Journal of Chemical Health Risks

www.jchr.org

JCHR (2025) 15(1), 652-660 | ISSN:2251-6727

Characterization of Primary Protein Sequence based on Rumer's Class of Degeneracy and its Application

Sanjay Sharma[1], Ajoy Hatibaruah[1], Baranitharan Mathalaimuthu[1,2]*, Jeyaparvathi S[2], Amarnath Pandian S[3]

[1]Department of Mathematics, St. Joseph University, Nagaland-797115, India.

[2]Department of Zoology, St. Joseph University, Nagaland-797115, India.

[3]Department of Botany, St. Joseph University, Nagaland-797115, India.

(Received: 16 November 2024 Revised: 20 December 2024 Accepted: 04 January 2025)

KEYWORDS

degeneracy, rumer's class, dichotomous class , correlation coefficient, t-test, significant analysis

ABSTRACT:

Aim: The current rapid accumulation of protein sequences in various biological databases has created a challenge of analyzing their similarity and as a result, inferring evolutionary relationships across organisms.

Materials and methods: Based on Rumer's class of degeneracy, we develop a novel three dimensional (3D) representation of the protein sequence.

Results: Two enhanced mathematical invariants are used to facilitate the quantitative study of the graphical of the protein sequences. The effectiveness of our method is tested by comparing and contrasting the ND5 protein sequences of nine distinct species. A correlation and significance analysis is provided to compare our result with the previous published results.

Conclusion: The statistical finding demonstrates that our proposed method is most significant for analyzing protein sequence similarity.

1. Introduction

The advancement of the sequencing technique causes a rapid increment in the number of biological sequences in various biological databases which have challenged the molecular biologist and bio-informaticians for extracting the essential information effectively and reliably to understand the function and structure of Deoxyribonucleic acid(DNA), Ribonucleic acid (RNA) and protein sequences [1-3]. Many computational methods were developed to study the primary protein sequences where it has been narrated that proteins with identical sequences are usually homologous, meaning they have the same 3D structure and functions [4]. The initial stage in predicting the 3D structure of a protein sequence is to align the sequences. Alignment techniques are divided into two categories: alignment-based and alignment-free. The most frequently used computer tools for alignment-based methods are BLAST (basic local alignment search tool) and ClustalW [5-7]. The results of these algorithms provide a more precise approach to the protein alignment problem. Several alignment-free techniques for sequence comparison have been published. The majority of biological sequence analysis methods have drawbacks, such as low accuracy and lengthy processing periods [8,9]. Not only may graphical representations of protein sequences give a simple method to see the complicated link between protein sequences, but they can also be used to mathematically characterize the similarities between them which has recently become a popular method for protein sequence analysis [10,11]. Protein sequences are usually described using letter sequence representation (LSR). LSR is a letter-based representation of any protein sequence, with each of the 20 amino acids represented by a letter. Because it's difficult to get and compare various sequences using LSR, numerous approaches have been developed to convert protein sequence letters into mathematical representations and then analyze them [12]. He et al., proposed a 3D representation of the protein sequences based on Jeffrey's technique, taking into account the interaction between neighbor amino acids. Then they compare the similarities of two

Journal of Chemical Health Risks

www.jchr.org

JCHR (2025) 15(1), 652-660 | ISSN:2251-6727

matching proteins with a new descriptor that defines the 3D graphical representation of a protein, and a distance between two 3D graphical representations is introduced [1]. Based on 10 physicochemical characteristics of amino acids and the BLOSUM62 matrix, Qi et al., propose a 3-dimensional (3D) graphic representation of protein sequences [2]. Again based on the physical characteristics of amino acid side chains, Abo el Maaty et al., suggested a new 3D graphical representation. Then they numerical analyzed their 3D graph to look for similarities between nine ND5 protein sequences [4].

We attempted a novel 3D graphical representation of protein sequences based on Rumer's class of degeneracy of the amino acids. After obtaining the 3D graph of the protein sequences, we tried to do some numerical characterization by using two different mathematical descriptors (accumulative frequency distances and eigen value of the covariance matrix) to analyze the similarity/dissimilarity of nine distinct species' of ND5 protein sequences. In addition, we do a correlation and significance analysis to compare our findings with the percentage identity matrix(PIM) that we obtained from multiple sequence alignment Clustal Omega 2.1 and found that our result has better significant than previous published result on the same data sets.

2. Materials and Methods

2.1. From DNA to Protein

The genetic code is a biological mechanism that determines how DNA nucleotide sequences are transcribed into mRNA codon sequences, which are then translated into amino acid protein sequences [13]. The chemicals adenine, cytosine, guanine and thymine/uracil that make up the nucleotide bases of DNA are represented by the letters A, C, G, and T/U. A triplet (codon) is made up of three nucleotides in a row and codes for a single amino acid. As a result, each three-letter sequence represents amino acids [14]. The standard genetic code is redundant i.e. 64 codons code for 20 amino acids. For example, amino acid Glycine(G)has degeneracy four, so there exists a subset of four codons $\{GGU, GGC, GGA, GGG\}$. The standard genetic code has a degeneracy group of 1,2,3,4 and 6 as shown table below [15].

Proteins are made up of twenty amino acids (AA) Phenylalanine(P), Leucine(L), Isoleucine(I), Methionine(M), Valine(V), Serine(S), Proline(P), Threonine(T), Alanine(A), Tyrosine(Y), Histidine(H), Glutamine(Q), Asparagine(N), Lysine(K), Aspartic acid(D), Glutamic acid(E), Cysteine(C), Tryptophan(W), Arginine(R), Glycine(G), which are tiny chemical compounds with an amino group, a carboxyl group, a hydrogen atom, and a variable component known as a side chain are all linked to an alpha (central) carbon atom. Peptide bonds are used to connect several amino acids to form a long chain within a protein. The linear amino acid sequence inside a protein is characterized as the main structure, which regulates the folding and intramolecular interactions of the long chain of amino acids and significantly impacting the protein's distinctive three-dimensional (3D) form [16].

2.2. 3D graphical representation of the protein sequences

Yury Borisovich Rumer, a Russian theoretical physicist discovers that half of the genetic code contains a degeneracy of class 4 and the other half contains degeneracy of class 1, 2, 3. Rumer's class is a dichotomous class in which a codon can take the value 4 (first rumor's class) or one of the other values 1, 2, or 3(second rumor's class) [17].

To define Rumer's classes, we look at the first two bases of a codon and follow the following steps.

Step 1. If the middle base of a codon belongs to an amino class (A or C), a C denotes a codon of the first Rumer's class (the coded amino acid or stop signal is uniquely determined by the first two bases) and an A denotes a codon of the second Rumer's class.

Step 2. If a codon's middle base belongs to the keto class (G or U/T), the first base is required.

2.1 If a codon's initial base is strong (C or G), it falls into the first Rumer's class;

2.2 If the first base is weak (A or U/T), it falls into the second Rumer's class [18].

Journal of Chemical Health Risks

www.jchr.org

JCHR (2025) 15(1), 652-660 | ISSN:2251-6727

$$\begin{bmatrix} TCT \\ TCC \\ TCA \\ TCG \end{bmatrix}_S \quad \begin{matrix} CCT \\ CCC \\ CCA \\ CCG \end{matrix}_P \quad \begin{matrix} ACT \\ ACC \\ ACA \\ ACG \end{matrix}_T$$

$$\begin{bmatrix} CTT \\ CTC \\ CTA \\ CTG \end{bmatrix}_L \quad \begin{matrix} GTT \\ GTC \\ GTA \\ GTG \end{matrix}_V \quad \begin{matrix} CGT \\ CGC \\ CGA \\ CGG \end{matrix}_R$$

$$\begin{bmatrix} TAT \\ TAC \end{bmatrix}_Y \quad \begin{matrix} TAA \\ TAG \end{matrix}_{Stop} \quad \begin{matrix} CAT \\ CAC \end{matrix}_H$$

$$\begin{bmatrix} AAT \\ AAC \end{bmatrix}_N \quad \begin{matrix} AAA \\ AAG \end{matrix}_K \quad \begin{matrix} GAT \\ GAC \end{matrix}_L$$

$$\begin{bmatrix} TTT \\ TTC \end{bmatrix}_F \quad \begin{matrix} TTA \\ TTG \end{matrix}_L \quad \begin{matrix} TGT \\ TGC \end{matrix}_C$$

$$\begin{bmatrix} ATT \\ ATC \\ ATA \end{bmatrix}_I \quad ATG_M \quad \begin{matrix} AGT \\ AGC \end{matrix}_S$$

$$\begin{bmatrix} S & P & T & A \\ L & V & R & G \\ Y & Stop & H & Q \\ N & K & D & E \\ F & L & C & W \\ I & M & S & R \end{bmatrix}_{6\times4}$$

Using the concept of Rumer's class of degeneracy we have designed a 6×4 matrix and propose a new 3D representation of the protein sequences. We have arranged the twenty amino acids plus the stop codons according to Rumer's class of degeneracy in each row of the matrix given below.. Every element of the matrix has a corresponding index (i, j) where $i = 1,2,3,4,5,6$ and $j = 1,2,3,4$. Each row of the matrix follows Rumer's class degeneracy. The first row elements are amino acids S, P, T, A which are coded by the codon in which the middle base has amino class (C) so they fall in the first Rumer's class of degeneracy. The second-row elements are amino acids L, V, R, G which are coded by the codon in which the middle base has Keto class (T) but the initial base belongs to a strong class (C) so they fall in the first Rumer's class of degeneracy. The third and fourth row of the matrix has amino acids Y, H, Q, N, K, D, E and stop codon which is coded by the codon in which the middle base have amino class (A) so they fall in the second class of Rumer's degeneracy. The fifth and sixth-row elements are amino acids F, L, C, W, I, M, S, R which have Keto $(T \& G)$ middle base but their initial base have weak class $(T \& A)$ so they fall in the Rumer's second class of degeneracy.

Based on the index of the matrix we assign one pair of coordinates to each of the two amino acids which are given below and considering that the amino acids Serine, Leucine and Arginine which are distributed into Rumer's first and second class of degeneracy, we shall consider the first Rumer's class degeneracy in these three amino acids. That is to say, that we have assigned each of the amino acids to its corresponding index i.e. (y, z) coordinates, while the corresponding curve extending along the x-axes.

In detail, if P is a protein sequence of length n, then there is a one to one correspondence between the protein sequence and the 3D protein graph by using the mapping $\emptyset$, which is defined below. The mapping $\emptyset$ given in Equation 1 reduces the protein sequence P into a series of vectors $p_1, p_2, p_3, \ldots, p_n$ called characteristics plot. The curve connects all the characteristics plot and in turn, we get the 3D graphical representation of the protein sequence. For example, we consider two shorter segments of yeast protein Saccharomyces cerevisiae. From Table 2, Figures 1 and 2 shows their corresponding coordinates and the 3D representations.

Protein I: WTFESRNDPAKDPVILWLNGGPGC-SSLTGL

Protein II: WFFESRNDPANDPIILWLNGGPGC-SSFTGL

From the construction of our 6×4 matrix, we can say that our design is not unique, as each row element can be arranged in 4! ways and each row can be arranged in 6! ways i.e in $4! \times 6!$ ways. We have $4! \times 6!$ 3D curve that can be obtained from the same protein sequences. Here we design the matrix based on Rumer's classification and we only consider the above matrix to illustrate our scheme. The zigzag curve which we obtain doesn't represent the genuine protein molecular structure but we are interested in the numerical parameter that may facilitate the comparison of protein sequences.

$$S(01,1), P(0,1,2), T(0,1,3), A(0,1,4)$$

$$L(0,2,1), V(0,2,2), R(0,2,3), G(0,2,4)$$

$$Y(0,3,1), H(0,3,3), Q(0,3,4), N(0,4,1)$$

$$K(0,4,2), D(0,4,3), E(0,4,4), F(0,5,1)$$

$$C(0,5,3), W(0,5,4), I(0,6,1), M(0,6,2)$$

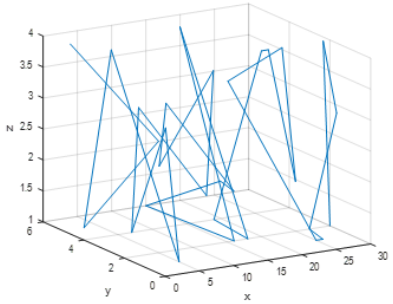

Figure 1 Protein I 3D graphical representation

Journal of Chemical Health Risks
www.jchr.org
JCHR (2025) 15(1), 652-660 | ISSN:2251-6727

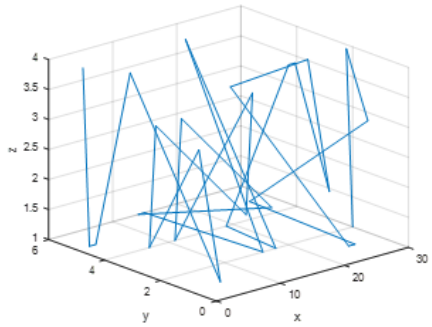

Figure 2 Protein II 3D graphical representation

$$D = \sum_{k=1}^{m} d_i + (n - m) \times \max(d_i) \qquad (1)$$

$$\emptyset(AA) = \begin{cases} (i,1,1) & if\ AA = S, \\ (i,1,2) & if\ AA = P, \\ (i,1,3) & if\ AA = T, \\ (i,1,4) & if\ AA = A, \\ (i,2,1) & if\ AA = L, \\ (i,2,2) & if\ AA = V, \\ (i,2,3) & if\ AA = R, \\ (i,2,4) & if\ AA = G, \\ (i,3,1) & if\ AA = Y, \\ (i,3,3) & if\ AA = H, \\ (i,3,4) & if\ AA = Q, \\ (i,4,1) & if\ AA = N, \\ (i,4,2) & if\ AA = K, \\ (i,4,3) & if\ AA = D, \\ (i,4,4) & if\ AA = E, \\ (i,5,1) & if\ AA = F, \\ (i,5,3) & if\ AA = C, \\ (i,5,4) & if\ AA = W, \\ (i,6,1) & if\ AA = I, \\ (i,6,2) & if\ AA = M, \end{cases} \qquad (1)$$

2.3. Numerical characterization of the Protein sequences

In this section, we give a quantitative characterization of the protein sequences for comparing their similarity by using two different mathematical descriptors. The distance between the two protein sequences is obtained based on their graphical representation according to our method. First, we calculated the Euclidean distances (d_i) between the corresponding amino acids of two different protein sequences followed by accumulative frequency distances from the first distance. Second, we use the matrix invariant (eigen value) of the covariance matrix to check the similarity between the protein sequences. In the next section, using the statistical tool we compare two of our mathematical descriptors with percentage identity matrix (PIM) obtained from multiple sequence alignment Clustal Omega 2.1.

Table 1: The information of ND5 Proteins of Nine species

Species	ID(NCBI)	Sequence length
Human	CAA24036	603
Gorilla	BAA07306	603
Pgymy Chimpanzee	BAA07315	603
Commom Chimpanzee	BAA07302	603
Fin Whale	CAA43449	606
Blue Whale	CAA51005	606
Rat	CAA32964	610
Mouse	CAA24088	607
Opossum	CAA82687	602

For protein sequences of unequal length($n > m$), we calculated the accumulative frequency by the given formula below

where d_i is the Euclidean distance between corresponding amino acids after deleting last $(n - m)$ amino acids bases from the larger sequence and max (d_i) is the maximum Euclidean distance between the corresponding amino acids. For example, in Table 2 in the last two columns, we have listed the Euclidean distances and the accumulative frequency distances and we take 12.59524 as the distance between the two protein sequences. We calculated the similarity/dissimilarity distance matrix of the ND5 protein sequence between nine species shown in Table 3 using Equation 2 shown in Table 4.

Journal of Chemical Health Risks
www.jchr.org
JCHR (2025) 15(1), 652-660 | ISSN:2251-6727

Again, we consider the matrix invariant eigen value of a covariance matrix to calculate the distance between two protein sequences. The covariance matrix for each protein sequence is obtained based on a graphical representation by our method, denoted by Cov(species_name) and is defined below.

$$Cov(species\_name) = \begin{bmatrix} cov(x,x) & cov(x,y) & cov(x,z) \\ cov(y,x) & cov(y,y) & cov(y,z) \\ cov(z,x) & cov(z,y) & cov(z,z) \end{bmatrix} \quad (2)$$

where x_i, y_i, z_i are data value of x, y and y; $\bar{x}, \bar{y}, \bar{z}$ are mean values of x,y and z; N = number of data values and

$$\mathrm{cov}(x,x) = \frac{\sum(x_i-\bar{x})(x_i-\bar{x})}{N-1}, \qquad \mathrm{cov}(y,y) = \frac{\sum(y_i-\bar{y})(y_i-\bar{y})}{N-1},$$

$$\mathrm{cov}(z,z) = \frac{\sum(z_i-\bar{z})(z_i-\bar{z})}{N-1},$$

$$\mathrm{cov}(x,y) = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{N-1} = \mathrm{cov}(y,x)$$

$$\mathrm{cov}(x,z) = \frac{\sum(x_i-\bar{x})(z_i-\bar{z})}{N-1} = \mathrm{cov}(z,x)$$

$$\mathrm{cov}(y,z) = \frac{\sum(y_i-\bar{y})(z_i-\bar{z})}{N-1} = \mathrm{cov}(z,y)$$

The Eigen value vector distance matrix between two different protein sequences of equal and unequal length is calculated and shown in Table 5.

3. Result and Discussion

The analysis of the similarity/dissimilarity distance matrix in Tables 4 and 5 of the ND5 protein sequences of nine species is based on the assumption that the smaller the corresponding distance (accumulative frequency and eigenvalue vector), the more similar the two protein sequences are. From Tables 4 and 5 we can observe that the distance among the species (human, gorilla, common chimpanzee, pygmy chimpanzee) is relatively small. The smallest distance in Tables 4 and 5 is between the Fin whale and Blue whale. Tables 4 and 5 show that the nine species can be divided into four groups: (human, gorilla, common chimpanzee and pygmy chimpanzee), (fin whale and blue whale), (rat and mouse) and opossum. The ND5 protein sequence of the opossum is substantially different from the rest of the eight species because it is the most distant species from the other mammals. In addition, we used Clustal Omega 2.1 program, a multiple sequence alignment, to find the percentage of identity/similarity of the ND5 protein sequences of nine mammalian species and obtain a distance matrix in Table 6.

We compare our results in Tables 4 and 5 with the PIM, we provide a correlation and significance analysis. The better the linear connection, the bigger the correlation coefficient r. Table 7 shows the correlation coefficient between our techniques and the PIM, which is computed from the relevant rows. The correlation coefficient corresponding to the first-row human is $|r| = 0.7094$, corresponding to the second-row gorilla and so on.

From Table 7 we observe that the accumulative frequency and eigen value vector distance matrix has maximum higher coefficient of correlation value with the percentage identity matrix obtained by Clustal Omega 2.1 as compared to the rest of the published result correlation value. Since the data set is small i.e. (n=9) it is easy to produce a high correlation coefficient. Therefore, we performed a significance analysis on the correlation coefficients to see if the connection between

Table 2: Similarity/Dissimilarity Distance Matrix based on the accumulative distance

	Human	Gorilla	P. Chimp	C. Chimp	F Whale	B Whale	Rat	Mouse	Opossum
Human	0	120	125.0990	120	142.9984	142.0984	164.4222	148.3238	131.2161
Gorilla		0	179.8388	173.1091	584.6965	565.42011	694.6413	660.31078	1094.4359
P Chimp			0	85.3082	556.0492	546.0499	677.3440	654.2374	1072.2456
C Chimp				0	555.2570	547.8846	683.7223	648.4901	1085.1565
					0	61.7508	645.1750	627.4859	1067.1617

Journal of Chemical Health Risks
www.jchr.org
JCHR (2025) 15(1), 652-660 | ISSN:2251-6727

F Whale				
B Whale	0	650.2136	623.7781	1065.8267
Rat		0	393.8350	1098.5716
Mouse			0	1073.7460
Opossum				0

Table 3: Similarity/Dissimilarity Distance Matrix based on Eigenvalue vectors

	Human	Gorilla	P. Chimp	C. Chimp	F Whale	B whale	Rat	Mouse	Opossum
Human	0	0.0100	0.0470	0.0084	302.3000	302.3000	707.6000	403.3001	100.4004
Gorilla		0	0.0462	0.0178	302.2000	302.3000	707.6000	403.3002	100.4000
P Chimp			0	0.0540	302.2001	302.3000	707.6000	403.3001	100.4003
C Chimp				0	302.2000 2	302.3000	707.6000	403.3002	100.4005
F Whale					0	0.1006	405.4000	101.1004	402.6000
B Whale						0	405.3000	101.0004	402.7000
Rat							0	304.3000	808.0000
Mouse								0	503.7000
Opossum									0

The result of the percentage identity matrix (PIM) of the ND5 protein sequences obtained from the Clustal Omega 2.1 is illustrated as distance matrix in Table 6 below

two sets of data is high enough or if it is most likely accidental. We perform a t-test for the correlation coefficient $|r| \geq 0.7$ and computed the statistical significance which is given in Table 8. Our sample size is 9 so the degree of freedom is 7. At t-values greater than 2.365 indicates a level of significance less than 0.05 indicating that r value have does not occur by chance. From Table 8, observing the t-values we can say that our two approaches has much more correlation with the PIM as compared with the previous published result and PIM on the same data sets.

4. Conclusion

Applying different mathematical tools, we can compute different comparison techniques for large and unequal protein sequence which may provide evidence for analyzing the functional, structural and evolutionary

Journal of Chemical Health Risks
www.jchr.org
JCHR (2025) 15(1), 652-660 | ISSN:2251-6727

Table 4: Distance matrix of ND5 protein sequences of nine species based on a percentage of similarity

	Human	Gorilla	P. Chimp	C. Chimp	F. Whale	B whale	Rat	Mouse	Opossum
Human	100	90.05	93.20	93.37	68.33	68.16	63.18	63.85	62.71
Gorilla		100	90.88	90.71	67.33	67.50	62.52	63.35	60.87
P Chimp			100	95.02	68.82	68.82	63.18	63.85	62.88
C Chimp				100	68.66	68.66	62.85	63.52	62.04
F Whale					100	96.53	65.51	64.69	61.20
B Whale						100	65.68	65.18	61.04
Rat							100	78.25	60.54
Mouse								100	62.21
Opossum									100

Table 5: The correlation coefficient for ND1 Protein sequences of nine species based on our approaches as compared with PIM

	Our approach (Accumulative frequency) Table 4 and PIM % $(\|r\|\ values)$	Our approach (Eigenvalue) Table 5 and PIM % $(\|r\|value$	Reference [**19**]Table 3 & PIM% $(\|r\|values$	Reference [**19**]Table 4&PIM% $(\|r\|values)$	Reference [**20**]Table 2 and PIM% $(\|r\|values$	Reference [**20**] Table 3and PIM% $(\|r\|value$	Reference [**21**]Table 3 and PIM% $(\|r\|values$	Reference [**21**]Table 4 and PIM% $(\|r\|values)$
Human	0.7094	0.7719	0.9298	0.7342	0.2842	0.1595	0.9211	0.6361
Gorilla	0.9395	0.7567	0.9221	0.7712	0.3530	0.0762	0.7697	0.8401
P Chimp	0.9361	0.7760	0.8608	0.7759	0.2557	0.1830	0.9320	0.8299
C Chimp	0.9231	0.7714	0.9320	0.8004	0.2776	0.1896	0.9307	0.5337
F whale	0.8136	0.8376	0.8099	0.6086	0.5103	0.0834	0.8001	0.3029
B whale	0.8116	0.8429	0.7649	0.6376	0.5069	0.0838	0.7540	0.6194
Rat	0.7268	0.9005	0.6308	0.8701	0.8733	0.7366	0.7010	0.5669
Mouse	0.7042	0.6332	0.6085	0.6086	0.8885	0.7652	0.6820	0.5776
Opossum	0.7250	0.4281	0.5083	0.8123	0.9954	0.9939	0.6135	0.4905

Journal of Chemical Health Risks

www.jchr.org

JCHR (2025) 15(1), 652-660 | ISSN:2251-6727

Table 6: The t-values computed for the correlation coefficient |r|≥0.7, based on which the statistical significance is determined

	Our approach (Accumulative frequency) Table 4 and PIM %	Our approach (Eigen value) Table 5 and PIM %	Reference [**19**]Table 3 & PIM%	Reference [**19**]Table 4&PIM%	Reference [**20**]Table 2 and PIM%	Reference [**20**] Table 3and PIM%	Reference [**21**]Table 3 and PIM%	Reference [**21**]Table 4 and PIM%
Human	2.6630	3.2124	6.6867	2.8613	-	-	6.2626	-
Gorilla	7.2590	3.0631	6.3060	3.2056	-	-	3.1901	4.0980
P Chimp	7.0418	3.2553	4.4749	3.2547	-	-	6.8037	3.9364
C Chimp	6.3536	3.2076	6.8051	3.5327	-	-	6.7346	-
F whale	3.7025	4.0573	3.6533	-	-	-	3.5300	-
B whale	3.6758	4.1457	3.1419	-	-	-	3.0378	-
Rat	2.8002	5.4800	-	4.6711	4.7434	2.8822	2.6010	-
Mouse	2.6248	-	-	-	5.1238	3.1452	-	-
Opossum	2.7853	-	-	3.6850	27.5776	23.8647	-	-

relationship. We proposed a 3D representation of protein sequences based on the twenty amino acids Rumer's class degeneracy and obtained a 3D graphical representation of the protein sequences which provide a powerful tool for quantitative analysis of protein sequence. We then applied our graphical representation of amino acids to check the similarity/dissimilarity of the ND5 protein sequences. We facilitate two improved mathematical descriptors for measuring the distances between two different equal and unequal protein sequences and obtained two different distance matrices. Further, we compare our results with the percentage identity matrix obtained from multiple sequence alignment Clustal Omega 2.1 through a statistical linear correlation and significance analysis. From the statistical result we can concluded that our two approaches are most significant than the previous published results on the same ND5 data sets.

5. Refrences

[1] Ping-an Hea, Jinzhou Weia, Yuhua Yaob, and Zhixin Tie, "A novel graphical representation of proteins and its application," Physica A , vol. 391 , p. 93–99, 2012.

[2] Zhao-Hu Qi, Ke-Cheng Li, Jin-Long Ma, and Yu-Hua Yao, "Novel Method of 3-Dimensional Graphical Representation for Proteins and Its Application," Evolutionary Bioinformatics, vol. 14, p. 1–8, 2018, https://doi.org/10.1177/1176934318777755.

[3] Wenbing Houa, Qiuhui Pana, and Mingfeng Hea, "A novel representation of DNA sequence based on CMI coding," Physica A , vol. 409, pp. 87-96, 2014.

[4] Mervat M. Abo-Elkhier, Marwa A. Abd Elwahaab, and Moheb I. Abo El Maaty, "Measuring Similarity among Protein Sequences Using a New Descriptor," Hindawi BioMed Research International, vol. 2019 2796971, Nov 2019, doi: 10.1155/2019/2796971.

[5] S F Altschul, W Gish, W Miller, E W Myers, and D J Lipman, "Basic local alignment search tool," Journal of molecular biology, vol. 215, no. 3, pp.

403-410, Oct. 1990, https://doi.org/10.1016/S0022-2836(05)80360-2.

[6] S Altschul et al., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," Nucleic Acids Research, vol. 25, no. 17, pp. 3389-3402, Sept. 1997, https://doi.org/10.1093/nar/25.17.3389.

[7] J D Thompson, D G Higgins, and T J Gibson, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.," Nucleic acids research, vol. 22, no. 22, pp. 4673-4680, Nov. 1994.

[8] M Randić, J Zupan, A T Balaban, D Vikić-Topić, and D Plavšić, "Graphical representation of proteins," Chemical reviews, vol. 111, no. 2, pp. 790-862, Feb. 2011.

[9] X Jin et al., "Similarity/dissimilarity calculation methods of DNA sequences: A survey," Journal of Molecular Graphics and Modelling, vol. 76, pp. 342-355, Sept. 2017.

[10] J Song and H Tang, "A new 2-D graphical representation of DNA sequences and their numerical characterization," Journal of biochemical and biophysical methods, vol. 63, no. 3, pp. 228-39, 2005, doi:10.1016/j.jbbm.2005.04.004.

[11] P Echenique, "“Introduction to protein folding for physicists," Contemporary Physics, vol. 48, pp. 81-108, Sept. 2007, https://doi.org/10.1080/00107510701520843.

[12] H Hu and Z Li, "Graphical Representation and Similarity Analysis of Protein Sequences Based on Fractal Interpolation," EEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 14, no. 1, pp. 182-192, Jan-Feb 2017, doi: 10.1109/TCBB.2015.2511731.

[13] R Sanchez, "Symmetric Group of the Genetic-Code Cubes. Effect of the Genetic-Code Architecture on the Evolutionary Process," MATCH Communications in Mathematical and in Computer Chemistry, vol. 79, no. 3, pp. 527-560, 2018.

[14] L C Brody. National Human Genome Research Institute. [Online]. https://www.genome.gov/genetics-glossary/Genetic-Code

[15] J Malpas and Donald Davidson. (2016, March) The Stanford Encyclopedia of Philosophy. [Online]. https://plato.stanford.edu/entries/information-biological/

[16] (2014) Scitable by nature education. [Online]. https://www.nature.com/scitable/topicpage/protein-structure-14122136/

[17] D L Gonzalez, S Giannerini, and R Rosa, "Strong short-range correlations and dichotomic codon classes in coding DNA sequences," Physical review. E, Statistical, nonlinear, and soft matter physics, vol. 78(5 Pt 1):051918, Nov. 2008, doi: 10.1103/PhysRevE.78.051918.

[18] D.L., Gonzalez, "The mathematical structure of the genetic code," The Codes of Life: The Rules of Macroevolution, vol.1. Springer, pp. 111-152, 2008.

[19] M. I. A el Maaty, M. M Abo-Elkhier, and M. A Abd Elwahaab, "3D graphical representation of protein sequences and their statistical characterization," Physica A: Statistical Mechanics and Its Applications, vol. 389, no. 21, pp. 4668-4676, (2010).

[20] A El-Lakkani and S El-Sherif, "Similarity analysis of protein sequences based on 2D and 3D amino acid adjacency matrices," Chemical Physics Letters, vol. 590, p. 192–195, 2013.

[21] J Wen and Y Zhang, "A 2D graphical representation of protein sequence and its numerical characterization," Chemical Physics Letters, vol. 476 , p. 281–286, 2009.