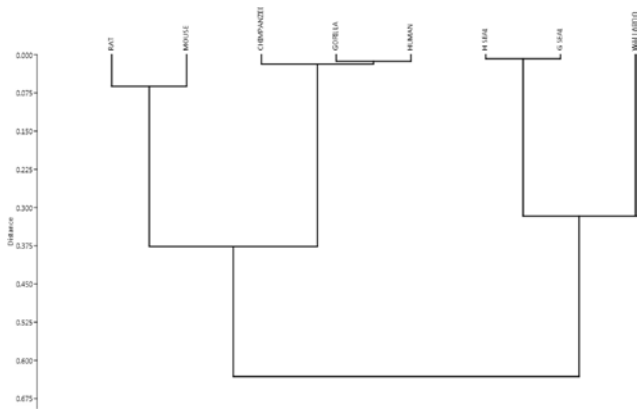

Figure 3: Dendrogram by PAST software based on the distance matrix in Table 3.

data set as us, with the exception of Randic, who used opossums instead of wallaroos to ensure consistency in comparison.

To show the efficiency of our proposed method, we further make the comparison on the same data set of ND6 protein sequences using MEGA-X (Molecular Evolutionary Genetics Analysis) software to obtain a phylogenetic tree of eight species. It gives a clear pictorial view of the evolutionary relationship. The phylogeny of the eight species shown in Figure 3 and Figure 4 are similar and almost in agreement with each other.

Based on our mathematical method we measure the distances between various species and compared our result with multiple sequence alignment method. Clustal W is an online tool which is widely used for the alignment of the nucleotides and protein sequences. The percentage identity matrix PIM(%) of eight ND6 protein sequences obtained by Clustal W 12.1 is shown below in Table 5.

The co-efficients of correlation, between each row of PIM(%) matrix in Table 5 and our result in Table 3 is calculated. Similary, we also calculate the co-efficients of correlation between the results of references[10,12,13,21] and the PIM(%) to compare with our result, which are listed on Table 6. The results in Table 6 shows that our results has higher positive coefficients of correlation with clustal W for all the species than the other compared results.

DISCUSSION

Every gene sequence has an identity, which we attempt to protect by providing a numerical characterization to the sequence. In this context, the question of sequence similarity/dissimilarity and distance between the sequences naturally arises. The notion that the same gene from various species shares a substaintial amount of information in their protein coding sequence and hence leads to significant homology is one of the main principles behind mathematical characterizatiom of protein sequences. Gupta *et al.* measures the similarity/ dissimilarity of ND6 protein sequences using a mathematical descriptor called the symmetric Kullbech-Leibler divergence and obtain a distance matrix between each pair of the protein sequence and concluded that evolutionary closely related species are expected to have small seperations compared to evolutionary disparate groups.[10] Xiao *et al.* analyse the protein sequence based on hydrophaty profile of amino acids by defining a 9D vector to each ND6 protein sequence whose elements are conditional probability of the internal (I), external (E) and ambivalent (A) groups of amino acids. They obtain the distance matrix by calculating the Euclidean between two vectors and their result shows effectual and feasible with evolutionary studies.[13] Yao *et al.* analyze the ND6 protein sequences based on six physicochemical properties of amino acids by transforming each protein to mathematical object called L/L matrix. They define a 6D vectors for each sequence whose elements are the leading eigen values of the L/Lmatrix and obtain distance matrix by calculating the euclidean distance between each pair of vectors. The result found to be match with the evolutionary chronology of the organisms.[12] Randic *et al.* constructed a 20x20 adjacency

matrix based on the selected properties of amino acids for each ND6 protein sequences. They define a 20D vector whose for each sequence whose entries are the main diagonal elements and obtained a distance matrix

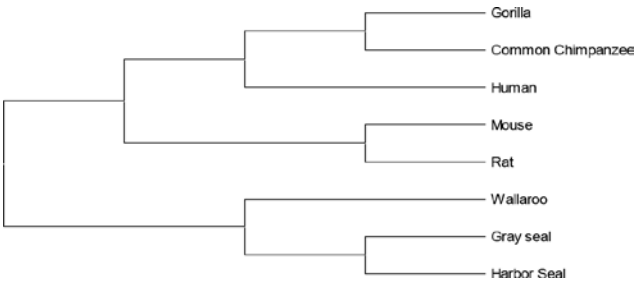

Figure 4: The phylogenetic tree constructed by MEGA-X for eight different ND6 protein sequences.

by calculating the Euclidean distances between pairs of vectors showing that their result reflects strong evolutionary relationship among various pairs of the protein sequence which are true according to the known fact of evolution.[21] In this paper, we analyze the ND6 protein sequences using a mathematical descriptor called *similar factor* and *similar matrix*, which assign each protein sequence a numerical value by maintaing a correct biological geometry. Taking logarithimic value of each entires of similar matrix we obtain a distance matrix and considered the pair to be similar which have largest distance. We observe from the Table 3 that there are three groups (1) primates (gorilla, common chimpanzee and human) are closely related (2) rodents (rat and the mouse) are closely releated (3) carnivorus (H. seal and

Table 4: The published results of the similarity between the coding sequences of different species and the coding sequences of humans were compared.

Species	Gorilla	Chimpanzee	Wallaroo	Opossum	H.seal	G.seal	Rat	Mouse
This work	2.23720	2.23695	2.12428	-	2.14134	2.14061	2.16613	2.17240
From Table 2 in[10]	0.00575	0.0125	0.80224	-	0.0694	0.0736	0.196	0.12756
From Table 3 in[12]	0.0094	0.0118	0.0369	-	0.0247	0.0284	0.033	0.0262
From Table 2 in[13]	0.0338	0.0979	0.278	-	0.1797	0.1487	0.2071	0.1472
From Table 4 in[21]	8.25	6.92	-	16.79	12.81	13.11	14.63	15.03

Table 5 : PIM(%) of eight ND6 protein sequence based on the Clustal W 12.1.

Species	Human	Gorilla	Chimpanzee	H.seal	G.seal	Rat	Mouse	Wallaroo	Opossum
Human	100.00	96.55	95.98	58.62	58.05	50.00	52.91	45.45	42.42
Gorilla		100.00	95.40	58.05	57.47	48.84	51.74	45.45	41.21
Chimpanzee			100.00	57.47	56.90	50.00	52.33	44.85	41.21
H.seal				100.00	97.14	52.91	55.81	45.78	42.17
G.seal					100.00	54.07	56.40	46.99	43.98
Rat						100.00	80.23	42.94	41.72
Mouse							100.00	42.94	41.72
Wallaroo								100	71.86
Opossum									100

Table 6: Results of correlation coefficients for the eight ND6 protein sequences of our approach and the approaches in references,[10,13,12,21] as compared with Clustal W 12.1.

Species	Our approach (Table 3) and PIM %	Reference[10] (Table 2) and PIM%	Reference[13] (Table 2) and PIM%	Reference[12] (Table 3) and PIM%	Reference[21] (Table 4) and PIM%
Human	0.941276	-0.605526	-0.907428	-0.962590	-0.921412
Gorilla	0.942360	-0.582413	-0.908224	-0.974146	-0.934427
Chimpanzee	0.944372	-0.602945	-0.960261	-0.952977	-0.945383
H. seal	0.925073	-0.558593	-0.710608	-0.908080	-0.945289
G. seal	0.932469	-0.550333	-0.852766	-0.885930	-0.939565
Rat	0.881485	-0.634821	-0.770650	-0.828827	-0.912808
Mouse	0.858468	-0.593108	-0.832501	-0.941012	-0.864553
Wallaroo	0.881481	-0.951728	-0.766316	-0.852845	-0.963983

G. seal) are closely related. Based on the results in Table 4 marsupials (wallaroo and opossum), we may conclude that these animals do not have a close evolutionary connection with primates as they are not placental mammalian species. However, despite significant differences, there is a general qualitative agreement between similarities based on various descriptors.

CONCLUSION

The application of mathematical techniques for assessing protein sequences based on the graphical methods allow us for quick and automated ways of assessing the massive amounts of sequence data created everyday. We have shown a graphical and 2D representation of protein sequences based on degeneracy number and vector representation of the amino acids, such that it is free from the problem of overlapping and can be used as a powerful tool for visualizing ND6 protein sequences. The dendrogram obtained from our distance matrix and the phylogenetic tree constructed from MEGA-X are in almost qualitative agreement with each other. In comparison to our results with previously published results, our approach has a strong correlation with multiple sequence alignment methods and this shows the utility of our approach. In this paper, we have only used amino acids degeneracy numbers for the characterization of protein sequences. Many known properties of amino acids along with similar factors can be used for studying the numerical characterization of protein sequences.

ACKNOWLEDGEMENT

The authors would like to thank the Dean, Dibrugarh University, Dibrugarh, Assam valuable suggestions for improving the manuscript.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

REFERENCES

1. Liao B, Zeng C, Li F, Tang Y. Analysis of similarity/dissimilarity of DNA sequences based on dual nucleotides. MATCH Commun Math Comput Chem. 2008;59(3):647-52.
2. Yao YH, Dai Q, Li C, He PA, Nan XY, Zhang YZ. Analysis of similarity/ dissimilarity of protein sequences. Proteins. 2008;73(4):864-71. doi: 10.1002/ prot.22110, PMID 18536018.
3. Hamori E, Ruskin J. H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences. J Biol Chem. 1983 Jan;258(2):1318-27. doi: 10.1016/S0021-9258(18)33196-X, PMID 6822501.
4. Hamori E. Novel DNA sequence representations. Nature. 1985;314(6012):585-6. doi: 10.1038/314585a0, PMID 3990794.
5. Gates MA. A simple way to look at DNA. J Theor Biol. 1986;119(3):319-28. doi: 10.1016/s0022-5193(86)80144-8, PMID 3016414.
6. Leong PM, Morgenthaler S. Random walk and gap plots of DNA sequences. Comput Appl Biosci. 1995;11(5):503-7. doi: 10.1093/bioinformatics/11.5.503, PMID 8590173.
7. Nandy A, Nandy P. Graphical analysis of DNA sequence structure: II. Relative abundances of nucleotides in DNAs, gene evolution and duplication. Curr Sci. 1995;68:75-85.
8. El-Lakkani A, El-Sherif S. Similarity analysis of protein sequences based on 2D and 3D amino acid adjacency matrices. Chem Phys Lett. 2013;590:192-5. doi: 10.1016/j.cplett.2013.10.032.
9. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A. 1992;89(22):10915-9. doi: 10.1073/ pnas.89.22.10915, PMID 1438297.
10. Gupta MK, Niyogi R, Misra M. A 2D graphical representation of protein sequence and their similarity analysis with probabilistic method. MATCH Commun Math Comput Chem. 2014;72:519-32.
11. Yao Y, Yan S, Xu H, Han J, Nan X, He PA, Dai Q. Similarity/Dissimilarity analysis of protein sequences based on a new spectrum-like graphical representation. Evol Bioinform Online. 2014;10:87-96. doi: 10.4137/EBO. S14713, PMID 25002811.
12. Yao YH, Dai Q, Li L, Nan XY, He PA, Zhang YZ. Similarity/dissimilarity studies of protein sequences based on a new 2D graphical representation. J Comput Chem. 2010;31(5):1045-52. doi: 10.1002/jcc.21391, PMID 19777597.
13. Xie XL, Zheng LF, Yu Y, Liang LP, Guo MC, Song J, Yuan ZF. Protein sequence analysis based on hydropathy profile of amino acids. J Zhejiang Univ Sci B. 2012;13(2):152-8. doi: 10.1631/jzus.B1100052, PMID 22302429.
14. Huang G, Liao B, Li Y, Yu Y. Similarity studies of DNA sequences based on a new 2D graphical representation. Biophys Chem. 2009;143(1-2):55-9. doi: 10.1016/j.bpc.2009.03.013, PMID 19428172.
15. Elzanowski A, Ostell J. Natl Cent Biotechnol Inf (NCBI). 2019.
16. Wu TJ, Hsieh YC, Li LA. Statistical measures of DNA sequence dissimilarity under Markov chain models of base composition. Biometrics. 2001;57(2):441-8. doi: 10.1111/j.0006-341x.2001.00441.x, PMID 11414568.
17. Wu TJ, Burke JP, Davison DB. A measure of DNA sequence dissimilarity based on Mahalanobis distance between frequencies of words. Biometrics. 1997;53(4):1431-9. doi: 10.2307/2533509, PMID 9423258.
18. Stuart GW, Moffett K, Baker S. Integrated gene and species phylogenies from unaligned whole genome protein sequences. Bioinformatics. 2002;18(1):100-8. doi: 10.1093/bioinformatics/18.1.100, PMID 11836217.
19. Fichant G, Gautier C. Statistical method for predicting protein coding regions in nucleic acid sequences. Comput Appl Biosci. 1987;3(4):287-95. doi: 10.1093/bioinformatics/3.4.287, PMID 3134115.
20. Hammer Ø, Harper DA, PAST RPD. Paleontological statistics software package for education and data analysis. Palaeontol Electron. 2001;4(1):9.
21. Randic M, Novic M, Vracko M. On novel representation of proteins based on amino acid adjacency matrix. SAR QSAR Environ Res. 2008;19(3-4):339-49. doi: 10.1080/10629360802085082, PMID 18484502.

Cite this article: Sharma S, Ali T. Similarity/ Dissimilarity and Phylogenetic Analysis of Protein Sequences. Asian J Biol Life Sci. 2022;11(1):39-45.