ACS OMEGA

http://pubs.acs.org/journal/acsodf

Article

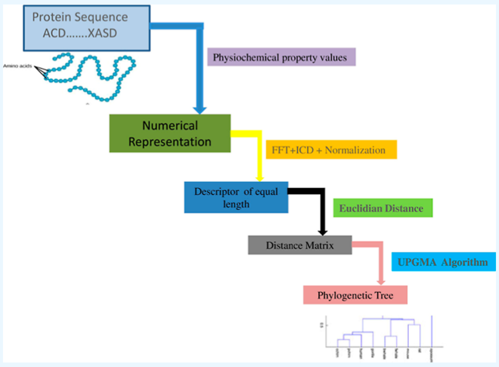
Mathematical Approach to Protein Sequence Comparison Based on Physiochemical Properties

Jayanta Pal,* Soumen Ghosh, Bansibadan Maji, and Dilip Kumar Bhattacharya

Cite This: *ACS Omega* 2022, 7, 39446−39455

Read Online

ACCESS | Metrics & More | Article Recommendations



ABSTRACT: The difficult aspect of developing new protein sequence comparison techniques is coming up with a method that can quickly and effectively handle huge data sets of various lengths in a timely manner. In this work, we first obtain two numerical representations of protein sequences separately based on one physical property and one chemical property of amino acids. The lengths of all the sequences under comparison are made equal by appending the required number of zeroes. Then, fast Fourier transform is applied to this numerical time series to obtain the corresponding spectrum. Next, the spectrum values are reduced by the standard inter coefficient difference method. Finally, the corresponding normalized values of the reduced spectrum are selected as the descriptors for protein sequence comparison. Using these descriptors, the distance matrices are obtained using Euclidian distance. They are subsequently used to draw the phylogenetic trees using the UPGMA algorithm. Phylogenetic trees are first constructed for 9 ND4, 9 ND5, and 9 ND6 proteins using the polarity value as the chemical property and the molecular weight as the physical property. They are compared, and it is seen that polarity is a better choice than molecular weight in protein sequence comparison. Next, using the polarity property, phylogenetic trees are obtained for 12 baculovirus and 24 transferrin proteins. The results are compared with those obtained earlier on the identical sequences by other methods. Three assessment criteria are considered for comparison of the results—quality based on rationalized perception, quantitative measures based on symmetric distance, and computational speed. In all the cases, the results are found to be more satisfactory.

■ INTRODUCTION

Genome sequence analysis is comparatively more frequent than protein sequences since proteins have a higher degree of spatial complexity, in the sense that protein is a sequence of 20 amino acids, whereas genome is a sequence of four nucleotides. Researchers need to create algorithms that can handle massive volumes of data in a timely manner due to the rapid rise in the protein database. Again, the lengths of the protein sequences in most of the species vary, so those techniques are to be developed that are quick and flexible enough to handle protein sequences of different lengths. Traditional alignment-based methods have become outdated in this situation due to the tremendous temporal complexity.[1−3] Alignment-free methods are later employed to overcome the obvious problems with the alignment-based strategy, as taken up in refs 4 and 5. The authors of ref 6 provide a sizable literature for alignment-free methods up to 2003. Again, several instances of alignment-free techniques are available in refs 7, 8, and 9, and they are also very appropriate. It is observed that, among the several methods, the graphical representation is the oldest. Another unique sequence analysis technique is to apply the composition vector (CV). An excellent summary of this method is provided in ref 10. The four components of the CV approach are as follows.

(1) To count how many times, for a given value of "*k*" each *k*-string appears in the sequence.
(2) To create the CV using the occurrence counts for all sequences considered.
(3) To calculate the distance between any two CVs for getting the distance matrix.
(4) To build the phylogenetic tree using the distance matrix.

The advantage of using the CV approach in genome sequence comparison is that the frequency vector produced from the K-mer occurrence count is 4k. The same in protein sequence, however, is 20k, which is exceptionally large. Therefore, the CV technique is limited to its application in protein sequence comparison due to this enormous volume, despite its notable success in genome sequence analysis.[10−12] In recent work, the authors in ref 13 effectively apply such methods in comparing protein sequences. In any case, the CV

Received: September 21, 2022
Accepted: September 27, 2022
Published: October 17, 2022

technique is not frequently utilized for comparing protein sequences. It may be noted that protein sequence comparison rarely employs probabilistic methods. However, the authors of ref 14 obtain probability descriptors from a special type of 2D representation of DNA sequences and apply it to DNA sequence comparison. However, the definition of the probability vector is incorrect. The reason is that it could not fulfill the basic criteria that all the components of the probability vector must be positive. In fact, it could not be assured that all components except the first two components are positive. Again, the authors in ref 15 extend this comparison method to protein sequences, which also suffers from the same problem in defining the probability vector. 3-mer type of representations of amino acids with emphasis on the arrangements and positions of the neighbors is taken up for protein sequence comparison in ref 16. The advantage of the method is that it can be applied to sequences of equal and unequal lengths. Again, groups of amino acids of different cardinality based on their physiochemical properties are commonly used in studying protein sequences.[17−19] Authors in ref 19 also compare protein sequences based on pair of groups with the same cardinality three. However, one pair of classified categories lacks adequate information. There are some comparisons of protein sequences based on the property values of the amino acids, both physical and chemical. The properties considered are as small as two in number and as large as 12 in number. All the properties are either physiochemical or chemical in nature. In ref 20, a model is presented based on position-feature for comparing sequence of proteins by measuring graph energy. The representation's physiochemical attributes include pI and pK_a values. The results acquired by this method show significant meaning when compared to those obtained by other methods. In ref 21, one more physical property side-chain mass and one more chemical property hydrophobicity are considered to obtain two dimensional representation. As a result the "2D-MH" curve is generated, with "*H*" representing hydrophobic quantity, and "*M*" representing mass of the amino acid side chain. In ref 22, 2D mapping is utilized to create two-dimensional representation. The coordinates are calculated based on only two chemical properties viz., the pK_a NH3 value and pK_a COOH value. In ref 23, only six parameters corresponding to all the 20 amino acids are explored in a unique two-dimensional graphical form. This graphical representation is used to compute the distance between two sequences that are to be examined. Advantage of the method is that it remains unaffected for taking difference in length of the sequence under examination. In ref 24, 12 fundamental physicochemical properties are used to demonstrate a new mapping technique for comparing some sequences of protein. Principal component analysis (PCA) is used to calculate the proportion of amino acids present on 12 principal axes only. Accordingly a simplified 2D representation of sequence of proteins is obtained. Finally, a 20-dimensional vector is selected as a descriptor for all of the sequences under comparison. In ref 25, authors propose a 2D graphical depiction of sequence of proteins based on six random physicochemical attributes. In ref 26, the authors use 10 distinct physiochemical properties of amino acids. This is also another random selection. In ref 27, author uses indices of only three physicochemical parameters, such as PH, PI, and Hp, to transform a protein sequences into a 23-dimensional vector. Finally, the similarity or dissimilarity of ND5 proteins from nine different species is calculated using Euclidean distance. In ref 28, a completely new form of iterated function systems is introduced based on four entirely random and arbitrarily chosen physicochemical properties of amino acids, viz., pK1, h, pK2, and pI. This results in a 2D graphical representation. This is applied in the similarity/dissimilarity analysis of 9 ND5 and 8 ND6 sequences of proteins. The method of PCA is also used in comparison of protein sequences in ref 29. It uses nine different properties for representation. These are the molecular weight (mW), hydropathy index (hI), solubility (S), van der Waals radius of side chains (vR), the pK_a value for terminal amino acid groups COOH (Pk1), the pK_a value for terminal amino acid groups NH+ 3 (pK2), isoelectric point (pI), the number of triplet codons (cN), and frequency of Ahuman proteins (F). In a recent paper,[30] the authors use 12 physicochemical properties of amino acids for representation of protein sequences. First of all, a N × L matrix is formed using N amino acids each having L = 12 property values arranged along the columns. Out of this N × L matrix, again a N x L matrix is created for each amino acid in the sequence by changing the values of each row by the addition of chosen row with some suitable normalized distance of the other rows. Each such N × L matrix for each amino acid is now subjected to PCA analysis to get the respective first Eigen vector expressed in columns. All such column vectors, each of length N, form a N × N matrix. Again by applying PCA analysis, its first Eigen vector is calculated; finally, this N component Eigen vector is taken as the descriptor. The method is applicable to compare sequences of equal length as the length of the descriptor depends on the length of the sequence. A very recent paper[31] also uses various physiochemical properties, but this is not applied in protein sequence comparison.

Another important method is the application of FFT to the processing of biological signals. FFT is always a popular tool for image and signal processing.[32] Gene prediction and hierarchical analysis are the principal uses of FFT in DNA comparison.[33,34] Again, this is accomplished by using the FFT of a DNA sequence's spectrum to display the distribution and a periodic pattern of the sequences. The inter coefficient difference (ICD) method is applied to the spectrum of the binary representation of DNA sequences to compare genome sequences.[35] Authors in ref 36 use a different method based on the Fourier power spectrum to compare DNA sequences. The usage of FFT in the Voss kind of representation-based similarity analysis of protein sequences is recently identified by authors in ref 37. This study uses the ICD method for protein sequences to assess the spectrum.

Thus, a binary sequence can be used to describe the Voss type of representation for DNA and protein sequences. Based on such representations, genome sequences and protein sequences can be compared under use of FFT. Now, it is already noted that in the case of protein sequences, some of the numerical representations come from the physiochemical properties of their amino acids. Therefore, it remains open to see whether FFT can be used to compare protein sequences based on their representations using one single property, either chemical or physical in nature. Naturally, the question remains whether the above ICD method is still applicable to compare protein sequences represented by such non-binary sequences.

In the current work, a numerical representation of protein sequences is suggested using two properties of amino acids, viz., molecular weight as the physical property and polarity as the chemical property separately. The purpose is to compare protein sequences and to see whether the known result that polarity is always a better choice than molecular weight still remains true.

From the above study, it is noticed that researchers have developed numerous strategies for studying protein sequences, which have been successfully employed in analyzing protein sequences across different species. Among them, alignment-based approaches lost their utility in this context due to their temporal complexity. For instance, the time required for multiple sequence alignment, as described in ref 3, is NP-hard. On the other hand, alignment-free methods offer better outcomes in terms of time requirements, but most of them are unable to cope with sequences of unequal length. Again, physiochemical properties of amino acids are considered in numerous studies for protein sequence comparison in the literature. However, these properties are chosen randomly, and no comparison is made between the various physiochemical properties, so as to decide which one or ones perform better in protein sequence comparison. This gap in research encourages in developing a new nondegenerate technique that is both quick and suitable for comparing protein sequences of unequal lengths. This leads to the following contributory scope of the current work.

(a) To propose a simple and effective representation of protein sequences using only one physical (molecular weight) and one chemical (polarity) property of amino acids.

(b) To use FFT in the analysis of protein sequences of various lengths.

(c) To apply the ICD method to compare protein sequences represented by non-binary sequences.

(d) To compare one physical (molecular weight) and one chemical (polarity) property of amino acids to determine the most promising one in protein sequence comparison.

(e) To compare protein sequences based on the better one as found above.

The present work attempts to compare the protein sequences, represented by one physical property, molecular weight, and one chemical property, polarity of the 20 amino acids. The comparison is made from the spectrum obtained using FFT on the represented sequences. The present paper first examines the outcomes in all the cases on 9 ND4, 9 ND5, and 9 ND6 proteins. Motivated by the finding that polarity (hydropathy index) is better than molecular weight for comparing protein sequences, the paper uses polarity (hydropathy index)-based representation to compare protein sequences of 12 baculovirus and 24 TF proteins. Also, the results are compared to those obtained earlier using other known similarity study methodologies on the same species to validate the suggested method. Without any genetic involvement, the proposed FFT-based method enables the discovery of the biological reference in a time-efficient manner. It can also be used to compare protein sequences of varying lengths.

The rest of the paper is presented in several sections. Methodology and Mathematical Background with the proposed method of representation of protein sequence are discussed in Section 2. Section 3 compares the proposed method's results with those of other methods. The advantage of the proposed method is discussed in Section 4. Finally, conclusion is drawn in Section 5.

2. METHODOLOGY AND MATHEMATICAL BACKGROUND

In the current work, we first compare one physical property and one chemical property by considering the phylogenetic tree produced for 9 ND4, 9 ND5, and 9 ND6 protein sequences based on these properties. We use polarity value and molecular weight to give numerical representations. It is found that the polarity value gives a better result. In the second part of our work, we use the polarity values to represent protein sequences and analyze the similarity of 12 baculovirus and 24 TF proteins. These are explained in the subsequent sections.

2.1. Representation of Protein Sequence. The proposed method gives two different numerical representations of the protein sequence of length N, considering the values of two physiochemical properties separately, as shown in Table 1.

Table 1. Values of Molecular Weight and Polarity of 20 Amino Acids

amino acid	abbreviations		molecular weight	hydropathy index (polarity)
alanine	Ala	A	89	1.8
cysteine	Cys	C	121	2.5
aspartic acid	Asp	D	133	−3.5
glutamic acid	Glu	E	147	−3.5
phenylalanine	Phe	F	165	2.8
glycine	Gly	G	75	−0.4
histidine	His	H	155	−3.2
isoleucine	Ile	I	131	4.5
lysine	Lys	K	146	−3.9
leucine	Leu	L	131	3.8
methionine	Met	M	149	1.9
asparagine	Asn	N	132	−3.5
proline	Pro	P	115	−1.6
glutamine	Gln	Q	146	−3.5
arginine	Arg	R	174	−4.5
serine	Ser	S	105	−0.8
threonine	Thr	T	119	−0.7
valine	Val	V	117	4.2
tryptophan	Trp	W	204	−0.9
tyrosine	Tyr	Y	181	−1.3

At this stage, the proposed method adds additional zeros if required to make all the sequences of equal length. This is called zero padding.[35,38] Then, FFT is applied to these numerical values using eq 1.

$$U_i(k) = \sum_{n=0}^{N-1} u_i(n)\, e^{-i\left(\frac{2\pi}{N}\right)kn}, \qquad k = 0,1,2, ..., N-1 \tag{1}$$

where N is the sequence length and k stands for frequency. The following section explains the representations using molecular weight and polarity.

2.1.1. Representation Using Molecular Weight and Polarity. The protein sequence consists of a combination of 20 amino acids, each of which is represented by using the value of the molecular weight and polarity separately, as shown in Table 1. It produces two representations, one using molecular weight and the other using polarity. A sequence having length N is represented by a series of N positive integers in case of molecular weight and N real numbers in case of polarity. Then, FFT is used on this numerical representation of the time series, producing N set of conjugate complex numbers $x + iy$ and $x - iy$. Subsequently, the amplitudes of these FFT values are calculated using eq 2

$$Z_i = \sqrt{x_i^2 + y_i^2} \; ; \; i = 1,2, ..., N/2 \tag{2}$$

Finally spectrum of length $N/2$ is calculated from these amplitudes.

2.2. ICD Method to Calculate the Descriptor. In the present work, ICD method is carried out by extracting amplitudes parts from the FFT values and then taking the absolute differences between every pair of consecutive amplitudes. For example, from a time series of sample length "N", $N/2$ (let it be M) amplitude values of the spectrum are extracted. Finally, after applying ICD to these amplitude values of length "M", a sequence (let it be A) of length "$M - 1$" is calculated, taking the absolute differences between every pair of consecutive amplitude values. As a final step, the resulting sequence A is normalized by dividing each element of the sequence by the norm $\|A\|$ to produce the descriptor of length "$M - 1$". Naturally, two sets of descriptors are obtained corresponding to the numerical values of two types of physiochemical properties, molecular weight and polarity for every protein sequences. Considering m number of protein sequences, we are having m number of descriptors each of length $M - 1$ for each of the property.

2.3. Formation of the Distance Matrix. Consider S_1 and S_2 as the two protein sequences under comparison of length N, after appending zeroes, if required. Furthermore, assume that D_{1i} and D_{2i} are the corresponding descriptors of length $M - 1$. It may be noted that even if S_1 and S_2 are of different lengths, D_{1i} and D_{2i} descriptors are of the same length.

$$d(S_1, S_2) = \sqrt{\sum (D_{1i} - D_{2i})^2} \; , \qquad i = 1,2, ..., M - 1 \tag{3}$$

Next, the distance matrix is formed by calculating the distance between each pair of sequences using eq 3. It is a symmetric matrix of size $m \times m$. The lesser the distance, the more similar are the pair of proteins. Finally, this distance matrix is used to obtain a phylogenetic tree applying the UPGMA algorithm using Molecular Evolutionary Genetics Analysis 11 (MEGA11) software.[39] Two phylogenetic trees are obtained corresponding to the two numerical values of molecular weight and polarity. These two phylogenetic trees are compared to identify the most promising one. It is found that polarity is the best one. Next, the paper conducts a similarity study for protein sequences of 12 baculovirus and 24 TF proteins using property of polarity. Details algorithm of the proposed method is given in Table 2.

2.4. Calculation of Time Complexity. Time complexity is calculated based on the time required in the following main steps.

(a) Protein sequence representation using the proposed method requires linear time. Therefore, considering the maximum length of the protein sequence as "N"; it results in O(N) in the worst case.

(b) FFT calculation on the representation of the numeric sequence of length "N" requires $O(N \log N)$ time.

(c) Descriptor calculation requires linear time, that is $O(N)$.

If we consider "m" number of species, then the above three steps will be executed "m" numbers of time, giving a complexity of $O(m \times N \log N)$. It is obvious that "m" is much less than "N." Therefore, it can be assumed that the complexity remains $O(N \log N)$. Again, constructing the distance matrix requires $O(m^2)$ time where "m" is the number of species. As "m" is comparatively very small with respect to "N", the complexity remains $O(N \log N)$ only.

Table 2. Algorithm of the Proposed Method

```
Procedure FFT-ICD
Let $m$ be the number of species whose protein sequences are compared for the similarity study. Input protein sequence one by one and apply the procedure of linear complexity to get the maximum length of the sequences; say it is $N$
$i = 1$
while ($i \le m$) do
    consider protein sequence of $i$th species
    represent the protein sequence using the value of physiochemical property considered
    append additional "0" if required at the end of the represented sequence to make it equal to $N$. It gives an "$N$" component vector. Apply FFT on this time-series numerical representation to produce $N$ conjugate complex numbers
    calculate the magnitude of these FFT values and takes the square to get the spectrum of length $N/2$
    apply the ICD method to reduce the length further by taking the absolute differences between every pair of consecutive spectrum values
    then divide it by the sequence's norm to get the protein sequence's descriptor
    $i: = i + 1$
end while
calculate distance matrix of size $m \times$ m from these m number of descriptors using Euclidean distance
as a last step phylogenetic tree is formed by applying UPGMA Algorithm using MEGA11 software from this distance matrix
end FFT-ICD
```

3. RESULTS AND DISCUSSION

Distance matrix is calculated using the proposed technique, applying molecular weight and hydropathy index (polarity) as physiochemical properties for 9 ND4 protein sequences. Then, phylogenetic trees are constructed from these distance matrices to compare two physiochemical properties, molecular weight and hydropathy index (polarity). It is found that the polarity value produced the best results produces with the molecular weight. The same is verified with 9 ND5 and 9 ND6 protein sequences as well. Phylogenetic trees produced for these 9 ND4, 9 ND5, and 9 ND6 protein sequences are presented in Figures 1−6. The proposed method is also used in the similarity study of 12 baculoviruses and 24 TF proteins using polarity (hydropathy index) as the physiochemical property, which seems to be the most promising property between molecular weight and hydropathy index (polarity). Phylogenetic trees produced using the proposed method for 12 baculovirus and 24 TF proteins are shown in Figures 7 and 8, respectively.

3.1. Qualitative Assessment. ClustalW is a commonly used technique for aligning protein sequences, although traditional whole-genome alignment-based methods appear to be time-consuming. However, ClustalW offers a good reference platform for comparing the viability of phylogenetic trees obtained through any faster non-alignment-based approach. In this work, we first construct the phylogenetic trees for 9 ND4, 9 ND5, and 9 ND6 protein sequences by the proposed method using polarity value and molecular weight separately. Then, we calculate the symmetric distance (SD) in reference with the phylogenetic tree constructed by ClustalW, as shown in Table 3 for all the data set. Phylogenetic trees

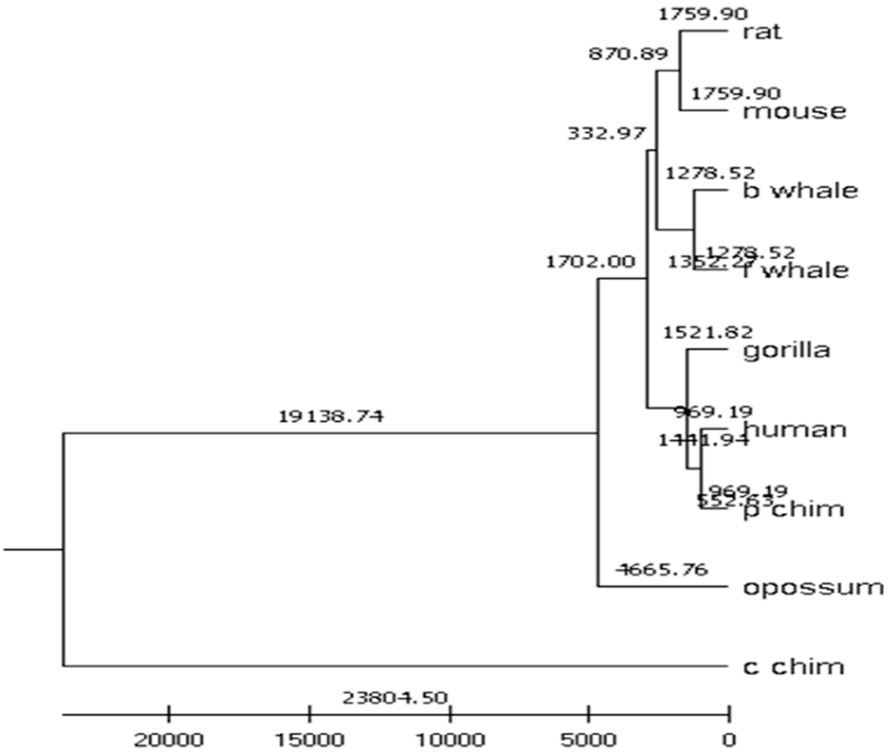


Figure 1. Phylogenetic tree based on molecular weight as physiochemical property for 9 ND4 protein sequences.

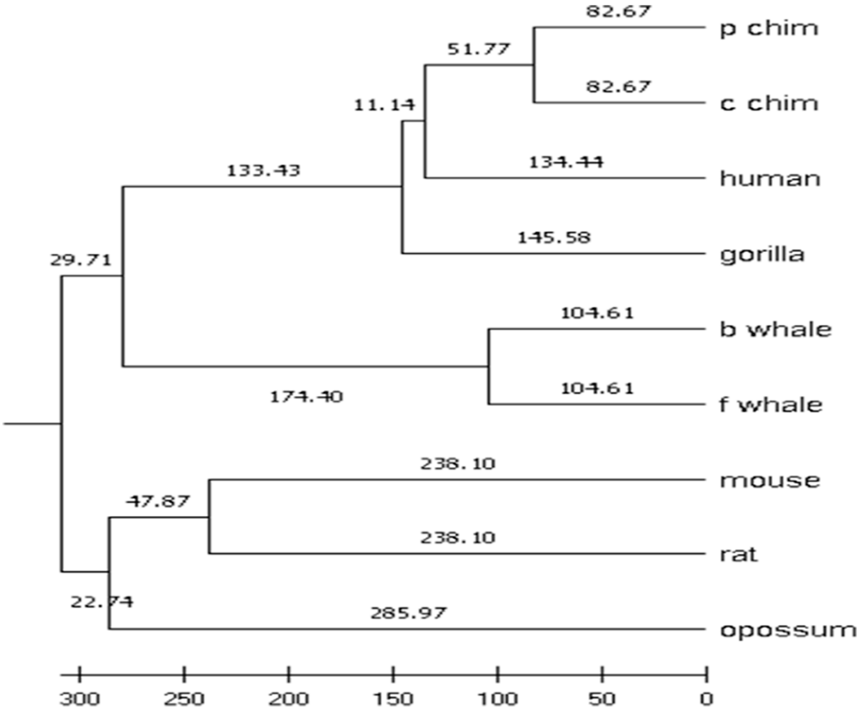


Figure 2. Phylogenetic tree based on polarity as physiochemical property for 9 ND4 protein sequences.

using the ClustalW method are constructed by using the method available at https://www.genome.jp/tools-bin/clustalw. It is found that in all the cases of 9 ND4, 9 ND5, and 9 ND6, phylogenetic tree produced by the proposed method using the polarity value is similar to that of the ClustalW method as the SD values are zero. Therefore, as a next step, we analyze the similarity study of 12 baculovirus and 24 TF protein sequences by calculating the SD value in reference with the ClustalW method and subsequently comparing with two established methods, as shown in Table 4. It is found that our method using polarity value outperforms in comparison with the compared methods. Based on the proposed method, the topology of the phylogenetic tree for 12 baculoviruses and 24 TF proteins agree with the known biological reference. The phylogenetic tree so produced

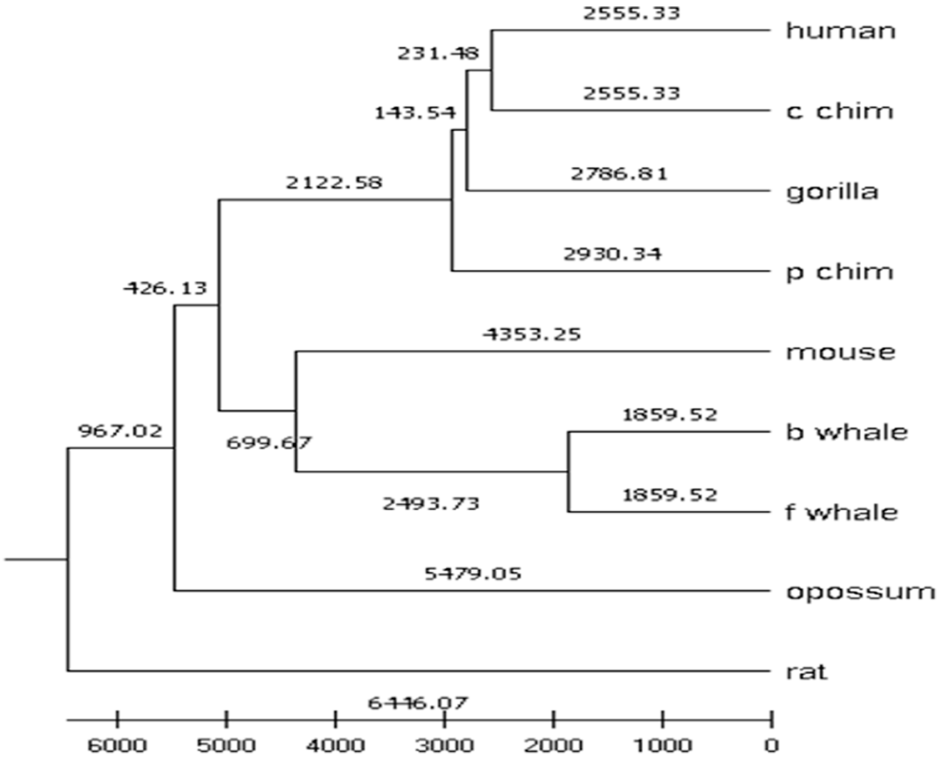


Figure 3. Phylogenetic tree based on molecular weight as the physiochemical property for 9 ND5 protein sequences.

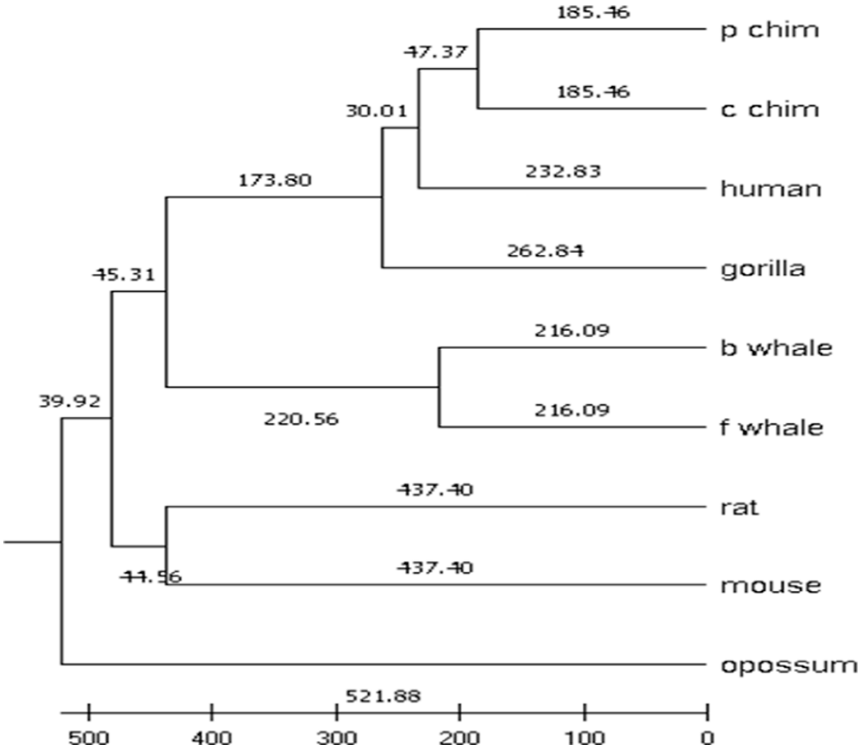


Figure 4. Phylogenetic tree based on polarity as the physiochemical property for 9 ND5 protein sequences.

suggests that the proposed method effectively infers evolutionary analysis of species.

3.2. Quantitative Assessment. The SD is famous for calculating distances between trees based on their topologies. A tree's branches are separated into two distinct groups. The treedist program in the PHYUP package proposed in ref 40 is applied to obtain SD between two trees. The number of partitions that are not shared by both trees is counted using the SD approach. Table 4 lists the SD of the phylogenetic trees constructed based on various strategies. It is evident from Table 4 that the SD of the proposed method is smaller than the method proposed by Yao et al.[19,41] and Yu et al.[20]

3.3. Computational Speed. Table 5 displays the computing time of the proposed approach for computing the

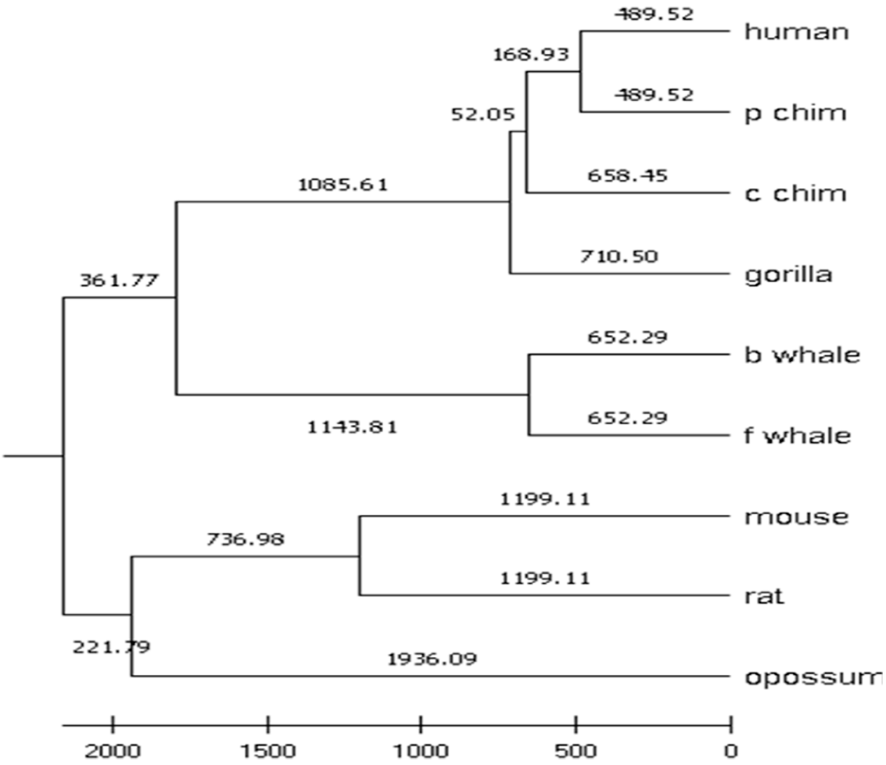


Figure 5. Phylogenetic tree based on molecular weight as the physiochemical property for 9 ND6 protein sequences.

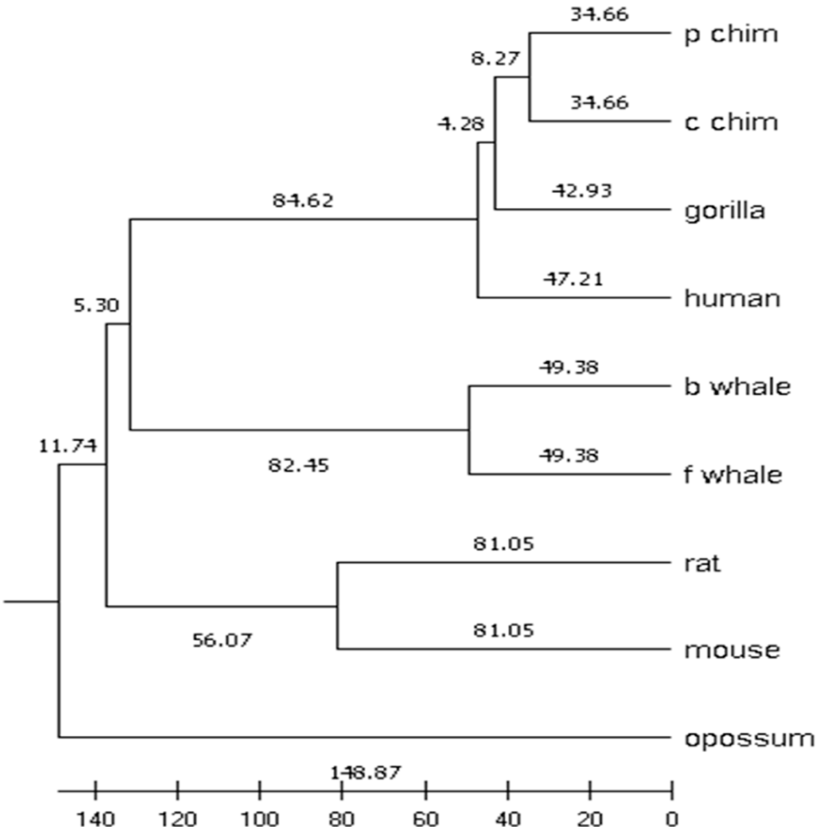


Figure 6. Phylogenetic tree based on polarity as the physiochemical property for 9 ND6 protein sequences.

distance matrix. To compute the overall execution time, 20 sample runs with the same data set are performed, and the average is calculated to report the required time. This procedure is repeated for each data set to establish our proposed model's computational efficiency. Hardwarewise, an Intel Core i5 CPU with 4 GB of RAM is utilized for all computational tasks. Matlab R2016b, a standard tool for mathematical calculations, is used to implement the proposed

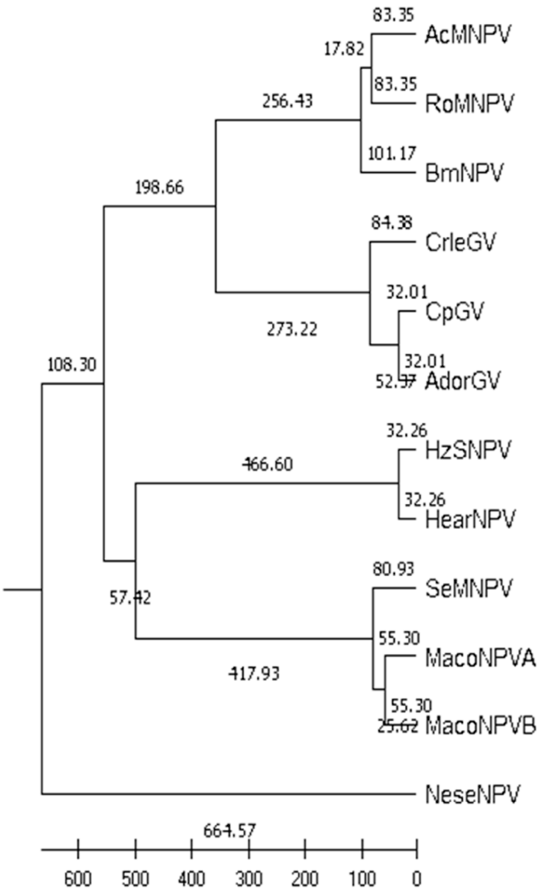


Figure 7. Phylogenetic tree based on polarity as the physiochemical property for 12 baculovirus.

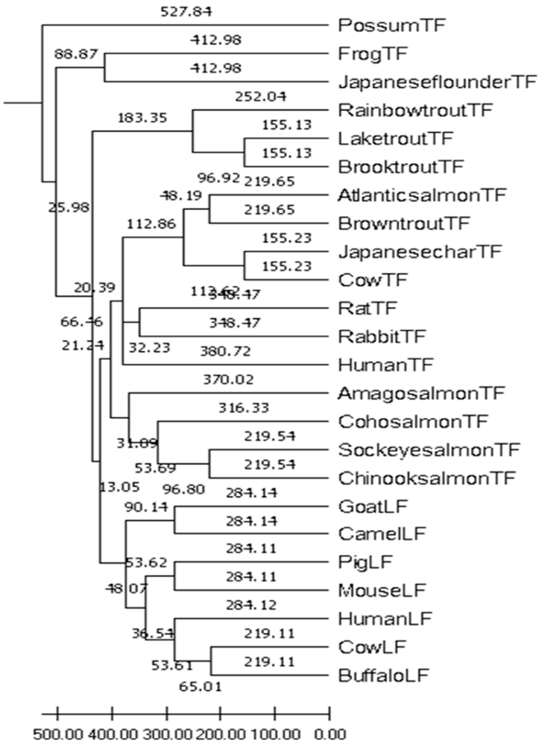


Figure 8. Phylogenetic tree based on polarity as the physiochemical property for 24 TF proteins.

Table 3. SD Value of the Proposed Method in Reference with ClustalW

data set	SD: proposed method (polarity)	SD: proposed method (molecular weight)
9 ND4	0	3
9 ND5	0	3
9 ND6	0	2

Table 4. SD Value of the Proposed Method Using Polarity in Reference with ClustalW and Other Methods

data set	SD: proposed method	SD: other methods	compared method
baculovirus	2	4	Yao et al. (2013)[19]
baculovirus	2	8	Yao et al. (2014)[41]
TF protein	8	12	Yu et al. (2017)[20]

Table 5. Computational Time of the Distance Matrix of the Proposed Method

data set used	time required (in s)
baculovirus	0.2116
TF protein	0.25249

model. Observations indicate that the proposed model requires significantly less time to construct the distance matrix.

3.4. Similarity Study of 12 Baculovirus.

The proposed method is also used to investigate 12 baculovirus, with the results displayed as a phylogenetic tree in Figure 7. The 12 baculovirus are grouped into four categories. Group-1 contains AcMNPV, BmNPV, and RoMNPV, which form the *Alphabaculovirus* group. Group-2 contains HearNPV, HzSNPV, MacoNPVA, MacoNPVB, and SeMNPV. Group-3 contains AdorGV, CpGV, and CrleGV, which constitute the *Betabaculovirus* group. Finally, group-4 contains only one virus, NeseNPV, called *Gammabaculovirus*. The proposed method correctly clusters all the groups prominently compared with the method described in ref 19 using the geometrical center and sequence-segmented method with $k = 5$ and in ref 41 using a new spectrum-like graph representation. This result establishes the usefulness of the proposed method.

3.5. Similarity Study of 24 TF Proteins.

24 TF protein sequences are known as 24 vertebrates taken from the NCBI database. As demonstrated in Figure 8, TF and lactoferrin proteins are easily separated using the proposed technique. These results are compared with those obtained in ref 20 using position-feature energy matrix and reflect better outcomes. This result shows the applicability of the proposed method. It is also at par with the known biological reference.

4. ADVANTAGE OF THE PROPOSED METHOD

From the above discussion, the advantage of the proposed method can be summarized as follows

(a) It is computationally efficient in terms of time complexity which is $O(N \log N)$.

(b) It can handle extensive data sets; the present paper efficiently analyses the sequence of 12 baculovirus and 24 TF proteins of the maximum length being 1253.

(c) It applies equally to protein sequences of equal and unequal length.

(d) proposed method is capable of producing evolutionary relationship among different species using only polarity value.

(e) It produces better results than those obtained in ref 19 20, and 41.

5. CONCLUSIONS

Researchers have developed several comparison methods based on physiochemical properties for decades. In such studies, a maximum of 12 physiochemical properties and a minimum of two chemical properties are used in protein sequence comparison. In this work, an attempt has been made to establish whether a single property like polarity or molecular weight is sufficient to give the same classification. This is carried out using FFT on the represented protein sequence based on polarity value and molecular weight separately. For comparing the results, phylogenetic trees are constructed. Phylogenetic trees so produced based on polarity are found to be consistent among all the species, which is not the case for molecular weight. This outcome validates the known fact that polarity is more significant than other physiochemical properties. Therefore, the proposed method uses polarity values to represent protein sequences of 12 baculovirus and 24 TF proteins. Again, the phylogenetic trees so produced for this protein sequence by the proposed method using polarity values alone are found to be similar, as compared with their known biological reference. Therefore, it can be concluded that using FFT spectrum analysis on a numerical time series obtained from polarity values of amino acids to compare protein sequences is a very significant attempt as it can deal with the protein sequences of varying lengths. The developed method is also time efficient which is $O(N \log N)$, considering N as the maximum length of the protein sequence under comparison.

AUTHOR INFORMATION

Corresponding Author

Jayanta Pal – *Department of ECE, National Institute of Technology, Durgapur 713209, India; Department of CSE, Narula Institute of Technology, Kolkata 700109, India;* orcid.org/0000-0001-9300-2248; Email: jayanta.pal@gmail.com

Authors

Soumen Ghosh – *Department of IT, Narula Institute of Technology, Kolkata 700109, India*

Bansibadan Maji – *Department of ECE, National Institute of Technology, Durgapur 713209, India*

Dilip Kumar Bhattacharya – *Department of Pure Mathematics, University of Calcutta, Kolkata 700073, India*

Complete contact information is available at:
https://pubs.acs.org/10.1021/acsomega.2c06103

Funding

We wish to confirm that there has been no financial support for this work that could have influenced its outcome.

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Zielezinski, A.; Vinga, S.; Almeida, J.; Karlowski, W. M. Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biol.* **2017**, *18*, 186.

(2) Bernard, G.; Chan, C. X.; Chan, Y. B.; Chua, X. Y.; Cong, Y.; Hogan, J. M.; Maetschke, M. A.; Ragan, M. A. Alignment-free inference of hierarchical and reticulate phylogenomic relationships. *Briefings Bioinf.* **2019**, *20*, 426–435.

(3) Just, W. Computational complexity of multiple sequence alignment with SP-score. *J. Comput. Biol.* **2001**, *8*, 615–623.

(4) Phillips, A.; Janies, D.; Wheeler, W. Multiple sequence alignment in phylogenetic analysis. *Mol. Phylogenet. Evol.* **2000**, *16*, 317–330.

(5) Katoh, K.; Misawa, K.; Kuma, K. I.; Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **2002**, *30*, 3059–3066.

(6) Vinga, S.; Almeida, J. Alignment-free sequence comparison—a review. *Bioinformatics* **2003**, *19*, 513–523.

(7) Pinello, L.; Lo Bosco, G.; Yuan, G. C. Applications of alignment-free methods in epigenomics. *Briefings Bioinf.* **2014**, *15*, 419–430.

(8) Domazet-Lošo, M.; Haubold, B. Alignment-free detection of local similarity among viral and bacterial genomes. *Bioinformatics* **2011**, *27*, 1466–1472.

(9) Ghosh, S.; Pal, J.; Maji, B.; Bhattacharya, D. K. Condensed Matrix Descriptor for Protein Sequence Comparison. *Int. J. Anal. Mass Spectrom. Chromatogr.* **2016**, *4*, 1–13.

(10) Chan, R. H.; Wang, R. W.; Yeung, H. M. Composition vector method for phylogenetics—a review. *9th International Symposium on Operations Research and Its Applications*; ORSC & APORC: Chengdu, China, August, 2010; p 13–20.

(11) Chu, K. H.; Qi, J.; Yu, Z. G.; Anh, V. O. Origin and phylogeny of chloroplasts revealed by a simple correlation analysis of complete genomes. *Mol. Biol. Evol.* **2004**, *21*, 200.

(12) Gao, L.; Qi, J. Whole genome molecular phylogeny of large dsDNA viruses using composition vector method. *BMC Evol. Biol.* **2007**, *7*, 41.

(13) Yu, C.; He, R. L.; Yau, S. S. T. Protein sequence comparison based on K-string dictionary. *Gene* **2013**, *529*, 250–256.

(14) Yu, C.; Deng, M.; Yau, S. S. T. DNA sequence comparison by a novel probabilistic method. *Inf. Sci.* **2011**, *181*, 1484–1492.

(15) Gupta, M. K.; 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*, 519–532.

(16) El-Lakkani, A.; Lashin, M. An efficient method for measuring the similarity of protein sequences. *SAR QSAR Environ. Res.* **2016**, *27*, 363–370.

(17) Li, C.; Xing, L.; Wang, X. 2-D graphical representation of protein sequences and its application to coronavirus phylogeny. *BMB Rep.* **2008**, *41*, 217–222.

(18) Ghosh, S.; Pal, J.; Maji, B.; Bhattacharya, D. K. A sequential development towards a unified approach to protein sequence comparison based on classified groups of amino acids. *Int. J. Eng. Technol.* **2018**, *7*, 678.

(19) Kong, F.; Yao, Y. H.; Dai, Q.; He, P. A. A sequence-segmented method applied to the similarity analysis of long protein sequence. *MATCH Commun. Math. Comput. Chem.* **2013**, *70*, 431–450.

(20) Yu, L.; Zhang, Y.; Gutman, I.; Shi, Y.; Dehmer, M. Protein sequence comparison based on physicochemical properties and the position-feature energy matrix. *Sci. Rep.* **2017**, *7*, 46237.

(21) Wu, Z. C.; Xiao, X.; Chou, K. C. 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.* **2010**, *267*, 29–34.

(22) Randić, M. 2-D graphical representation of proteins based on physicochemical properties of amino acids. *Chem. Phys. Lett.* **2007**, *440*, 291–295.

(23) Zhang, Y.; Zhan, Y.; Xu, C. A novel method of 2D graphical representation for proteins and its application. *MATCH Commun. Math. Comput. Chem.* **2016**, *75*, 431–446.

(24) Qi, Z. H.; Jin, M. Z.; Li, S. L.; Feng, J. A protein mapping method based on physicochemical properties and dimension reduction. *Comput. Biol. Med.* **2015**, *57*, 1–7.

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

(26) Yu, C.; Cheng, S. Y.; He, R. L.; Yau, S. S. T. Protein map: an alignment-free sequence comparison method based on various properties of amino acids. *Gene* **2011**, *486*, 110.

(27) Zhang, Y. P.; Ruan, J. S.; He, P. A. Analyzes of the similarities of protein sequences based on the pseudo amino acid composition. *Chem. Phys. Lett.* **2013**, *590*, 239−244.

(28) Ma, T.; Liu, Y.; Dai, Q.; Yao, Y.; He, P. A. A graphical representation of protein based on a novel iterated function system. *Phys. A* **2014**, *403*, 21−28.

(29) Ping, P.; Zhu, X.; Wang, L. Similarities/dissimilarities analysis of protein sequences based on PCA-FFT. *J. Biol. Syst.* **2017**, *25*, 29−45.

(30) Mahmoodi-Reihani, M.; Abbasitabar, F.; Zare-Shahabadi, V. A novel graphical representation and similarity analysis of protein sequences based on physicochemical properties. *Phys. A* **2018**, *510*, 477−485.

(31) Mahmoodi-Reihani, M.; Abbasitabar, F.; Zare-Shahabadi, V. In silico rational design and virtual screening of bioactive peptides based on QSAR modeling. *ACS Omega* **2020**, *5*, 5951−5958.

(32) Wu, Y. L., Agrawal, D.; El Abbadi, A. A comparison of DFT and DWT based similarity search in time-series databases. *9th Int. Conf. Inf. Knowl. Manag. Proc.* November, 2000; pp 488−495.

(33) Yin, C.; Yau, S. S. T. Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence. *J. Theor. Biol.* **2007**, *247*, 687−694.

(34) Tiwari, S.; Ramachandran, S.; Bhattacharya, A.; Bhattacharya, S.; Ramaswamy, R. Prediction of probable genes by Fourier analysis of genomic sequences. *Bioinformatics* **1997**, *13*, 263−270.

(35) King, B. R.; Aburdene, M.; Thompson, A.; Warres, Z. Application of discrete Fourier inter-coefficient difference for assessing genetic sequence similarity. *EURASIP J. Bioinf. Syst. Biol.* **2014**, *2014*, 8.

(36) Hoang, T.; Yin, C.; Zheng, H.; Yu, C.; Lucy He, R. L.; Yau, S. S. T. A new method to cluster DNA sequences using Fourier power spectrum. *J. Theor. Biol.* **2015**, *372*, 135−145.

(37) Pal, J.; Ghosh, S.; Maji, B.; Bhattacharya, D. K. Use of FFT in protein sequence comparison under their binary representations. *Comput. Mol. Biosci.* **2016**, *06*, 33.

(38) Aamir, K. M.; Maud, M. A.; Loan, A. On Cooley-Tukey FFT method for zero padded signals. *Proceedings of the IEEE Symposium on Emerging Technologies*; IEEE, September, 2005; pp 41−45.

(39) Tamura, K.; Stecher, G.; Kumar, S. MEGA11: molecular evolutionary genetics analysis version 11. *Mol. Biol. Evol.* **2021**, *38*, 3022−3027.

(40) Felsenstein, J. *PHYLIP (phylogeny inference package)* Distributed by the author. Department of Genome Sciences, University of Washington: Seattle Version, 3, 2005.

(41) Yao, Y.; Yan, S.; Xu, H.; Han, J.; Nan, X.; He, P. A.; Dai, Q. Similarity/dissimilarity analysis of protein sequences based on a new spectrum-like graphical representation. *Evol. Bioinf.* **2014**, *10*, 87.

https://doi.org/10.1021/acsomega.2c06103