MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **72** (2014) 519-532

ISSN 0340 - 6253

A 2D Graphical Representation of Protein Sequence and Their Similarity Analysis with Probabilistic Method

Manoj Kumar Gupta[1], Rajdeep Niyogi[2], Manoj Misra[3]

[1]*Department of Computer Science and Engineering, I.T.S Engineering College, Greater Noida 201308, India*

[23]*Department of Electronics & Computer Engineering, Indian Institute of Technology Roorkee, Roorkee-247667, India*

{[1]manojdec, [2]rajdpfec, [3]manojfec}@iitr.ernet.in

(Received June 20, 2013)

Abstract

In this paper, a new 2D graphical representation of protein sequences is proposed. The coordinates of twenty amino acids are confined to only the first quadrant. The assignment of coordinates to each of the twenty amino acids is based on their molecular weights. This graphical representation is used to construct the probabilistic distribution of protein sequences. The distribution corresponding to each protein sequence is then used to analyze the similarity by using the relative entropy (Kullback–Leibler divergence). The proposed method is tested on ND6 protein sequences taken from eight different species. The phylogeny shown by the method is in agreement with previous studies on the same dataset. The proposed method does not require any alignment of protein sequence as compared to traditional alignment methods. Moreover, our results are consistent with the alignment methods.

[1] Corresponding Author

1. Introduction

An important problem in bioinformatics is the analysis and comparison of DNA and protein sequences efficiently to depict their evolutionary relationship. The principal objective of evolutionary studies is to follow the state of the system through long periods of time. However, it is impractical to repeat these evolutionary events in the laboratory. Therefore, the approaches for comparing biological sequences are mainly based on computational and statistical methods. Several mathematical approaches have been proposed by researchers which accompanied the translation of letters in protein sequence to 2D or 3D graphical representations. The representation convoys the mathematical objects which characterize the sequence numerically. These mathematical objects are used to compare the respective sequences.

The pioneering work by Hamori and Ruskin [1] in 1983 established that the graphical techniques for DNA sequence is a powerful tool for visualizing and analyzing sequence data. It has been subsequently used by other researchers [2,3,4,5,6]. Recently, numerous graphical representations have been proposed, for analyzing the similarity of DNA sequences [7,8,9,10,11,12,13,14] and protein sequences [15,16,17,18,19,20,21,22,23,24] incorporating the algorithmic and computational statistics. However, the center of attention of these problems is the consideration of the sequence as a string. Hence, all four nucleotide bases in DNA sequences and twenty amino acids in protein sequences are equally treated. One limitation of this approach is that it restricts better insight of their chemical structures and physicochemical properties. In fact, the physicochemical properties of amino acids are found to have strong effects on amino acid substitution rates [25,26]. Hence these properties directly determine the estimation of distance between two amino acid sequences.

Conventionally, the comparisons of protein sequences are made on the basis of alignment methods. A score function (PAM or BLOSSUM matrix) is used to derive the probabilities related with the insertion, deletion and substitution of amino acid in the compared protein sequences [27]. This score function helps in determining the alignment of the protein sequences. Though, the alignment methods have high computational cost. Alignment-free by graphical representation is equally contributing in results and their computational cost is also very low. The graphical representations based on physicochemical property imply their biological relevance.

In this paper, a new two-dimensional (2D) graphical representation of proteins based on a physicochemical property of amino acid is proposed which has been discussed in the following section. In this study, we have assigned the coordinates to twenty amino acids in the first quadrant only. This allows the construction of probabilistic distribution for the protein sequence. Subsequently, the symmetric Kullback–Leibler divergence [28] is used to perform similarity analysis among protein sequences. The approach is tested with NADH dehydrogenase subunit 6 (ND6) proteins taken from eight different species. The phylogenetic results obtained are consistent with previous studies. The results have been discussed in Sec. 3 and finally the work has been concluded in Sec. 4.

2. Material and Methods

For this study, the probabilistic distribution method [7] is used for the numerical characterization of graphical curve. The characterization facilitates the quantitative comparison of protein sequence.

2.1. A new 2D Graphical Representation of Protein Sequence

The new graphical representation that we are proposing is based on one of the chemical properties of amino acid (Molecular weight). We have considered the molecular weight of each amino acid for constructing the graphical descriptors. Using the chemical properties of amino acid for their graphical representation will give better insight in comparative studies of protein sequences. Initially, all the twenty amino acids are arranged in descending order of

Table 1. Amino acids with their molecular weights and normalized values (y-coordinate).

Amino Acid	Molecular weights	y-coordinates	Amino Acid	Molecular weights	y-coordinates
Tryptophan (W)	204.23	0.9	Asparagine (N)	132.12	0.4534
Tyrosine (Y)	181.19	0.7573	Leucine (L)	131.18	0.4475
Arginine (R)	174.20	0.712	Isoleucine (I)	131.18	0.4474
Phenylalanine (F)	165.19	0.6581	Cysteine (C)	121.16	0.3855
Histidine H (H)	155.16	0.596	Threonine (T)	119.12	0.3728
Methionine (M)	149.21	0.5592	Valine (V)	117.15	0.3606
Glutamate (E)	147.13	0.5463	Proline (P)	115.13	0.3481
Lysine (K)	146.19	0.5405	Serine (S)	105.09	0.2859
Glutamine (Q)	146.15	0.5402	Alanine (A)	89.09	0.1868
Aspartate (D)	133.10	0.4594	Glycine (G)	75.07	0.1

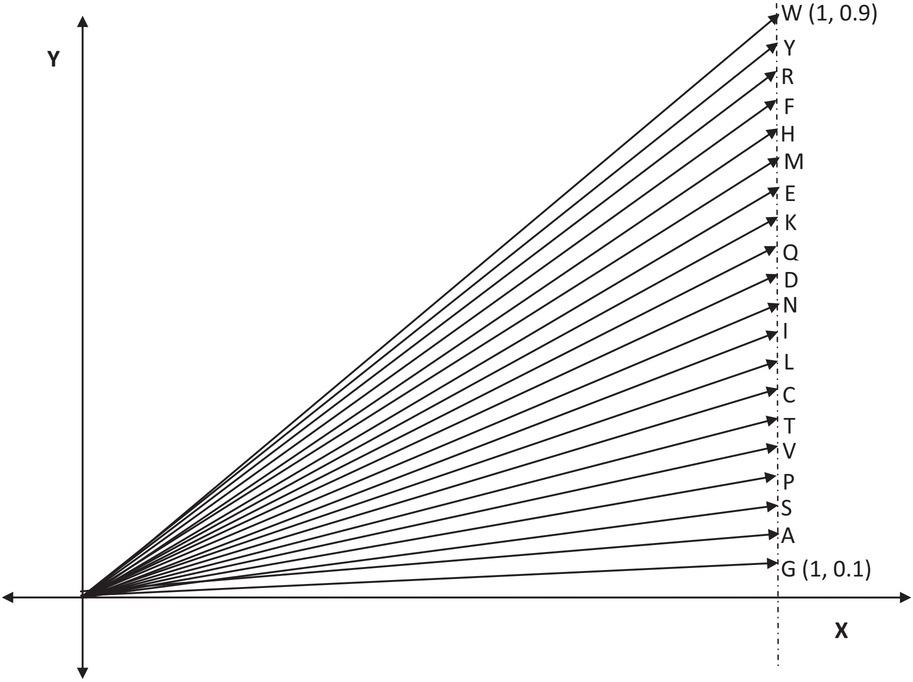

Fig. 1. Twenty vectors corresponding to each of the twenty amino acids

their molecular weights as shown in Table 1. In order to get the new descriptors by using the molecular weights, it needs to be normalized. The molecular weights are normalized using following formula:

$$Y = a + \frac{(X-m1)(b-a)}{(m1-m2)} \qquad [1]$$

where, X is the molecular weight of each amino acid as shown in Table 1, *m1* is the maximum value of molecular weight of Tryptophan (W), *m2* is the minimum value of molecular weight of Glycine (G) in Table 1. The values of a and b are taken as 0.1 and 0.9 respectively. Due to the constraint of obtaining the probabilistic distribution, the normalized values should be between 0 and 1.

These normalized values (i.e. Y in Eq. [1]) between 0.1 and 0.9 are taken as the y-coordinates of amino acids are shown in Table 1. The x-coordinate is 1 for all these values. In order to avoid the problem of degeneracy the y-coordinate values should be distinct. However, due to the equal molecular weights of Leucine (L) and Isoleucine (I), their calculated normalized values are also identical. In order to have distinct values of y-coordinates, initially the two amino acids are arranged in descending alphabetical order. The y-coordinate of L remains unchanged. We have used a difference of 0.0001 between the y-coordinates of L and I to establish that these ordinate are distinct. Therefore, the y-coordinate of Isoleucine (I) is taken greater than Leucine (L) ordinate by 0.0001. Selecting on the basis of alphabetical order is equivalent to random order. The proposed graphical descriptors are based on one of the chemical properties, here it is molecular weight. According to our method the coordinates of the descriptors would be same since the molecular weights of Isoleucine and Leucine are same. The problem could persist while selecting other physicochemical property. For example, Kyte and Doolitle [29] have reported the same hydrophobicity values for the two amino acid pairs. In their scale, the hydrophobicity values vary from 0 (for glutamine) to 2.65 (for tryptophan), but two pairs of amino acids happen to have the same hydrophobicity: serine and threonine (0.05), and alanine and histidine (0.61). The presence of duplicate values does not allow a unique ordering of the 20 natural amino acids, and hence their descriptors.

In view of the above difficulties, an alternative strategy is to take a pool of physicochemical properties of amino acids and select any two of them as the basis for construction of a graphical representation of proteins. For more effectiveness, complementary properties may be preferred, that is structurally or functionally related properties. We would pursue it in our future work.

The vectors corresponding to 20 amino acids are shown in Fig. 1. The points in the graphical representation of protein sequence are obtained by the sum of vectors representing amino acids. The graph representing a protein sequence does not form a circuit as it progresses in an incrementing fashion along the positive x-axis. Therefore, the problem of degeneracy does not exist. Otherwise, in case of degeneration the protein sequence will not be recoverable.

We have shown graphically in Fig. 2, the ND6 proteins of common chimpanzee, gorilla, human, and wallaroo, which are based on the vector system shown in Fig. 1. It could be closely observed in Fig. 2, the sharp curve between 40 to 60th amino acid in the graph of common chimpanzee, gorilla, and human where as the smooth curve in case of wallaroo. Since human, chimpanzee, and gorilla belongs to the same group called primates, their representations are more similar as compared to wallaroo. Similarly, In Fig. 2, at 100th, 120th, and 160th x-coordinate in the graph of chimpanzee, gorilla, human their corresponding y-coordinate values are less than 40, 50, and 65 respectively, where as y-coordinate values for wallaroo greater than 40, 50 and 65 respectively. Therefore, it has been observed, the graphical representation helps visually in inspecting the relatedness among the protein sequences. But the visual inspection does not distinguish appropriately unless the degree of

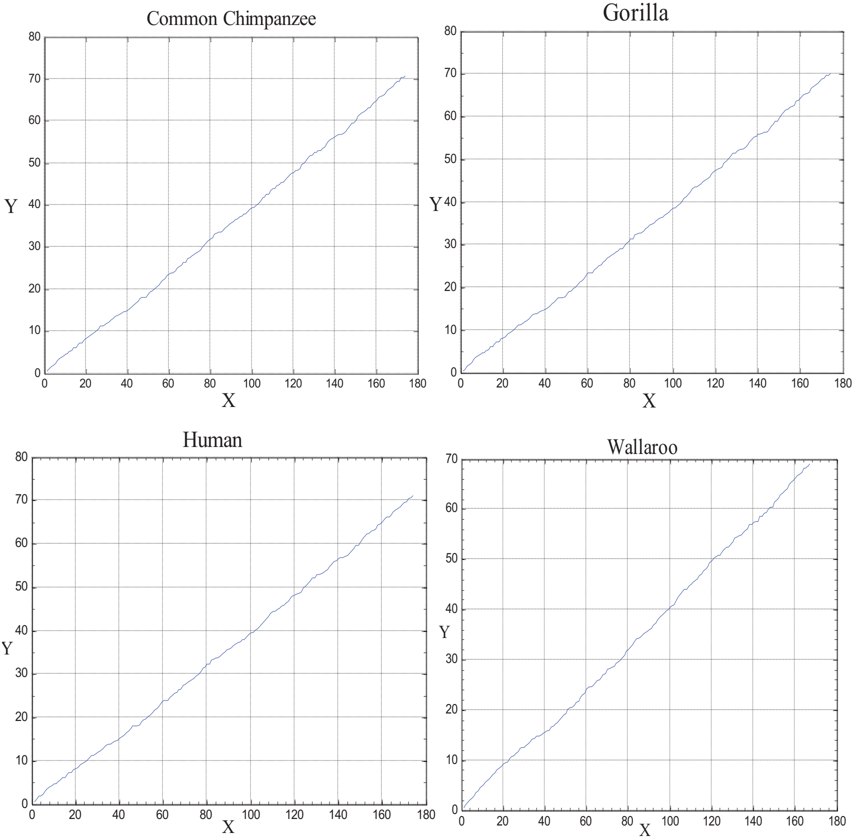

Fig. 2. Graphical representation of protein sequences. Graphical representations of ND6 protein sequences of four species (Common chimpanzee, Gorilla, Human and Wallaroo) based on the vector system shown in Fig. 1. X-value stands for the number of amino acids in the protein sequence. Y-value is the cumulative y-values according to the third column of Table 1.

relatedness has been quantified. Protein sequences with the intention of using them for quantitative comparative study of the degree of similarity/dissimilarity needs numerical characterization, which has been discussed in following section.

2.2. Probabilistic distribution method and Symmetric Kullback-Leibler divergence

Wu, Burke, and Davison [30] introduced the Mahalanobis distance [31] and Standardized Euclidean distance (SED) into the study of DNA sequence similarity. They have shown that both distances had better sensitivity and selectivity than commonly used Euclidean distance (ED). The primary reason behind improvement is Mahalanobis distance is true statistical distance and accounts for both variance and covariance between the frequencies of words. Furthermore, in case of difficulty in computing the Mahalanobis distance, standardized Euclidean distance (SED) still performs better performance than the Euclidean distance as SED accounts for the variances of frequencies of words. Huang et al. [12] have applied Evolutionary Angle Distance [32] in addition to MD, SED, and ED for computing the distances for examining the similarity of sequences on random data set. They concluded with the ED is more sensitive to mutation rate than the other three distance measures. Correlation coefficient is also one of the similarity measures used by many researchers. In addition, one fundamental equation of information theory is used to quantify the proximity of two probability distributions called Kullback-Leibler divergence. The term relative entropy was first defined by Kullback and Leibler [33]. It is known under a variety of names, including the Kullback–Leibler divergence, K-L distance, information divergence, cross entropy, and information for discrimination, and has been studied in detail by Csisz´ar [34] and Amari [35]. The relative entropy is a measure of the distance between two distributions. If P_1 and P_2 are two discrete probability distributions then the Kullback–Leibler divergence (KLD) or the relative entropy, denoted as H (P_1, P_2) of P_1 with respect to P_2 is defined in Eq. [3]. Similarly H (P_2, P_1) is the relative entropy of P_2 with respect to P_1. Since KLD is asymmetric, we simply have defined other distance measures which is equal to the mean of H (P_1, P_2) and H (P_2, P_1) as shown in Eq. [4]. This does not affect the asymmetric property of KLD.

The probabilistic distribution method has been used to numerically characterize the protein sequence. Yu et al. [7] proposed a method for the probability distribution for protein sequence. The probability distribution of protein sequence of length n is defined as

$$p_i = \frac{x_i - \overrightarrow{y_i}}{\frac{1}{2}n(n+1) - y_i} \qquad [2]$$

where n is the length of protein sequence, x_i and y_i are the coordinate of amino acid at position i, and $\overrightarrow{y_i}$ is the y-coordinate value at the i^{th} amino acid in the protein graphical curve.

Since the sequences are of varying length, the probabilistic distribution of protein sequence is transformed to normalize the probability distribution by specific N, where N is the length of the smallest protein sequence in the given set of sequences, and $N \leq n$. By using the Eq. [2] of probability distribution, we get the probability distributions ($p_1, p_2, \ . \ . \ . \ , p_N$) for each of the (n- N + 1) subsequences of length N. Subsequently, average over the probabilistic distributions is calculated to obtain a normalized probability distribution for the protein sequence. After obtaining the normalized probability distribution, a similarity/dissimilarity measure between two discrete probability distributions $P_1 = (p_1, p_2, . . . , p_n)$ and $P_2 = (q_1, q_2, . . . , q_n)$ is calculated using Kullback–Leibler divergence [36] (a dissimilarity measure), denoted by $H(P_1,P_2)$ of P_1 with respect to P_2 is defined as

$$H\ (P_1, P_2) = \sum_{i=1, p_i \in P_1, q_i \in P_2}^{n} p_i log \frac{p_i}{q_i} \qquad [3]$$

The dissimilarity measure is zero, if the probability distribution of two protein sequence are same i.e., P_1=P_2. However, the dissimilarity measure, a distance metric does not give the correct result due to $H(P_1,P_2) \neq H(P_2,P_1)$; and it also does not satisfy the triangle inequality. Therefore, the symmetric Kullback-Leibler divergence [28] is used and it is defined as

$$\mathrm{d}\ (\mathrm{P}_1, \mathrm{P}_2) = \frac{\mathrm{H}\ (\mathrm{P}_1, \mathrm{P}_2) + \mathrm{H}\ (\mathrm{P}_2, \mathrm{P}_1)}{2} \qquad [4]$$

We can see clearly that this metric is symmetric i.e., $d(P_1,P_2) = d(P_2,P_1)$. Conclusively, a distance matrix is attained showing the distance between every pair of protein sequence. Following the distance matrix, phylogenetic tree is constructed by Unweighted Pair Group Method with Arithmetic Mean (UPGMA) [37] of MEGA5 package [38].

2.3. Dataset

The following eight ND6 proteins are compared for the proposed graphical representation as: human (Homo sapiens, AP_000650), gorilla (Gorilla gorilla, NP_008223), common chimpanzee (Pan troglodytes, NP_008197), harbor seal (Phoca vitulina, NP_006939), gray seal (Halichoerus grypus, NP_007080), rat (Rattus norvegicus, AP_004903), mouse (Mus musculus, NP_904339), and wallaroo (Macropus robustus, NP_007405). The same dataset has also been used in [17,22,39,40,41].

3. Result and Discussion

The proposed graphical representation of protein sequence and their probabilistic distribution has been used to construct the phylogeny of eight ND6 proteins. The details of ND6 proteins are mentioned above (Sec 2.3) along with their accession numbers. All these sequences are downloaded from Genbank (http://www.ncbi.nlm.nih.gov). As mentioned in Sec. 2.2, the value of N is taken as the length of smallest protein sequence length, i.e. N=167. Subsequently, the probability distributions of the eight protein sequences are transformed into eight normalized probability distributions. The similarity matrix for the protein sequences is constructed by using symmetric Kullback-Leibler divergence and is shown in Table 2. The smaller values in the matrix show higher degree of similarity between the species. For example, the smallest value beween H.seal and G.seal shows higher similarity between them, than any other species.

Similarly, human, gorilla and chimpanzee are more closely related with each other than other species. The phylogenetic tree is obtained from UPGMA method available in MEGA5 package [38], which uses the similarity matrix as an input as depicted in Fig. 3. We can also observe in Fig. 3 that gorilla, human and chimpanzee are more closely related with each other than any other species in the tree. Similarly, mouse, rat, and H.seal, G.seal have identical recent common ancestor. A wallaroo, which is intermediate in size between a kangaroo and a wallaby, is most dissimilar among the given set of species as shown in Fig. 3. The phylogeny obtained is consistent with the previous studies [17,22,39,40,41] on ND6 proteins.

For comparison of our result with others, we list some published results on examining the degree of similarity of human and other several species in Table 3. The results taken from the published work are on same data set as of us, except the Opossum species instead of walleroo by Randic, so that, there exist uniformity in comparison. As it can be observed from the Table

3, there are three groups: 1) gorilla and common chimpanzee are closest to human falls in first group and the corresponding columns values are highlighted in bold; (2) H.seal and G.seal are in second group; and (3) the group of rat and mouse is distant from human in evolutionary relationship. According to the values for wallaroo and opossum in the Table 3, it should be concluded that these species are farthest from human in evolutionary relationship among the given set of species. We can articulate that there exists an overall agreement among similarities obtained by other approaches, despite some variation among them.

Table 2. The similarity matrix for eight ND6 protein sequence.

	Chimpanzee	H.seal	Gorilla	G.seal	Mouse	Rat	human	Wallaroo
1.0e-005								
Chimpanzee								
H.seal	0.05612							
Gorilla	0.01524	0.06004						
G.seal	0.05893	0.00093	0.06371					
Mouse	0.15404	0.10013	0.11832	0.10360				
Rat	0.21588	0.12131	0.17952	0.12390	0.02642			
human	0.01250	0.06948	0.00575	0.07363	0.12756	0.19609		
Wallaroo	0.81362	0.60299	0.77079	0.59363	0.69457	0.68894	0.80224	

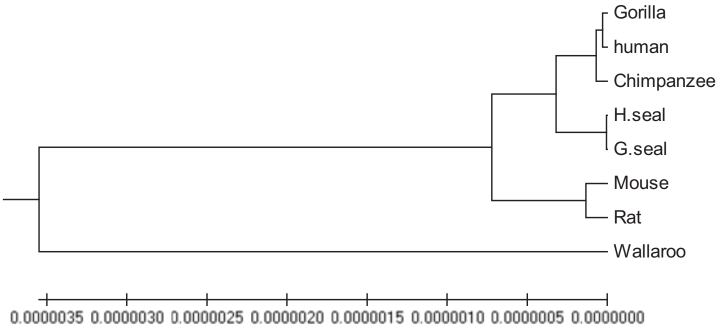

Fig. 3. The phylogenetic tree by UPGMA method for eight ND6 protein sequence.

Table 3. The comparison among other published work of similarity between the coding sequences of several species with the coding sequence of human

Species-->	Gorilla	Chimpanzee	Wallaroo	Opossum	H.seal	G.seal	Rat	Mouse
This Work	**0.00575**	**0.0125**	0.80224	-	0.0694	0.0736	0.196	0.12756
From Table 3 in [22]	**0.0094**	**0.0118**	0.0369	-	0.0247	0.0284	0.033	0.0262
From Table 2 in [17]	**0.0338**	**0.0979**	0.278	-	0.1797	0.1487	0.2071	0.1472
From Table 4 in [39]	**8.25**	**6.92**	-	16.79	12.81	13.11	14.63	15.03

In order to illustrate the effectiveness of the proposed method, we have implemented the same dataset of protein sequence using Clustal omega (a multiple sequence alignment program). The phylogenetic tree obtained from Clustal omega is shown in Fig. 4. The phylogeny of eight species shown in Fig. 3 and Fig. 4 are similar. The lineages in tree have not been considered in the present study and the trees have been compared on their phylogeny. The similarity of the trees are mostly observed on the basis of the evolutionary relationship among the species; like human, gorilla, and chimpanzee falls in one group, similarly mouse and rat, H.seal and G.seal are two individual groups showing close phylogeny, and wallaroo is distantly related from human group as shown in both the Fig. 3 and Fig. 4.

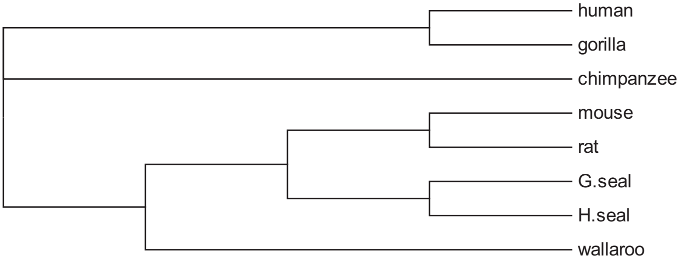

Fig. 4. The phylogenetic tree from Clustal omega for eight ND6 protein sequence.

The complexity of the proposed method is O(*Nmn*), where *N* is the length of smallest sequence length, *m* is the number of protein sequences and *n* is the length of largest sequence length. It may be noted that the complexity of multiple sequence alignment method is O(n^m). According to the evolutionary results (Fig. 3), it seems that our method may be useful for evolutionary analysis. The proposed 2D graphical representation of protein sequence does not form close loop path (i.e. circuit) and also does not suffer with the problem of degeneracy. Ordering amino acids based on their physicochemical properties offer better insights into similarity analysis of protein sequences than random ordering of amino acids. Though the proposed method uses the molecular weight physicochemical property, other properties such as aromaticity, aliphaticity, hydropathy, hydroxythiolation [19] etc, can also be explored for other type of informative representation.

4. Conclusion

We have proposed a new 2D graphical representation of protein sequence which does not form circuit and free from the problem of degeneracy. The method has been applied to ND6 proteins of eight different organisms and the results obtained match well with the evolutionary chronology of the organisms. Phylogeny obtained from our method and Clustal omega is similar. The proposed approach uses the physicochemical property of amino acid. This enables better insight into similarity analysis of protein sequences. It is observed that the final phylogeny obtained by alignment-free using graphical techniques takes significantly less computational cost as compared to multiple sequence alignment methods which have exponential time complexity. In this paper, we have used only one physicochemical property i.e. molecular weight of amino acid. As part of ongoing work, we would like to explore some other properties like aromaticity, hydropathy, isoelectric point, etc. for similarity analysis.

Acknowledgement: The authors would like to thank the referee for the valuable comments.

References

[1] E. Hamori, J. Ruskin, H Curves, A novel method of representation of nucleotide series especially suited for long DNA sequences, *J. Biol. Chem.* **258** (1983) 1318–1327.

[2] M. A. Gates, Simpler DNA sequence representations, *Nature* **316** (1985) 219–219.

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

[4] P. M. Leong, S. Morgenthaler, Random walk and gap plots of DNA sequences, *Comput. Appl. Biosci.* **11** (1995) 503–507.

[5] M. Randić, M. Vračko, A. Nandy, S. C. Basak, On 3-D graphical representation of DNA primary sequences and their numerical characterization, *J. Chem. Inf. Comput. Sci.* **40** (2000) 1235–1244.

[6] M. Randić, Condensed representation of DNA primary sequences, *J. Chem. Inf. Comput. Sci.* **40** (2000) 50–56.

[7] C. Yu, M. Deng, S. S. T. Yau, DNA sequence comparison by a novel probabilistic method, *Inform. Sci.* **181** (2011) 1484–1492.

[8] Y. H. Yao, Q. Dai, X. Y. Nan, P. A. He, Z. M. Nie, S. P. Zhou, Y. Z. Zhang, Analysis of similarity/dissimilarity of DNA sequences based on a class of 2D graphical representation, *J. Comput. Chem.* **29** (2008) 1632–1639.

[9] C. Yu, Q. Liang, C. Yin, R. L. He, S. S. T. Yau, A novel construction of genome space with biological geometry, *DNA Res.* **17** (2010) 155–168.

[10] S. Ding, Q. Dai, H. Liu, T. Wang, A simple feature representation vector for phylogenetic analysis of DNA sequences, *J. Theor. Biol.* **265** (2010) 618–623.

[11] M. K. Gupta, R. Niyogi, M. Misra, A new adjacent pair 2D graphical representation of DNA sequences, *J. Biol. Syst.* **21** (2013) # 1350005 (15 pages).

[12] G. Huang, H. Zhou, Y. Li, L. Xu, Alignment–free comparison of genome sequences by a new numerical characterization, *J. Theor. Biol.* **281** (2011) 107–112.

[13] N. Jafarzadeh, A. Iranmanesh, A novel graphical and numerical representation for analyzing DNA sequences based on codons, *MATCH Commun. Math. Comput. Chem.* **68** (2012) 611–620.

[14] J. Song, Analysis of similarity of DNA sequences based on a novel 3-D graphical representation, in: T. Chang (Ed.), *Advances in Biochemical Engineering*, Inf. Engin. Res. Inst., Newark, 2012, pp. 29–36.

[15] H. J. Yu, D. S. Huang, Novel 20-D descriptors of protein sequences and it's applications in similarity analysis, *Chem. Phys. Lett.* **531** (2012) 261–266.

[16] Z. H. Qi, J. Feng, X. Q. Qi, L. Li, Application of 2D graphic representation of protein sequence based on Huffman tree method, *Comput. Biol. Med.* **42** (2012) 556–563.

[17] L. Z. X. Xie, Y. Yu, L. Liang, M. Guo, J. Song, Z. Yuan, Protein sequence analysis based on hydropathy profile of amino acids, *J. Zhejiang Univ. Sci. B* **13** (2012) 152–158.

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

[19] C. Yu, S. Y. Cheng, R. L. He, S. S. T. Yau, Protein map: An alignment–free sequence comparison method based on various properties of amino acids, *Gene* **486** (2011) 110–118.

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

[21] Z. C. Wu, X. Xiao, K. C. Chou, 2D-MH: A web–server for generating graphic representation of protein sequences based on the physicochemical properties of their constituent amino acids, *J. Theor. Biol.* **267** (2010) 29–34.

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

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

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

[25] X. Xia, W. H. Li, What amino acid properties affect protein evolution? *J. Mol. Evol.* **47** (1998) 557–564.

[26] Z. Yang, R. Nielsen, M. Hasegawa, Models of amino acid substitution and applications to mitochondrial protein evolution, *Mol. Biol. Evol.* **15** (1998) 1600–1611.

[27] S. Henikoff, J. G. Henikoff, Amino acid substitution matrices from protein blocks, *Proc. Nat. Acad. Sci. USA* **89** (1992) 10915–10919.

[28] L. R. R. B. H. Juang, A probabilistic distance measure for hidden Markov models, *AT&T Techn. J.* **64** (1985) 391–408.

[29] J. Kyte, R. F. Doolittle, A simple method for displaying the hydropathic character of a protein, *J. Mol. Biol.* **157** (1982) 105–132.

[30] T. J. Wu, J. P. Burke, D. B. Davison, A measure of DNA sequence dissimilarity based on mahalanobis distance between frequencies of words, *Biometrics* **53** (1997) 1431–1439.

[31] R. Mazumder, A. Kolaskar, D. Seto, GeneOrder: comparing the order of genes in small genomes, *Bioinformatics* **17** (2001) 162–166.

[32] G. W. Stuart, K. Moffett, J. J. Leader, A comprehensive vertebrate phylogeny using vector representations of protein sequences from whole genomes, *Mol. Biol. Evol.* **19** (2002) 554–562.

[33] S. Kullback, R. A. Leibler, On information and sufficiency, *Ann. Math. Stat.* **22** (1951) 79–86.

[34] I. Csiszár, Information–type measures of difference of probability distributions and indirect observations, *Studia Sci. Math. Hung.* **2** (1967) 299–318.

[35] S. Amari, *Differential–Geometrical Methods in Statistics*, Springer, New York, 1985.

[36] T. M. Cover, J. A. Thomas, *Elements of Information Theory*, Wiley, New York, 1991.

[37] R. Sokal, C. Michener, A statistical method for evaluating systematic relationships, *Univ. Kansas Sci. Bull.* **38** (1958) 1409–1438.

[38] K. Tamura, D. Peterson, N. Peterson, G. Stecher, M. Nei, S. Kumar, MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods, *Mol. Biol. Evol.* **28** (2011) 2731–2739.

[39] 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.

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

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