SCIENTIFIC REPORTS

OPEN

Alignment-free similarity analysis for protein sequences based on fuzzy integral

Received: 12 March 2018
Accepted: 15 January 2019
Published online: 26 February 2019

Ajay Kumar Saw[1], Binod Chandra Tripathy[2] & Soumyadeep Nandi[3]

Sequence comparison is an essential part of modern molecular biology research. In this study, we estimated the parameters of Markov chain by considering the frequencies of occurrence of the all possible amino acid pairs from each alignment-free protein sequence. These estimated Markov chain parameters were used to calculate similarity between two protein sequences based on a fuzzy integral algorithm. For validation, our result was compared with both alignment-based (ClustalW) and alignment-free methods on six benchmark datasets. The results indicate that our developed algorithm has a better clustering performance for protein sequence comparison.

With the advent of the advanced sequencing techniques, researchers are generating a large number of protein sequences. This brings in a new challenge[1,2] for phylogenetic and comparative study of these protein sequences. Phylogenetic study and comparative analysis between taxa are an essential part of molecular biology and bioinformatics. These studies, traditionally depended on multiple or pairwise sequence alignments which are the well established classical approach and regarded as a standard method for sequence analysis. However, producing reliable multiple sequence alignments become extremely difficult when more dissimilar protein sequences are considered. The traditional alignment-based methods[3–5] are much empirical to select and create a sequence alignment score matrix, and variation of which may affect the alignment results. Various alignment-free tools[6–13] have been developed over the past two decades to overcome the alignment complexity for phylogenetic analysis. An alignment-free approach consist of two steps for comparing protein sequences. At the first step, the protein sequences are converted into a fixed-length feature vectors. Feature extraction is a series of process for extracting the required information from the query sequences, which is critical for the accuracy of an alignment-free method. At the second step, these extracted feature vectors are used as an input data in vectors similarity comparison algorithm to perform downstream analysis like phylogenetic analysis. Methods based on graphical representation, distance frequency matrix, numerical characterization, K-string dictionary etc., have been introduced to overcome the complication of the sequence alignment. Graphical representation[14,15] of protein sequences provides a simple way of viewing, sorting and comparing various sequences. It also provides mathematical descriptor which help in identifying differences among similar protein sequences quantitatively. Distance frequency of amino acid pairs suggest a new numerical characterization of protein sequence, which converts protein sequence into a distance frequency matrix[16]. Numerical characterization directly extracted from protein sequence would capture the essence of the amino acid composition and their distribution on the protein sequence in a quantitative aspect. In this approach, each sequence is mapped into a vector or matrix based on the numerical characterization extracted from the protein sequence. Subsequently, a similarity score is calculated by following distance measure tools, such as, Euclidean distance, Cosine distance, Manhattan distance, etc., among their corresponding vectors or matrices. K-string dictionary[17] approach permit users to use a much lower dimensional frequency or probability vector to represent a protein sequence. It also significantly reduces the space requirement for their implementation. Furthermore, after getting the lower dimensional frequency vectors, Singular Value Decomposition (SVD) is used to get a better protein vector representation which helps user to obtain a precise phylogenetic tree. However, these above mentioned methods are lagging behind in terms of accuracy. Thus, more discriminatory features are still needed to be developed. In addition to the accuracy, these method have another drawback and that is, computational complexity. Motivated by the aforementioned work, in this study, we proposed to use fuzzy integral algorithm[18,19] for analysis of protein sequence based on Markov chain[20]. Fuzzy integral

[1]Institute of Advanced Study in Science and Technology, Mathematical Sciences Division, Guwahati, 781035, India. [2]Tripura University, Department of Mathematics, Agartala, 799022, India. [3]Institute of Advanced Study in Science and Technology, Life Science Division, Guwahati, 781035, India. Correspondence and requests for materials should be addressed to S.N. (email: soumyadeep.nandi@gmail.com)

similarity[21,22] method assigns similarity score within the closed interval [0, 1] between two protein sequences. A protein sequence consists of twenty amino acids. By taking these 20 amino acids as a state space $M = \{A, I, L, M, F, P, W, V, D, E, N, C, Q, G, S, T, Y, R, H, K\}$, we have used $k^{th}-$step transition probability matrix, fuzzy measure[23], fuzzy integral to describe protein sequence. We have used fuzzy integral similarity for getting distance matrix, which is used in neighbor program in PHYLIP package[24] for constructing a phylogenetic tree. The advantage of our method is, it do not require any prior knowledge of homologous relationship (common ancestry) among the sequences, which makes it fully automated and robust. For validation of our developed algorithm, we implemented our approach on NADH Dehydrogenase-5 protein sequences, NADH Dehydrogenase-6 protein sequences, xylanases protein sequences in the F10 and G11 datasets, transferrin protein sequences, coronavirus spike protein sequences and beta-globin protein sequences. We compared the tree generated by our method with the trees generated by both alignment-free method, and alignment-based ClustalW method using MEGA package[25]. In addition, we used few standard statistical tools such as correlation coefficient (CC), Robinson-Foulds distance (RF-distance)[26] and receiver operating characteristic (ROC)[27–29] curve to compare distance matrices generated by our method with the other alignment-free methods. The main purpose of this study is to compare the performance among alignment-based and alignment-free protein clustering methods and to identify their strengths and weakness from the practical perspectives of the users.

Methods

Markov chain for protein sequence. Let $P = [p_{i,j}]$ represent the transition probability matrix of a discrete-time Markov chain[20]. Transition probability $p_{i,j}$ can be defined as follows:

$$p_{i,j} = p(Z_{n+1} = a_j | Z_n = a_i), \ \ 1 \leq i, \ j \leq M, \tag{1}$$

where Z_n represent the actual state at time $n (n = 1, 2, 3 \ldots)$, a_i is the i_{th} state within 20 distinct states. In the context of protein sequence, the number of states is $M = 20$, which corresponds to the twenty amino acids symbol set $M = \{A = a_1, I = a_2, L = a_3, M = a_4, F = a_5, P = a_6, W = a_7, V = a_8, D = a_9, E = a_{10}, N = a_{11}, C = a_{12}, Q = a_{13}, G = a_{14}, S = a_{15}, T = a_{16}, Y = a_{17}, R = a_{18}, H = a_{19}, K = a_{20}\}$. The state transition probabilities satisfy the following constraints

$$p_{i,j} \geq 0 \ \forall i, j \text{ and } \sum_{j=1}^{M} p_{i,j} = 1 \ \forall i.$$

We calculated the transition probability matrices based on the observed sequences. From each alignment-free protein sequence, we assumed that the frequency of occurrences of all possible amino acid pairs as the parameters of Markov chain. If $N_{a_i a_j}$ denotes the total number of adjacent amino acid pair (a_i, a_j), then $1^{st}-$step transition probability matrix from the state a_i to the state a_j is given by

$$p_{i,j} = \frac{N_{a_i a_j}}{\sum_{j=1}^{M} N_{a_i a_j}} \tag{2}$$

Above explanation is the 1^{st} step Markov chain and the k^{th} step Markov chain can be obtained through the 1^{st} step Markov chain. Let $P^k = [p^k_{i,j}]$ denote the transition probability matrix of a discrete-time Markov chain starting from state i after k steps to end with state j. Each state transition probability $p^k_{i,j}$ is given as follows:

$$p^k_{i,j} = p^k(Z_{n+k} = a_j | Z_n = a_i), \ \ 1 \leq i, \ j \leq M, \tag{3}$$

satisfy following constraints

$$p^k_{i,j} \geq 0 \ \forall i, j \text{ and } \sum_{j=1}^{M} p^k_{i,j} = 1 \ \forall i.$$

For three sets U, V and W, the following condition holds: $p[U \cap V | W] = p[U | V \cap W] p[V | W]$. Interpreting U as $Z_{n+k} = a_j$, V as $Z_{n+t} = a_r$ and W as $Z_n = a_i$, we have

$$\begin{aligned} p^k_{i,j} &= p[Z_{n+k} = a_j | Z_n = a_i] \\ &= \sum_{a_r \in \mathbf{M}} p[Z_{n+k} = a_j, Z_{n+t} = a_r | Z_n = a_i] \\ &= \sum_{a_r \in \mathbf{M}} p[Z_{n+k} = a_j | Z_{n+t} = a_r, Z_n = a_i] \\ &\quad \times p[Z_{n+t} = a_r | Z_n = a_i] \\ &= \sum_{a_r \in \mathbf{M}} p[Z_{n+k} = a_j | Z_{n+t} = a_r] \\ &\quad \times p[Z_{n+t} = a_r | Z_n = a_i] \\ &= \sum_{a_r \in \mathbf{M}} p^{k-t}_{r,j} p^t_{i,r}, \end{aligned} \tag{4}$$

which is known as the Chapman-Kolmogorov equation.

Hence, the matrix with element $p_{i,j}^k$ are $[p_{i,j}^k] = P^k$.

In the context of protein sequence, k^{th}-step transition probability matrix can be expressed as:

$$P^k = \begin{bmatrix} p_{1,1}^k & p_{1,2}^k & \cdots & \cdots & p_{1,20}^k \\ p_{2,1}^k & p_{2,2}^k & \cdots & \cdots & p_{2,20}^k \\ \vdots & \vdots & \cdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \cdots & \vdots \\ p_{20,1}^k & p_{20,2}^k & \cdots & \cdots & p_{20,20}^k \end{bmatrix}_{20\times 20}$$

which is subjected to $p_{i,j}^k \geq 0\ \forall i, j \in \{1, 2, \ldots 20\}$ and $\sum_{j=1}^{20} p_{i,j}^k = 1\ \forall i$. The $p_{i,j}^k$ can be determined by the equations (2) and (4). After the derivation of the k^{th}-step transition probability matrix, we optimized the step $k = h$, which is least positive integer under satisfying following condition for each protein sequence:

$$rmsd(P^{k=h} - P^{k=h+1}) \approx 0 \text{ (upto six decimal place)}, \tag{5}$$

where *rmsd* represent root mean square distance between two consecutive transition probability matrices. After optimizing the step of transition probability matrix, i.e, P^h. We noted that, all 20 rows in optimized transition probability matrix are approximately identical with each other. Therefore, we took a single row from the transition probability matrix P^h as a input for further investigation, which reduced our time complexity.

Fuzzy integral and fuzzy measure for the h^{th}– step amino acids sequence. Let $G = \{(ba_i)^h = x_i | i \in \{1, 2, 3, \ldots, 20\}, b \in \mathbf{M}\}$ be the finite set of h^{th}–step amino acids starting from amino acid b and ending with amino acid a_i, estimated from protein sequence. The finite set G is termed as feature vector.

Let $\nu, \tau \subseteq G$ and $R(G)$ be the power set of G. A fuzzy measure μ is a real valued function:

μ: $R(G) \rightarrow [0, 1]$, satisfy the following condition,

(i) $\mu(\phi) = 0$ and $\mu(G) = 1$
(ii) $\mu(\nu) \leq \mu(\tau)$ *if* $\nu \subseteq \tau$.

For a fuzzy measure μ, let $\mu(x_i) = \mu^i\ \forall\ x_i \in G$. The mapping $x_i \rightarrow \mu^i$ is known as fuzzy density function. The fuzzy density of single element $x_i \in G$, μ^i can be interpreted as the importance of x_i in determining the set G. Based on the fuzzy measure definition μ, the measure of a subset is not just only the summation of the measure of its elements but also included the measure of each combination. This information could be delivered by an expert or observed through the problem. However, when handing with larger set, this job may become computationally complex, difficult or even not feasible. λ-measures is the possible solution for solving this problem. λ-fuzzy measure[30] fulfills the criteria of fuzzy measure plus some additional property: for all $\nu, \tau \subset G$, $\nu \cap \tau = \phi$ and

$$\mu(\nu \bigcup \tau) = \mu(\nu) + \mu(\tau) + \lambda\mu(\nu)\mu(\tau), \quad \text{for some } \lambda > -1. \tag{6}$$

Furthermore, λ can be obtained by solving following equation:

$$\lambda + 1 = \prod_{i=1}^{20} (1 + \lambda\mu^i). \tag{7}$$

Therefore, we can construct fuzzy measure by applying equation(6) and equation(7), for this we only need to know the individual fuzzy densities of the elements μ^i ($\forall\ i \in \{1, 2, 3, \ldots, 20\}$).

Let ρ:$G \rightarrow [0, 1]$ represent a function that maps every element of G to its evidence. The function ρ must satisfy descending order, which is as follows: $\rho(x_1) \geq \rho(x_2) \geq \rho(x_3) \geq \ldots \geq \rho(x_{20})$. If suppose ρ function does not satisfy the above condition, then reorder G so that ρ function must satisfy descending order condition and we will proceed further calculation based on the modified descending order condition. Let μ:$R(G) \rightarrow [0, 1]$ be a fuzzy measure. Then the fuzzy integral of ρ with respect to the fuzzy measure μ is given by

$$I = max[min[\rho(x_i),\ \mu(A_i)]_{i=1}^{20}], \tag{8}$$

$$\text{where } A_i = \{x_1,\ x_2,\ \ldots, x_i\}. \tag{9}$$

The fuzzy integral examine the fact supplied by each element of a given set, and the assessment of each subset of elements (using a fuzzy measure) in its decision-making process. The combination of the important significance of the source and the extracted information makes the fuzzy integral appropriate for information fusion. This theory has capability to tackle uncertainties associated with issue related to the processing procedures and data extraction. Therefore, this theory has been extensively applied in pattern recognition[31] and classification.

Fuzzy integral similarity and distance matrix for protein sequence comparison. The fuzzy integral similarity is based on the h^{th}–step amino acids frequencies between the feature vector of the two sequences. Let ν and τ are feature vectors of the two sequences. We define fuzzy integral function ρ, which is given as:

$$\rho(x_i) = 1 - |x_i^\nu - x_i^\tau|, \tag{10}$$

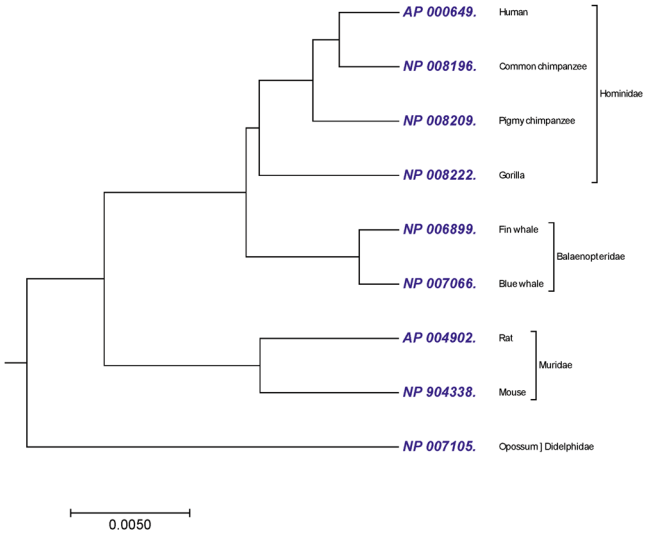

Figure 1. The phylogenetic tree of 9 sequences of NADH Dehydrogenase 5 protein constructed by our method using Fitch-Margoliash approach.

Methods	Correlation coefficients	Robinson-Foulds distance (RF-distance)
Our method	0.7378	2
Jayanta *et al.* (without grouping)[34] (Table 7*)	0.9734	0
Li *et al.*[37] (Table 4*)	0.962	0
Jayanta *et al.* (with grouping)[34] (Table 8*)	0.9403	4
Ma *et al.*[13] (Table 3*)	0.9304	0
Wen *et al.*[35] (Table 3*)	0.7324	4
Yao *et al.*[36] (Table 3*)	0.6908	4
Czerniecka *et al.*[38] (Table 8*)	0.61840	10

Table 1. Comparison of alignment-free methods with the ClustalW based on correlation coefficient (CC) and Robinson-Foulds distance (RF-distance) on the ND 5 dataset.

where $x_i \in G$ (i.e., the similarity of the h^{th}–step amino acid frequency x_i in the two feature vectors ν and τ).

Using fuzzy measure, we can determine the relative importance of subsets of amino acids being considered. Taking benefit of the λ–fuzzy measure properties described above, we can formulate μ using the fuzzy density of the individual element μ^i.

In this case,

$$\mu^i = max(x_i^{\nu}, x_i^{\tau}), \tag{11}$$

where $x_i \in G$ (i.e., the maximum level of h^{th}–step amino acid frequency starting from amino acid b and ending with amino acid a_i between two feature vectors with respect to their assigned position). Using equation (7), we calculated the value of λ and put the λ value in equation(6) to obtain the fuzzy measure μ. It can be easily verified that μ satisfy the properties (*i*) and (*ii*) of the fuzzy measure. Once we have ρ and μ, it is a straight forward using equation(8) to obtain the fuzzy integral[21].

Next we calculate difference between two feature vectors ν and τ, which is given as follows:

$$D(\nu, \tau) = 1 - I(\nu, \tau), \tag{12}$$

where $I(\nu, \tau)$ is fuzzy integral similarity between ν and τ

The above process is continued for all pairwise combinations taken from n number of protein sequences. Finally, a distance matrix was generated. This distance matrix contained the dissimilarity information related to n protein sequences. This distance matrix was used as an input data to the neighbor.exe program in PHYLIP package[24] for phylogenetic tree construction.

Algorithm

This section explains an algorithmic view of the developed method. The complete algorithm consists three stages.

Stage 1: Calculation of optimal-step transition probability matrix using Markov chain estimated from observed protein sequences:

Algorithm 1. Derivation of h^{th}-step transition probability matrix.

1. **read** all protein sequences **P** from input file
2. amino acids $\leftarrow \{a_1, a_2, ..., a_{20}\}$
3. **while** (**P**!=NULL) **do**
4. l $\leftarrow$ string length of Protein
5. **set** Protein array **P**[l]
6. **repeat** all possible amino acid pairs **do**
   i. c ,d $\in$ amino acids
   ii. **set** n[c][d] $\leftarrow 0$
   iii. **for**($i = 0$ to $i = l - 2\ step$) **do**
   iv. $j \leftarrow i + 1$
   v. **if**(**P**[i]==c and **P**[j]==d) **do**
   vi. n[c][d] $\leftarrow$ n[c][d]+1
   vii. **end if**
   viii. **end for**
7. **end repeat**
8. **for** $(1 <= i, j <= 20)$ **do**
9. **store** n$[a_i][a_j]$
10. $p[i][j] \leftarrow n[a_i][a_j] / \sum_{j=1}^{20} n[a_i][a_j]$
11. **end for**
12. $P \leftarrow [p_{ij}]_{1<=i,j<=20}$
13. $P^1 \leftarrow P$ known as 1^{st}-step
14. $P^2 \leftarrow P * P$ known as 2^{nd}-step
15. $P^k \leftarrow P * P..(k - times)..* P$ known as k^{th}-step
16. h_i^{th}-step $\leftarrow$ optimize k^{th}-step using equation (5) for each protein sequence
17. **return** feature vector $F_i \leftarrow h_i^{th}$-step transition probability matrix
18. **end while**

Stage 2: Fuzzy integral similarity between two feature vectors F_1 and F_2:

Algorithm 2. FISim (F_1, F_2).

1. **input** F_1, F_2
2. amino acids $\leftarrow \{a_1, a_2, ..., a_{20}\}$
3. **for** $i \in \{1, 2, 3, ..., 20\}$ **do**
4. $\rho[x_i] \leftarrow$ solve equation (10)
5. $\mu[x_i] \leftarrow$ solve equation (11)
6. $\{\rho[x_i]\}_{i=1}^{20} \leftarrow$ sort $\{\rho[x_i]\}_{i=1}^{20}$ in decreasing order
7. $\{\mu[x_i]\}_{i=1}^{20} \leftarrow \{\mu[x_i]\}_{i=1}^{20}$ positional arrangement based on $\{\rho[x_i]\}_{i=1}^{20}$
8. **end for**
9. $\lambda \leftarrow$ apply newton rapson method for solving equation (7)
10. $\{\mu(A_i)\}_{i=1}^{20} \leftarrow$ solve equation (6) using equation (9)
11. $T_i \leftarrow min[\rho(x_i), \mu(A_i)]$ for $i \in \{1, 2, 3..., 20\}$
12. FISim$(F_1, F_2) \leftarrow max_{i=1}^{20}[T_i]$
13. **return** FISim (F_1, F_2)

Stage 3: Integrate stage(1) and stage(2) for phylogenetic tree construction:

Algorithm 3. Distance matrix for phylogenetic tree construction.

1. **input** feature vectors $\{F_i\}_{i=1}^{n}$ of n protein sequences from **stage 1**
2. **thread** $\leftarrow 2$ (parallel computation)
3. **for** $(i = 1$ to $i = n - 1)$
4. **for** $(j = i + 1$ to $j = n)$ **do**
5. $d[F_j][F_i] \leftarrow d[F_i][F_j] \leftarrow 1 - (FISim(F_i, F_j))$, by **stage 2**
6. **end for**
7. $\{d[F_i][F_i]\}_{i=1}^{n} \leftarrow 0$
8. **return** distance matrix $[d[F_i][F_j]]_{n \times n}$ for phylogenetic construction using PHYLIP package.

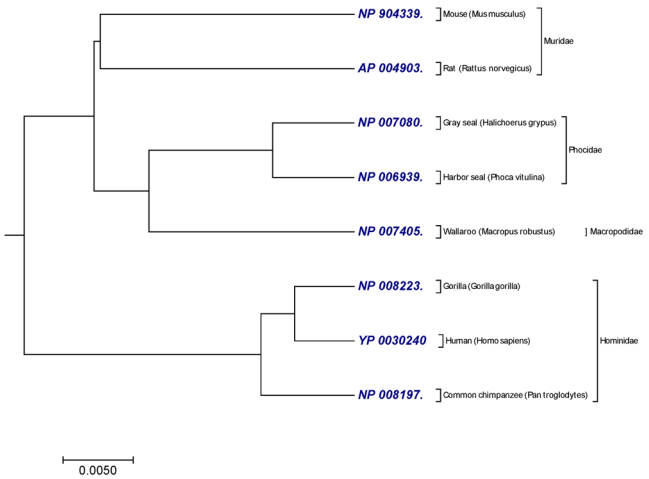

Figure 2. The phylogenetic tree of 8 sequences of NADH Dehydrogenase 6 protein constructed by our method using Fitch-Margoliash approach.

Methods	Correlation coefficients	Robinson-Foulds distance (RF-distance)
Our method	0.5982	2
Gupta *et al.*[42] (Table 2*)	0.7763	2
Czerniecka *et al.*[38] (Table 13*)	0.4609	6

Table 2. Comparison of alignment-free methods with the ClustalW based on correlation coefficient (CC) and Robinson-Foulds distance (RF-distance) on the ND 6 dataset.

Time complexities of proposed algorithm. For calculating computational complexity[32] of developed algorithm, we assumed that all operations took the same unit of time. Our algorithm was partitioned into three stages. For time complexity calculation: in the initial stage, transition probability matrices were calculated from the raw protein sequences. Time complexity of stage (1) is $O(nl + m^3\sum_{i=1}^{n} h_i)$, where n is the total number of protein sequences, m is the number of amino acids, l is the average length of protein sequences and h_i is the optimal-step of feature vector. In the second stage, fuzzy integral similarity is calculated between two feature vectors. Therefore, time complexity of stage (2) is $O(m2^m)$. In the third stage, we integrated both the stages for generating distance matrix. Here, we used parallel computation for reducing the time complexity. Therefore, total time complexity for generating distance matrix is:

$$\begin{aligned} &= \text{time complexity of stage 1} + ((n(n-1))/2t) * \text{time complexity of stage 2} + \varepsilon_t \\ &= O(nl + m^3\textstyle\sum_{i=1}^{n} h_i) + ((n(n-1))/2t) * O(m2^m) + \varepsilon_t \\ &= O(nl + m^3\textstyle\sum_{i=1}^{n} h_i) + O((n^2m2^m)/t) + \varepsilon_t \\ &= O(m^3\textstyle\sum_{i=1}^{n} h_i + nl + (n^2m2^m)/t) + \varepsilon_t, \end{aligned}$$

where t is the number of threads and ε_t is the extra time taken in job assigning to all t threads. We also calculated the computational speed of our method and ClustalW method on tested datasets, which is given below in conclusion section.

Results

To test our developed algorithm, we applied it to six sets of benchmark data. Different model might result different phylogenetic tree, therefore it is important to choose the most appropriate method. Here, we used Fitch-Margoliash or UPGMA (UPGMA = Unweighted Pair Group Method with Arithmetic Mean) approaches in PHYLIP package[24] for generating the phylogenetic tree. On the benchmark data, result generated using both the approaches has minor differences between them. However, we chose optimal tree based on taxonomic classification and compare with existing tools. The six benchmark datasets used in this study are as follows:

(i) NADH Dehydrogenase 5 (ND 5) protein sequences.
(ii) NADH Dehydrogenase 6 (ND 6) protein sequences.
(iii) xylanases protein sequences in the F10 and G11 datasets.
(iv) transferrin protein sequences.
(v) coronavirus spike protein sequences.
(vi) beta-globin protein sequences.

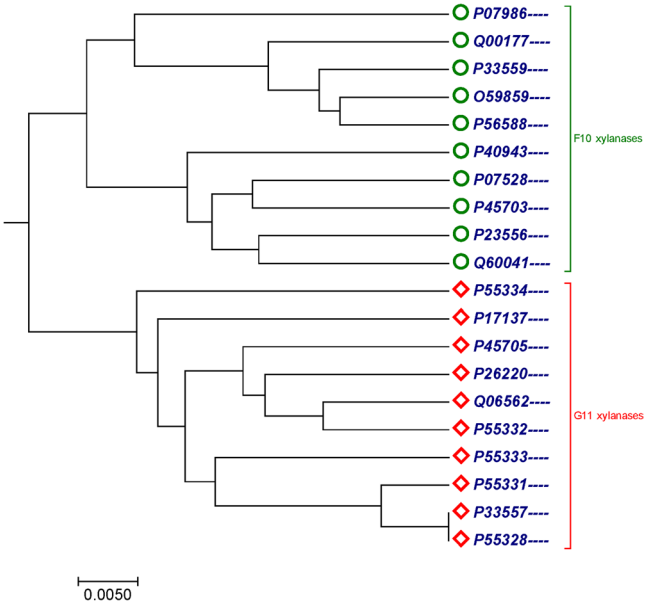

Figure 3. The Phylogenetic tree for 20 sequences of xylanases protein in the F10 and G11 datasets constructed by our method using Fitch-Margoliash approach.

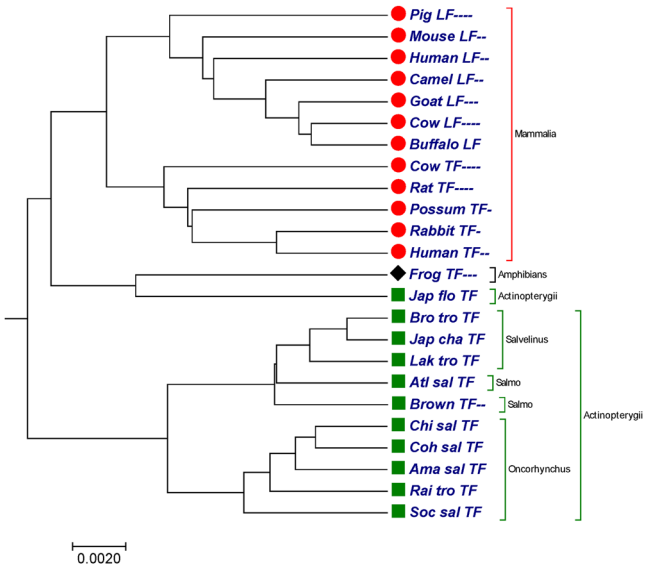

Figure 4. The phylogenetic tree for 24 sequences of transferrin protein constructed by our method using Fitch-Margoliash approach.

NADH Dehydrogenase 5 (ND 5) protein sequences. The proposed algorithm was tested on the benchmark dataset of 9 protein sequences of NADH Dehydrogenase 5 with nearly 600 amino acids (Table S1). All the sequences was obtained from the NCBI genome database. The MT-ND5 gene provides instructions for making a protein called NADH dehydrogenase 5. This protein is a part of a large enzyme complex known as complex I, which is active in mitochondria. Mitochondrially encoded NADH dehydrogenase 5 (complex I) in eukaryotes recognize as highly conserved subunit composition[33]. Therefore ND5 has been widely used for the analysis of the phylogenetic studies and their evolution. The phylogenetic tree generated by our method shown in Fig. 1, successfully grouped similar category based on taxonomic family classification. 9 sequences of ND5 protein belonged to mammals can be divided into following four categories based on their family; (i) *Hominidae* includes human, pigmy chimpanzee, common chimpanzee and gorilla; (ii) *Balaenopteridae* includes fin whale and blue whale; (iii) *Muridae* includes mouse and rat; and (iv) *Didelphidae* include opossum. From Fig. 1, it is clear that our method successfully clustered protein sequences separately based on their families. To illustrate the effectiveness of our method, we compared the phylogenetic tree generated by our approach with the phylogenetic tree generated by ClustalW using MEGA package[25] (Fig. S1) and phylogenetic trees generated by the previous studies[13,34–38] on the

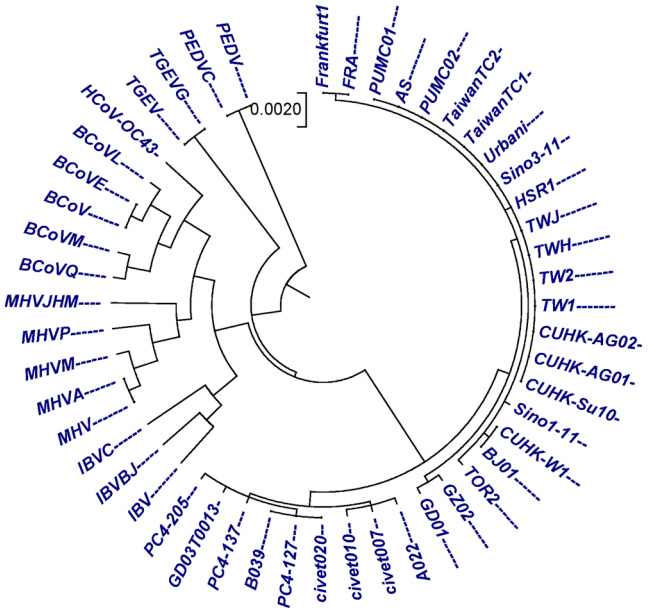

Figure 5. The phylogenetic tree for 50 sequences of coronavirus spike protein constructed by our method using Fitch-Margoliash approach.

same dataset. Figure 1 generated by our method did not clustered common chimpanzee and pigmy chimpanzee together as compared to Fig. S1. However, tree generated by our approach (Fig. 1) has advantage over[37,38]. In[37], phylogenetic trees construction based on the 20-D amino acid position ratio vector method and based on the 20-D amino acid content ratio vector method, four categories based on their family; *Hominidae*, *Balaenopteridae*, *Muridae* and *Didelphidae* are not separately clustred. Similarly in[38] and[37], phylogenetic trees construction based on the 20-D moment of inertia method and based on the 40-D amino acid position ratio and content ratio vector method, opossum is not separated as an outgroup.

We used correlation coefficient (CC) and Robinson-Foulds distance (RF-distance)[26] as a statistical tools for comparative analysis between two phylogenetic trees. As a general perception more CC means higher similarity between an inferred tree and a reference tree. Similarly, we often use the RF-distance[39,40] for analyzing topological similarity between two trees. RF-distance = 0 indicates that the test-tree topology is completely similar to that of the reference tree, while similarity level decreases as the RF-distance value increases. We obtained or calculated the CC and RF- distance of different alignment-free methods (Table 1) against the reference tree (ClustalW method). We used R-package for both CC and RF-distance calculation. In the Table 1, Jayanta *et al.*[34] (with grouping) method shows that, even the CC is very high (0.9403) as compared to our method CC (0.7378) but their corresponding the RF-distance is 4, which is higher than our method RF-distance which is 2 (i.e., tree from[34] (with grouping) is topologically less similar as compared to our tree to the reference tree). Similarly in Table 1, Wen *et al.*[35] and Yao *et al.*[36] having CC 0.7324 and 0.6908, respectively, which is nearer to CC of our method (CC = 0.7378). However, in terms of topological similarity, the RF- distance of Wen *et al.*[35] and Yao *et al.*[36] are 4 which is higher than RF-distance of our method. The above analysis shows that higher or closer CC does not always implies that the two phylogenetic trees are more similar or closer to each other.

NADH Dehydrogenase 6 (ND 6) protein sequences. The other benchmark dataset used in this study was 8 protein sequences of NADH Dehydrogenase 6 with nearly 175 amino acids (Table S2). All the sequences were obtained from the NCBI genome database. NADH-ubiquinone oxidoreductase chain 6 is a protein that in human is encoded by the mitochondrial NADH Dehydrogenase 6 gene. The ND6 protein is a subunit of NADH dehydrogenase (ubiquinone), which is found in the mitochondrial inner membrane and is the biggest of the five complexes of the electron transport chain[41]. 8 sequences of ND6 protein belong to mammals can be divided into following four categories based on their taxonomic family; (i) *Hominidae* includes human, common chimpanzee and gorilla; (ii) *Phocidae* includes harbor seal and gray seal; (iii) *Muridae* includes mouse and rat; and (iv) *Macropodidae* include wallaroo. As shown in the tree generated by our method (Fig. 2), the protein sequences belong to the families *Hominidae*, *Muridae* and *Phocidae* were correctly separated. Based on the taxonomic family classification, we compared our tree with the trees genetated in the previous studies[38,42] and tree generated by the ClustalW using MEGA package[25] (Fig. S2). The tree generated by our method has an advantage over[38], because it did not cluster (harbor seal, gray seal) and (mouse, rat) in separate clades. However, Fig. 2 shows consistency with[42] and Fig. S2 based on taxonomic family division.

We calculated CC and RF-distance from previous studies[38,42] with ClustalW. CC and RF-distance were also calculated between the our method and with ClustalW. In Table 2, Czerniecka *et al.*[38] method has lower CC (0.4609) than CC (0.5982) generated by our method compared with ClustalW method, and their corresponding RF-distance (RF = 6) is much higher than our method (RF = 2). Therefore, phylogenetic tree generated by our

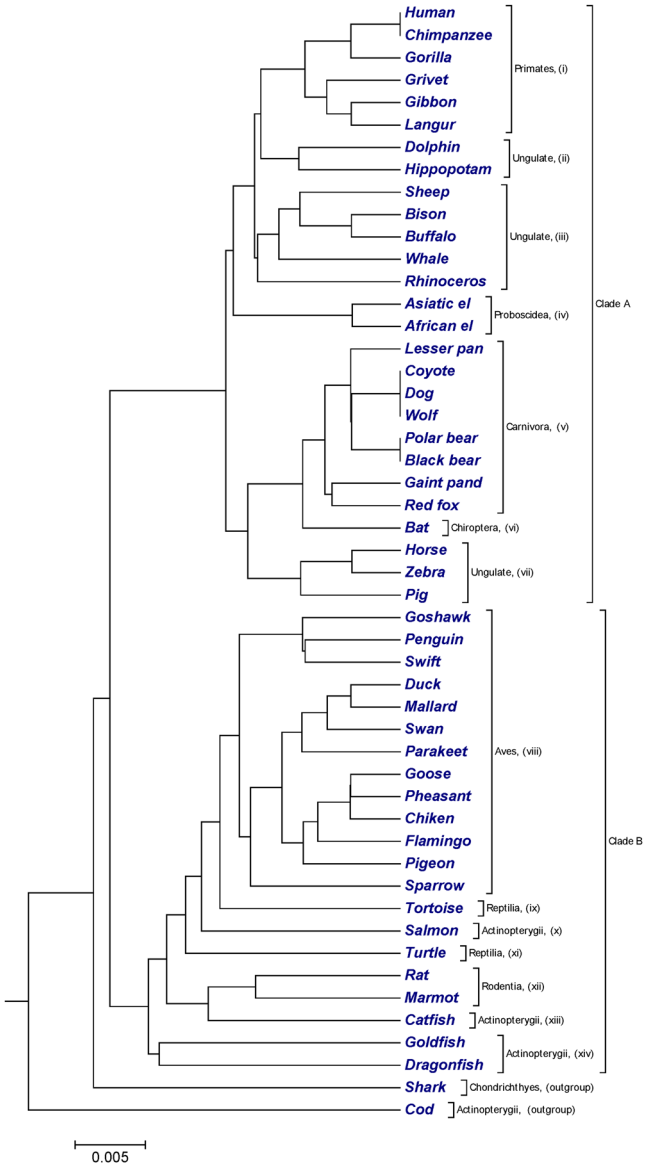

Figure 6. The phylogenetic tree for 50 sequences of beta-globin protein constructed by our method using UPGMA approach.

method (Fig. 2) is more topologically similar than phylogenetic tree generated by Czerniecka *et al.*[38] compared to reference tree (Fig. S2). However in Table 1, Gupta *et al.*[42] method has higher CC (0.7763) as compared to our method CC (0.5982) but both the methods has the same RF-distance = 2.

Xylanases protein sequences in the F10 and G11 datasets. The other benchmark dataset used for validation of the method was the 20 xylanases protein sequences in the F10 and G11 protein datasets with nearly 500 amino acids collected from[37]. Phylogenetic tree generated by our method (Fig. 3) accurately separated protein sequences belonging to G11 xylanases (red diomand) dataset from protein sequences belonging to F10 xylanases (green circle) dataset in separate branches. The phylogenetic tree generated in[37] did not separated protein sequences belonging to family F10 and G11 in two separate branches. Figure 3 showed that there is an improvement in building phylogenetic tree with our method than the method used in study[37]. However, our tree (Fig. 3) is consistent with the tree generated by ClustalW using MEGA package[25] (Fig. S3). We also calculated the CC and RF-distance between our method and ClustalW, which are 0.6998149 and 18.

Transferrin protein sequences. In this study the other benchmark dataset used was 24 protein sequence of transferrins (TFs) from vertebrates[43] with nearly 700 amino acids (Table S3). All the sequences were obtained from the NCBI genome database. Transferrins are the iron-binding proteins that are involved in iron storage and resistance to bacterial disease. Transferrins have high binding affinities for iron and keep the free iron in low concentration in blood and other bodily fluids[44]. The phylogenetic trees constructed by our method (Fig. 4),

Datasets	ND 5	ND 6	xylanases	transferrin	Coronavirus	beta-globin
Amino acids (approximate lengths)	600	175	500	700	1500	150
Number of sequences	9	8	20	24	50	50
Our method (execution time)	1 s	1 s	3 s	4 s	16 s	15 s

Table 3. Running time of our method.

successfully clustered transferrin protein sequences and lactoferrin protein sequences in separate clades. The tree generated by our approach (Fig. 4) divided the 24 sequences of transferrins (TFs) from vertebrates into three groups: *mammalia*(red circle), *actinopterygii*(green square) and *amphibians*(black diomand). Only Japanese flounder transferrin sequence belong to *actinopterygii* class was clustered with Frog transferrin sequence belong to *amphibians* class. In Fig. 4, sequences belong to genera *oncorhynchus* and *salvelinus* were clustered in separate clades, and sequences belong to genus *salmo* were placed close to each other.

Based on taxonomic division, comparison between alignment-free methods, the phylogenetic tree generated by our approach (Fig. 4) with phylogenetic tree generated in the previous studies[45,46] indicates improvement in our approach. In Fig. 4, sequences belong to *mammalia* class were clustered in a separate clade which were not observed in[45,46]. Moreover, species belong to genera *oncorhynchus* and *salvelinus* were grouped into separate clades, which is lacking in[46]. While comparing our tree (Fig. 4) with the benchmark tree constructed by[43] and tree constructed by ClustalW using MEGA package[25] (Fig. S4), we noticed that they are consistent among each other. The calculated the CC and RF-distance between our method and ClustalW are 0.7453224 and 20.

Coronavirus spike protein sequences. The other benchmark dataset used for the validation of our method was the 50 coronavirus spike proteins (Table S4) with nearly 1500 amino acids. Coronaviruses are diverse group of large, enveloped, positive-stranded RNA viruses belonging to the family Coronaviridae. Coronaviruses are responsible for respiratory and enteric diseases in human and other animals. According to the host type, Coronaviruses can be divided into four groups (Table S4). Group I and II contains mammalian coronaviruses, group III contain avian coronaviruses and group IV contain SARS-CoVs[47–49]. The spike protein which is common to all known coronaviruses, is crucial for viral attachment and entry into the host cell. To illustrate the use of the quantitative characterization of these sequences, we employed our method to analyse the 50 coronavirus spike proteins. Observing Fig. 5, we found that SARS-CoVs (group IV) appear to cluster together and formed a separate branch, which can be easily distinguishable from other three groups(I, II and III) of coronaviruses. Similarly, sequences belonging to groups II and III are placed at an independent branch. While sequences belong to group I, such as (TGEV, TGEVG) and (PEDVC, PEDV) formed separate clades, but they were close to each other. A closer look at the subtree of SARS-CoVs (group IVa) belonged to 03–04 interspecies epidemic are cluster together, while all the human SARS-CoVs formed another branch. Phylogenetic tree generated by our method (Fig. 5) is consistent with phylogenetic trees generated in the previous studies[42,50,51] and alignment based method ClustalW using MEGA package[25] (Fig. S5). The CC and RF-distance between our method and the ClustalW are 0.9555357 and 46.

Beta-globin protein sequences. 50 sequences of beta-globin protein (Table S5) of different species[52] with nearly 150 amino acids were extracted from GenBank. Based on the type of host, 50 sequences of beta-globin protein can be classified into following groups such as primates, proboscidea, ungulate, carnivora, rodentia, chiroptera, aves, actinoptergii, reptilia and chondrichthyes. The phylogenetic trees constructed by our method (Fig. 6) separated 50 sequences of beta-globin protein into two major clades: clade A and clade B. Clade A contained mammalian beta-globins and clade B contained beta-globins from avian, fish, and reptilian species. According to the taxonomy division, we categorized two major clades into several sub-clades. All primates, proboscidea, carnivora, chiroptera, aves and rodentia were successfully cluster into clades (i), (iv), (v), (vi), (viii) and (xii) respectively. Ungulate were clustered into clades (ii), (iii) and (vii). We observed an obvious limitation in Fig. 6 is that, our approach failed to cluster fish species into single clades based on taxonomy. However, the phylogenetic tree generated by our approach is consistent and generated a better result based on taxonomic characteristic of species while compared with previous studies[45,53]. Phylogenetic tree generated by ClustalW using MEGA package[25] (Fig. S6), successfully clustered fish species and reptilian species in separate clades, while our approach (Fig. 6) failed to cluster separately. However, from both figures, it is clear that phylogenetic tree generated by our method (Fig. 6) depicted more clear division in terms of branch length than phylogenetic tree generated by ClustalW (Fig. S6). The calculated CC and RF-distance between our method and ClustalW are 0.7294663 and 64.

Conclusion

This study focused on fuzzy integral similarity method based on Markov chain and applied this algorithm to protein sequence analysis. Sequence comparison is the fundamental and most frequent activity in bioinformatics. In sequence alignment method, two sequences are assigned an alignment score based on insertion, deletion and substitution of nucleotides or amino acids. However, sometimes alignment becomes misleading due to unequal length of sequences, gene rearrangements, inversion, transposition and translocation at substring level. In these scenarios, alignment-free methods are therefore a better alternative as it reduces the technical constraints of alignments. We have constructed transition probability matrix using Markov chain of each protein sequence. Subsequently, a fuzzy integral similarity method was used to assign similarity score belong to closed interval [0, 1] between two protein sequences. The benefit of our approach is that, it do not require any prior biological knowledge regarding homologous relationship (common ancestry) among the sequences which makes it fully automated and robust. We implemented our method on six benchmark datasets as discussed in the result section.

In Figs 1 and 2, our method successfully grouped NADH Dehydrogenase 5 and NADH Dehydrogenase 6 protein sequences into four categories based on the taxonomic family classification. However, in Fig. 1, common chimpanzee is closer to human than pigmy chimpanzee, which is contrast to the known fact of evolution. In xylanases protein sequences, tree generated by our approach (Fig. 3) correctly distinguished 20 sequences of xylanases protein belong to families G11 and F10 in separate clades. Similarly in Fig. 4, it is clear that, our method separated the transferrin protein sequences and the lactoferrin protein sequences into separate clades, which is desirable. A satisfactory improvement can be seen in the phylogenetic tree built by our algorithm at genus level (Fig. 4). Our tree (Fig. 4) successfully separated sequences belong to genera *oncorhynchus* and *salvelinus* in separate branches, and sequences belong to genus *salmo* were closest to each other. In coronavirus spike protein, phylogenetic tree generated by our approach (Fig. 5) nicely categorized four groups based on their host types (groups I, II, III and IV). Moreover, our method successfully categorized SARS-CoVs which belong to group IV into two subgroups, which corresponds to the 03–04 interspecies epidemic and human epidemic, respectively. Finally, we implemented our method on 50 sequences of beta-globin protein. An obvious default in Fig. 6 generated by our method is that our approach failed to cluster fish species into a single clade. However, we found consistency while comparing our tree (Fig. 6) with recently developed alignment-free method collected from[45,53].

Our programs were executed on a linux server with 24 dual core processor with 384 GB RAM. We enriched our programs by incorporating parallel computation, which can reduce the execution time of our program by increasing the number of threads, depending on the number of sequences. In our program, we implemented two threads as a default parameter. However, the user can manipulate the parameter to single thread or multiple threads. The execution time of our method with two threads is shown in Table 3. In the Table 3, the execution time of our method for 50 sequences of coronavirus spike protein is 16 seconds by using two threads, which can be reduced to 7 seconds and 5 seconds by using threads four and six, respectively. In this study, we implemented statistical tools such as CC, RF-distance and ROC[27–29] curve to compare the result generated by our method with the other alignment-free methods. We performed comparative study between the RF-distance and the CC for each method for the ND5 and ND6 datasets. Similarly, we plotted ROC curve and calculate area under the ROC curve (AUC) for distance matrices generated by our method and other alignment-free tools from Alfree repository[54]. The results of ROC and AUC analysis for all benchmark datasets are given in supplementary material. We are yet to attain an highly efficient alignment-free method for phylogenetic analysis. However, our method shows an improvement over the other existing alignment-free methods in terms of sequence clustering. Based on the observed progress, this method would be useful for the researcher to develop hypothesis that can be examined further in details. Before continuing our research work for further improvement, we would like to emphasize that this is a probabilistic approach in nature. It can later be modified by including more biological evidence. Overall, our goal in this study was to bring a new methodology or algorithm to the proteomics study. This proposed algorithm can be used to guide the development of more powerful measures for sequence analysis.

Data Availability

We wrote code in C-programming which is available via our institute website.

References

1. Liu, N. & Wang, T. Protein-based phylogenetic analysis by using hydropathy profile of amino acids. *FEBS Lett.* **580**, 5321–5327 (2006).
2. Xu, Q. *et al*. Statistical analysis of interface similarity in crystals of homologous proteins. *J. Mol. Biol.* **381**, 487–507 (2008).
3. Gotoh, O. An improved algorithm for matching biological sequences. *J. Mol. Biol.* **162**, 705–708 (1982).
4. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453 (1970).
5. Liu, X., Yang, X., Wang, C., Yao, Y. & Dai, Q. Number of distinct sequence alignments with k-match and match sections. Comput. *Biol. Medicine* **63**, 287–292 (2015).
6. Vinga, S. & Almeida, J. Alignment-free sequence comparison—a review. *Bioinforma.* **19**, 513–523 (2003).
7. Elloumi, M. Comparison of strings belonging to the same family. *Inf. Sci.* **111**, 49–63 (1998).
8. Pham, T. D. & Zuegg, J. A probabilistic measure for alignment-free sequence comparison. *Bioinforma.* **20**, 3455–3461 (2004).
9. Song, K. *et al*. New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing. *Briefings Bioinforma.* **15**, 343–353 (2014).
10. Kantorovitz, M. R., Robinson, G. E. & Sinha, S. A statistical method for alignment-free comparison of regulatory sequences. *Bioinforma.* **23**, i249–i255 (2007).
11. Zhang, Y. & Chen, W. A new measure for similarity searching in dna sequences. *Match Commun. Math. Comput. Chem.* **65**, 477–488 (2011).
12. Hide, W., Burke, J. & Vision, D. B. D. Biological evaluation of d2, an algorithm for high-performance sequence comparison. *J. Comput. Biol.* **1**, 199–215 (2009).
13. Ma, T., Liu, Y., Dai, Q., Yao, Y. & He, P. A graphical representation of protein based on a novel iterated function system. *Phys. A: Stat. Mech. its Appl.* **403**, 21–28 (2014).
14. Hamori, E. & Ruskin, J. H curves, a novel method of representation of nucleotide series especially suited for long dna sequences. *J. Biol. Chem.* **258**, 1318–1327 (1983).
15. El-Lakkani, A. & El-Sherif, S. Similarity analysis of protein sequences based on 2d and 3d amino acid adjacency matrices. *Chem. Phys. Lett.* **590**, 192–195 (2013).
16. Mu, Z., Wu, J. & Zhang, Y. A novel method for similarity/dissimilarity analysis of protein sequences. *Phys. A: Stat. Mech. its Appl.* **392**, 6361–6366 (2013).
17. Yu, C., He, R. L. & Yau, S. S.-T. Protein sequence comparison based on k-string dictionary. *Gene* **529**, 250–256 (2013).
18. Ralescu, D. & Adams, G. The fuzzy integral. *J. Math. Analysis Appl.* **75**, 562–570 (1980).
19. Torra, V. & Narukawa, Y. The interpretation of fuzzy integrals and their application to fuzzy systems. Int. J. Approx. *Reason.* **41**, 43–58 (2006).

20. Medhi, J. *Stochastic Processes*. (New Age Science, 2009).
21. Garcia, F., Lopez, F. J., Cano, C. & Blanco, A. Fisim: A new similarity measure between transcription factor binding sites based on the fuzzy integral. *BMC Bioinforma.* **10**, 224 (2009).
22. Zhang, S., Zhang, Y. & Gutman, I. Analysis of dna sequences based on the fuzzy integral. *MATCH Commun. Math. Comput. Chem.* **70**, 417–430 (2013).
23. Sims, J. R. & Zhenyuan, W. Fuzzy measures and fuzzy integrals: An overview. *Int. J. Gen. Syst.* **17**, 157–189 (1990).
24. Felsenstein, J. Phylip–phylogeny inference package(version 3.2). *Cladistics* **5**, 164–166 (1989).
25. Kumar, S., Stecher, G. & Tamura, K. Mega7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* **33**, 1870–1874 (2016).
26. Robinson, D. & Foulds, L. Comparison of phylogenetic trees. *Math. Biosci.* **53**, 131–147 (1981).
27. Swets, J. Measuring the accuracy of diagnostic systems. *Sci.* **240**, 1285–1293 (1988).
28. Nemes, S. & Hartel, T. Summary measures for binary classification systems in animal ecology. *North-Western J. Zool.* **6**, 323–330 (2010).
29. Sonego, P., Kocsor, A. & Pongor, S. Roc analysis: applications to the classification of biological sequences and 3d structures. *Briefings Bioinforma.* **9**, 198–209 (2008).
30. Sugeno, M. *Fuzzy Measures and Fuzzy Integrals—a Survey*, 251–257 (Morgan Kaufmann, 1993).
31. Chaira, T. *Fuzzy Measures in Image Processing*, 587–606 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2008).
32. Devi, S. G., Selvam, K. & Rajagopalan, S. P. An abstract to calculate big o factors of time and space complexity of machine code. In *International Conference on Sustainable Energy and Intelligent Systems (SEISCON 2011)*, 844–847 (2011).
33. Cardol, P. Mitochondrial nadh:ubiquinone oxidoreductase (complex i) in eukaryotes: A highly conserved subunit composition highlighted by mining of protein databases. *Biochimica et Biophys. Acta (BBA) - Bioenerg.* **1807**, 1390–1397 (2011).
34. Das, J. K., Choudhury, P. P., Chaturvedi, N., Tayyab, M. & Hassan, S. S. Ranking and clustering of drosophila olfactory receptors using mathematical morphology. *Genomics*, https://doi.org/10.1016/j.ygeno.2018.03.010 (2018).
35. Wen, J. & Zhang, Y. A 2d graphical representation of protein sequence and its numerical characterization. *Chem. Phys. Lett.* **476**, 281–286 (2009).
36. Yao, Y.-H. *et al.* Analysis of similarity/dissimilarity of protein sequences. *Proteins: Struct. Funct. Bioinforma.* **73**, 864–871 (2008).
37. Li, Y., Song, T., Yang, J., Zhang, Y. & Yang, J. An alignment-free algorithm in comparing the similarity of protein sequences based on pseudo-markov transition probabilities among amino acids. *Plos One* **11**, 1–14 (2016).
38. Czerniecka, A., Bielinska-Waz, D., Waz, P. & Clark, T. 20d-dynamic representation of protein sequences. *Genomics* **107**, 16–23 (2016).
39. Leimeister, C.-A., Sohrabi-Jahromi, S. & Morgenstern, B. Fast and accurate phylogeny reconstruction using filtered spaced-word matches. *Bioinforma.* **33**, 971–979 (2017).
40. Leimeister, C.-A. & Morgenstern, B. kmacs: the k -mismatch average common substring approach to alignment-free sequence comparison. *Bioinforma.* **30**, 2000–2008 (2014).
41. Donald Voet, J. G. V. & Pratt, C. W. *Fundamentals of Biochemistry: Life at the Molecular Level*, 5th edition. (Wiley, 2016).
42. Gupta, M., Niyogi, R. & Misra, M. An alignment-free method to find similarity among protein sequences via the general form of chou's pseudo amino acid composition. *SAR QSAR Environ. Res.* **24**, 597–609 (2013).
43. Ford, M. J. Molecular evolution of transferrin: Evidence for positive selection in salmonids. *Mol. Biol. Evol.* **18**, 639–647 (2001).
44. Loehr, T. M. *Iron Carriers and Iron Proteins*. (VCH, New york, 1989).
45. Yu, L., Zhang, Y., Gutman, I., Shi, Y. & Dehmer, M. Protein sequence comparison based on physicochemical properties and the position-feature energy matrix. *Sci. Reports* **7** (2017).
46. Wu, H., Zhang, Y., Chen, W. & Mu, Z. Comparative analysis of protein primary sequences with graph energy. *Phys. A: Stat. Mech. its Appl.* **437**, 249–262 (2015).
47. Gao, L., Qi, J., Wei, H., Sun, Y. & Hao, B. Molecular phylogeny of coronaviruses including human sars-cov. *Chin. Sci. Bull.* **48**, 1170–1174 (2003).
48. Gorbalenya, A. E., Snijder, E. J. & Spaan, W. J. M. Severe acute respiratory syndrome coronavirus phylogeny: toward consensus. *J. Virol.* **78**, 7863–7866 (2004).
49. Ksiazek, T. G. *et al.* A novel coronavirus associated with severe acute respiratory syndrome. *New Engl. J. Medicine* **348**, 1953–1966 (2003).
50. Li, C., Xing, L. & Wang, X. 2-d graphical representation of protein sequences and its application to coronavirus phylogeny. *BMB Rep.* **41**, 217–222 (2008).
51. Hou, W., Pan, Q., Peng, Q. & He, M. A new method to analyze protein sequence similarity using dynamic time warping. *Genomics* **109**, 123–130 (2017).
52. Yau, S.-T., Yu, C. & He, R. A protein map and its application. *DNA Cell Biol.* **27**, 241–250 (2008).
53. Xu, C., Sun, D., Liu, S. & Zhang, Y. Protein sequence analysis by incorporating modified chaos game and physicochemical properties into chou's general pseudo amino acid composition. *J. Theor. Biol.* **406**, 105–115 (2016).
54. Zielezinski, A., Vinga, S., Almeida, J. & Karlowski, W. M. Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biol.* **18**, 186 (2017).

Acknowledgements

The authors thank Ilya Ioshikhes, Per Stenberg and Andrew Lynn for critical reviewing and helpful comment to improve the manuscript. Ramalingaswami Fellowship from Department of Biotechnology, Ministry of Science and Technology, Government of India, supported S.N. (BT/RLF/Re-entry/48/2013).

Author Contributions

A.S. and B.T. developed the idea, A.S. and S.N. written code in programming language and analyzed the results, and A.S. wrote the manuscript text. B.T. and S.N. guided the study. All authors reviewed the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-39477-8.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.