Investigation of protein sequence similarity based on physio-chemical properties of amino acids

Ranjeet Rout[1], Hari Pandey[2], Sanchit Sindhwani[3], Saiyed Umer[4], and Smitarani Pati[3]

[1]National Institute of Technology Srinagar
[2]Edge Hill University
[3]National Institute of Technology Jalandhar
[4]Aliah University

May 6, 2020

Abstract

Comparison of protein sequence similarity is a significant study. By virtue of this method, we can expose the evolutionary relationship among protein sequences. So, it is required to design effective computational algorithms that can compare the similarities among the colossal amount of sequences. The aim of this research is to develop efficient tools in the field of protein sequences comparison and phylogenetic study. The proposed method performs a feature generation process based on the physio-chemical properties of amino acids that best describes the revolutionary relationship among the species in a protein family. The protein sequences are transferred into an Eighty dimensional feature vector among the group of amino acids. Finally, four different datasets were used to validate the accuracy of the proposal and a correlation coefficient of 0.94417 of ND5 dataset using ClustalW has been found. This is much higher than some of the methods. At last the result explains the effectiveness in the similarity analysis among genome sequences.

Investigation of protein sequence similarity based on physio-chemical properties of amino acids

[1]Ranjeet Kumar Rout, [2]Hari Mohan Pandey*, [3]Sanchit Sindhwani, [4]Saiyed Umer, [5]Smitarani Pati

[1]Computer Science & Engineering, National Institute of Technology Srinagar,

Hazratbal-190006 J&K, India.

[2]Department of Computer Science, Edge Hill University, Ormskirk, Lancashire, UK.

[3]Computer Science and Engineering, National Institute of Technology

Jalandhar-144011,Punjab, India.

[4]Computer Science & Engineering, Aliah University, West Bengal, India.

[5]Instrumentation & Control Engineering, National Institute of Technology Jalandhar,

Punjab-144011, India.

Short Running Title : Protein Sequence Similarity based on amino acids

***Correspondence:**

Hari Mohan Pandey

Department of Computer Science, Edge Hill University, Ormskirk, Lancashire, UK.

Pandeyh@edgehill.ac.uk

Contact: +447414981569

Investigation of protein sequence similarity based on physio-chemical properties of amino acids

Abstract: Comparison of protein sequence similarity is a significant study. By virtue of this method, we can expose the evolutionary relationship among protein sequences. So, it is required to design effective computational algorithms that can compare the similarities among the colossal amount of sequences. The aim of this research is to develop efficient tools in the field of protein sequences comparison and phylogenetic study. The proposed method performs a feature generation process based on the physio-chemical properties of amino acids that best describes the revolutionary relationship among the species in a protein family. The protein sequences are transferred into an Eighty dimensional feature vector among the group of amino acids. Finally, four different datasets were used to validate the accuracy of the proposal and a correlation coefficient of 0.94417 of ND5 dataset using ClustalW has been found. This is much higher than some of the methods. At last the result explains the effectiveness in the similarity analysis among genome sequences.

Keywords: Sequence similarity, amino acids, Physio-chemical property, Markov Chain transition matrix.

Introduction

Research in the field of Computational Biology has seen significant growth in the last decades. This has derived rich data concerning protein sequence, structures, and gene expressions and aided in prediction and analysis of DNA. With this massive generation of protein sequences it is prudent to develop efficient tools for research in the field of Phylogenetic study. It is required to design effective computational algorithms that can compare the similarities among the colossal amount of sequences. Studies show that few protein sequences do not possess noteworthy sequence alignment similarities and are a great impediment to sequence comparisons and analysis [14]. This may be because of the unequal sequence lengths, inversion, transposition and translocation at sub-string level [45]. Thus, applying alignment-free methods will be a more realizable and cost-effective approach as they concentrate more on feature vectors for identifying attributes. These methods are realized in two steps. First, fixedâ\euro“length feature vectors are derived from protein sequences and then in the second step these vectors are provided as input to the similarity comparison algorithms. There are various approaches for creating serviceable datasets like predicting transcriptional activity of multiple site p53 mutants [31], predicting drug-target interaction networks [30], HIV cleavage sites in proteins [15], body fluids [33], antimicrobial peptides [54], colorectal cancer related genes[39], S-nitrosylation modification sites, protein sub cellular locations [40], and many more that prove to be very useful in the Phylogenetic study.

The sequence alignment is a very convoluted process and there are numerous efficacious methods to reduce its complexities and provide reliable results. One great method is for sequence comparison is graphical representation [20] that delineates the protein sequences aided with mathematical descriptors that help in recognizing the similarities between them. Numerical characterization [44] of protein sequences expresses the crux about their amino acid compositions. Each sequence is mapped to a distance frequency matrix and a similarity score is computed applying any effective distance measuring tool. With the k-string dictionary [60] protein sequences can be rendered with comparatively lower dimensional frequency on low cost. This is then inattentive by implementing singular value decomposition that provides a more precise vector representation of the protein sequences with the help of a tree. Fuzzy integrals [49] methods for similarity comparison earmark similarity scores within close intervals [0,1] for two selected sequences. A protein sequence inheres 20 amino acids. Protein sequences can be delineated by employing transition probability matrix, fuzzy measures and fuzzy integrals. Distance matrix can be derived by identified fuzzy integral similarities and a phylogenetic tree can be constructed with the data. The Chou's pseudo amino acid composition [25] can also be utilized for alignment free similarity comparison. On the basis of the acquired proportion of amino acids, the distance between the foremost and every other amino acids, and the organization of the amino acids a 60-dimensional feature vector is derived. The phylogenetic tree is contrived out of this matrix. This proved to be economic in terms of space and time complexity as compared to other alignment free methods. Pseudo-Markov transition probabilities [43] among the 20 amino acids can also derive similarities among protein sequences. This method encodes the protein sequence into a$440-D$ (dimensional) feature vector. This vector is comprised of a$400-D$ Pseudo-Markov transition probability vector among the 20amino acids, a $20-D$ content ratio vector and a $20-D$ position ratio vector of amino acids in the sequence. The protein sequences are compared by calculating the Euclidean distance between these vectors. Deploying Markov Chain parameters [52] is also a very effective method for similarity comparison. Markov chain parameters are evaluated based on the frequencies of occurrence of all the realizable pairs of amino acids for every alignment free gene sequences. These features can derive the similarity between two gene sequences utilizing a fuzzy integral algorithm. This algorithm has an advantage of more appropriate clustering performance for gene sequence comparison. $H-$curve [26] is very effective in analysis of local as well as global features of long protein sequences. The information derived from the nucleotide sequences is mapped from four letter language into a 3D space function called the H-curve. This curve also provides the integral information about the DNA. Alternating word frequency and normalized Lempel-Ziv complexity [65] also promise a cost-effective sequence similarity analysis. We can also approximate the protein sequence alignments to analyze the similarities and the max. segment pair score[1]. This method is extremely constructive and robust for analysis of long protein sequence databases, motif searches and gene identification searches. Protein fold prediction methods [56] use classifiers for deriving robust successive evolution knowledge from$\mathrm{PSI}-\mathrm{BLAST}$ and $\mathrm{PSI}-\mathrm{PRED}$ profiles and give a comprehensive feature set. This information is crucial for protein function analysis and structure prediction. A random forest (RF) classifier [55] pertained to a feature set can be used for both sequence and structure prediction using three large datasets. Positionâ\euro“based features [16] are also considered an efficient and reliable method for describing the distribution of amino acids. Studies show that evolution information provides an efficient way in protein sequence analysis [19, 38, 58, 64]. Methods of phylogenetic analysis contribute to function prediction and can be used in identifying similarities between life forms by analyzing their medicinal qualities [35].

Methods of alignment-free sequence analysis are very effective in phylogenetic classification of protein sequences, horizontal gene transfer recognition and discovering recombined sequences. These methods are also economic in terms of computations as they are usually of linear complexity and influenced by length of query sequence [8]. As compared to alignment based methods, these methods are not subject to presumptions of evolutionary trajectories of sequence changes. These methods are mathematically justifiable through linear algebra and information theory. Most of the methods can easily be applied using standard tree-building software [22, 37]. The alignment-free algorithms are expanding their applications in phylogenomics and horizontal gene transfer [6], population genetics [27] and relations between genome and epigenome [48]. These methods have evolved and improved their performance in the last decades [7] still there are confrontations for the number of effective bench marking approaches for alignment free similarity analysis [68]. The sample data-sets available [18] is outpacing the storage and processing capacities of the computers used today for research. Alignment free methods proliferate over primary next-generation sequencing applications[5, 53, 67, 50] and can efficiently derive biological data from next-generation raw data.

These alignment-free techniques have also some limitations such as it is difficult to classify with the concerned properties of a protein sequence to a specific cluster. But still, they are much better for pattern recognition with the known protein clusters. They have much better potential as when compared to alignment-based techniques which are applicable to a number of applications in bio-informatics. In this paper these problems have been tried to resolve based on by introducing three different quantitative methods and the physiochemical properties of amino acids which leads the desired results. Based on theses methods a feature vector consist of 80 features have been generated for different species. Here, Euclidean distance has been used to measure the distance between two feature vector P and S. From which some proximity results have been observed among the species and reported in more precise fashion. As a final point, our proposed technique is more precise than a few offered techniques for comparison analysis on the ND5 and ND6 dataset in intricate level, and phylogenetic tree obtained using this method are find accurate on the F10 and G11 dataset.

In brief, the contributions of this research work are summarized as follows:

1. Characterization of amino acids based on their chemical properties: From the large range of physio-chemical properties of amino acids, side chain effect renders an important role for formation of tertiary structure of proteins. According to the side chain effect of amino acids, these amino acids are classified into eight different groups. So, each primary protein sequences renders into another structure for further analysis.
2. *Procedure for obtaining feature vector based of Markov Chain transition matrix:* Identification of feature vectors based on the transition probability among the amino acid group is possible. A unique procedure is devised to generate a feature vector of 64features based the characteristic of physio-chemical properties of amino acids of ND5, ND6 protein families and ten species each from G10 and F11 protein families have been studied in this research.
3. Procedure for obtaining feature vector based of content ratio and distribution ratio of amino acid groups: Based on physiochemical properties of amino acid groups an eight dimension content ratio vector explaining the frequency of each group and another eight dimension vector describing the position distribution of each group have devised. Further, we study the content ratio and distribution ratio among the eight groups of amino acids of ND5, ND6, G10 and F11 protein families. In the following section these points are elaborated briefly.

The rest of this article is arranged as follows: In Section 2, definition of different fundamental parameters with the appropriate description of the employed method. The experimental results and discussions have been established with the usefulness of our proposed method in Section 3. Section 4 finish off this paper with highlighting the key factors of the intact analysis.

The intended Methods and Materials

In this section, three different novel methods have been proposed to analyze primary protein sequences in intricate level on the basis of Markov chain transition matrix, content ratio and distribution of chemical groups of amino acids. These methods which are discussed in this section, have been explained for the sake of clarity on primary gene sequence obtained from different gene family.

Characterization of amino acids based on physio-chemical properties

The structure of protein sequences majorly depends on the physio-chemical properties of amino acids. These properties provide information about the coding region of the gene sequence [32] and about the function of the gene coded by the region [2]. The methods of similarity analysis in DNA and protein sequence serve as effective tools in phylogenetics research [46]. The primary protein sequence is consist of twenty amino acids represented by characters as show in Table 1. They act an important role in the determination of three dimensional structure of proteins and hence the biological processes are depends upon the physio-chemical properties of amino acids [3, 10, 12, 51]. According to the side chain effect of amino acids listed in Table 1, these twenty amino acids can be classified into eight different groups as shown in Table 1 discussed in [17]. The distribution of eight types of amino acids describes protein primary structures. For better understanding the feature vector of protein primary structure the classification of amino acids is define in Equation 1.

$P(S(i)) = \{$

$$\begin{array}{cc} D & amp; if\ S(i)=\{D,E\} \\ R & amp; if\ S(i)=\{R,H,K\} \\ Y & amp; if\ S(i)=\{F,Y,W\} \\ A & amp; if\ S(i)=\{I,L,V,A,G\} \\ P & amp; if\ S(i)=\{P\} \\ C & amp; if\ S(i)=\{M,C\} \\ S & amp; if\ S(i)=\{S,T\} \\ N & amp; if\ S(i)=\{Q,N\} \\ & amp; \end{array} \qquad (1)$$

Where $S(i)$ represents i^{th} character in the given protein sequence and $P(S(i))$ will represent the corresponding substitute of amino acid $S(i)$. For example, for a given protein sequence$S(i)$ = AGMEQQTMPHERCSNPTTGHIRTF, the feature sequence of$S(i) = P(S(i) =$ AACDNNSCPRDRCSNPSSARARSY. Composition and distribution of amino acids are two key factors of Protein sequences. This has been used in different areas such as protein similarity, classification on the basis of structure[63, 66, 36] or the chemical composition, identification of patterns among protein sequences. There are mainly two ways of representing the Protein sequences. As per proposed methods [41], these are discrete and Sequential. Both have their own shortcomings in their ways of representation. The sequential way of representation fails when protein does not have much sequence similarity to known protein sequences. In other way the loss of ordering is the main drawback. Thus a much novel way for the same has been proposed i.e a 80-D vector involving both of the features has been taken into consideration.

Construction of 64-D vector using Markov chain transition matrix

Let $S = s_1, s_2, s_3, \ldots, s_n$ be a gene sequence of length$n$ characterized on amino acids $A = a_1, a_2, \ldots, a_8$ a set of 8 Alphabets representing each amino acid group as defined in equation 1. For $1 <= i <= 8$, a amino acid is said to appear at some position k in the protein sequence S, if $S(k) = A(i)$and for $1 <= j <= 8$, a pair of amino acid$A\,(i)\,A\,(j)$ is said to occur at adjacent position k, l in the protein sequence S, if$S\,(i)\,S\,(j) = A\,(i)\,A\,(j)$. Correspondingly, the 64-dimension vector is defined$(P_{11}, P_{12}, \ldots, P_{88})$. Here P is a(8×8)-matrix with elements$\{P_{i,j} : i, j = 1, 2, \ldots, 8\}$. A random process$(X_0, X_1, \ldots, X_8)$ with finite state space$S = \{s_0, s_1 \ldots, s_8\}$ is said to be a Markov chain transition matrix P [24]. If for all n, all$i, j \in 1, \ldots, 8$ and all $i_1, i_2, \ldots, i_8$, here$P(X_{n+1} = s_j | X_1 = s_{i_1}, X_1 = s_{i_1}, \ldots, X_n = s_{i_n}, X_n = s_{i_n})$ =$P(X_n = s_j | X_n = s_i) = P_{i,j}$.

The elements of matrix P are called transition probability. The element $P_{i,j}$ is the conditional probability of being in statess_j given that we are in state s_i, where$\{i, j\} \in \{1, \ldots, 8\}$ for eight group of amino acids is defined in equation 2.

$P_{\text{ij}} = \{$

$$\begin{array}{cc} \frac{n_{\text{ij}}}{n_i} & amp; \text{if}A_i \neq S_N \\ \frac{n_{\text{ij}}}{n_i - 1} & amp; \text{if}A_i = S_N \\ & amp; \end{array} \qquad (2)$$

Here, $P_{\text{ij}} = 0$ if $n_i = 0$ for some A_i or for some A_i which has frequency equal to one and appears at the end of the sequence.

$\sum_{j=1}^{8} n_{\text{ij}} = \{$

$$\begin{array}{cc} n_i & amp; \text{if}A_i \neq S_N \\ n_i - 1 & amp; \text{if}A_i = S_N \\ & amp; \end{array}$$

$\sum_{j=1}^{8} n_{\text{ij}} = \{$

$$\begin{array}{cc} n_i & amp;\text{if} A_i \neq S_i \\ n_i - 1 & amp;\text{if} A_i = S_j \\ & amp; \end{array}$$

In the context of amino acid sequence, Markove chain transition probability matrix can be expressed as:

$$P_{\text{ij}} =$$

$$\begin{bmatrix} P_{1,1}, P_{1,2}, \cdots, \cdots, P_{1,8} \\ P_{2,1}, P_{2,2}, \cdots, \cdots, P_{2,8} \\ P_{3,1}, P_{3,2}, \cdots, \cdots, P_{3,8} \\ \vdots, \vdots, \cdots, \cdots, \vdots \\ P_{8,1}, P_{8,2} \cdots\cdots, P_{8,8} \\ \end{bmatrix}$$

Construction of 8-D Content Ratio Vector based on the physiochemical properties of amino acids

The primary structure of a protein sequence is consist of 20 amino acids. Considering that a protein sequence is composed of 8 amino acids, as defined in equation 1,for each amino acid present in the given sequence content ratio C_i where $1 <= i <= 8$ is defined as in equation 3.

$C_i = \frac{c_i}{N} \quad wh$ere $\sum_{i=1}^{8} n_i = N(3)$

It is very much clear that this vector$C(C_1, C_2, C_3, \ldots, C_8)$ will sum equal to 1. This parametric quantity uniquely is not an adequate parameter for gene sequence comparison because gene sequences having the same number of amino acids placed at different positions are not quite similar in nature. Therefore, this add another 8 feature values in the 80-D feature vector.

Construction of 8-D Distribution Ratio Vector based on the physio-chemical properties of amino acids

The two parameters discussed above can not define any protein sequence uniquely. Thus a new parameter of distribution for each amino acid i, $1 <= i <= 8$, is introduced which will differentiate between the gene sequences even if they have same content ratio and probability of adjacent amino acids. This vector will contribute to third part of80-D vector differentiating the protein sequences. The distribution vector is defined as follows:

$\delta_i = \sum_{j=1}^{\alpha_i} \frac{(\tau_i - \eta_i)^2}{\alpha_i}, \quad wh$ere $\eta_i = \frac{\beta_i}{\alpha_i}$ and $\beta_i = \sum_{j=1}^{\alpha_i} \tau_j(4)$

Here, τ_j represents as the distance of j^{th} amino acid from the first position amino acid in the gene sequence. This signifies the distance between each amino acid of the sequence from the first position amino acid. The variable α_i and β_i is defined as the count and sum of the position of i^{th} amino acid in the protein sequence respectively. The variable β_i represents the distance between eight group of amino acids from the first group of the amino acid. But this parameter is sometimes appear to be same for dissimilar protein sequences as well. For example, an amino acid group position in a protein sequences is at 4^{th} and 6^{th} position in one sequence and in other is at 3^{rd} and 7^{th} from first amino acid group in the sequence. Here, in both the cases the distance from first amino acid group in both of the sequences is 10, but are at different places. In order to consider the protein sequences uniquely this distribution parameter has been taken into consideration. Similarly, η_i is represents as the ratio of sum of positions of i^{th} amino acid to the count of i^{th} amino acid in the protein sequence. We carry out several experiments to validate the accuracy of the proposed method in the following sections.

Analyzing Protein Sequences

In this proposed method, we applied Euclidean Distance to compute the distance among the feature vectors of protein sequences. Euclidian Distance is one of the simplest and most effective method, which has been used in many fields for measuring the distance like gene identification [23], tertiary protein structure comparison and constructions [47]. However, there are many other methods used for protein sequence comparison [9]. In our method we considered $A$and B be two protein Sequences, and $V(A)$ and $V(B)$ be two vectors representing their corresponding 80-D vectors. Euclidean distance between these two vectors is defined as below:

$d(A,B) = \sqrt{\sum_{i=1}^{80}(V_S[i]-V_T[i])^2}(5)$

Where $V(A)[i]$ and $V(B)[i]$ represent the i_{th} entry in two vectors $V(A)$ and $V(B)$. Smaller distance between sequences refers to the closeness of the sequences.

1.

Experimental result and discussions

Data set used with specification

In this article, the proposed methodology tested on four datasets likeND5 proteins of nine different species taken for analysis like Human ($\mathcal{H}$), P-Chimpanzee (PC), C-Chimpanzee (CC), Gorilla ($\mathcal{G}$) , Fin Whale ($\mathcal{FW}$), Blue Whale($\mathcal{BW}$), Rat ($\mathcal{R}$), Mouse($\mathcal{M}$), and Opossum($\mathcal{O}$) which are listed in Table 2. These sequences have length between 602 to 610 base pairs(bps). NADH dehydrogenase sub-unit 6 protein family also taken into consideration including Human ($\mathcal{H}$), Chimpanzee($\mathcal{C}$), Gorilla ($\mathcal{G}$), Wallaroo($\mathcal{W}$), Harbor-seal ($\mathcal{HS}$), Gray-seal(GS), Rat ($\mathcal{R}$) and Mouse($\mathcal{M}$) for similarity analysis *GenBank(www.ncbi. nlm.nih.gov)* . These data set are standard bench mark data used for sequence similarity analysis for validation of different computational procedures. These data sets are used before in other approaches [59, 21, 28, 61]. Further, another two datasets $F10$ glycoside hydrolase family with NCBI accession IDs: O59859, P56588, P33559, Q00177, P07986, P07528, P40943, P23556, P45703, and Q60041 and $G11$ of glycoside hydrolase family with NCBI IDs: P33557, P55328, P55331, P45705, P26220, P55334, Q06562, P55332, P55333, and P17137 are also considered to validate the proposed method.

Analysis of similarity between nine different proteins of ND5

For illustration of our proposed method, similarity among all the species of ND5 dataset. We calculated the Euclidean distance between all the nine ND5 protein sequences as shown in Table 3. The data have been collected from the*GenBank(www.ncbi.nlm.nih.gov)* namely: Human (Homo sapiens, AP_000649), Gorilla (Gorilla gorilla, NP_008222), Common Chimpanzee (Pantroglodytes, NP_008196), Pygmy Chimpanzee (Pan paniscus, NP_008209), Fin Whale (Balaenopteraphysalus, NP_006899), Blue Whale (Balaenopteramusculus, NP_007066), Rat(Rattusnorvegicus, AP_004902), Mouse (Mus musculus, NP_904338), and Opossum (Didelphis virginiana, NP_007105) as shown in Table 2. From Table 3, we observed that the Euclidean distance between CC,PC, $\mathcal{H}$ and $\mathcal{G}$ are quite small as comparison to other species in the same family. So, these four species are more similar with each other. The distance between$\mathcal{FW}$ and $\mathcal{BW}$ is also small that they are more similar with each other. There is also a small distance between $\mathcal{R}$ and $\mathcal{M}$ indicates the evolutionary closeness between them. Where as the Opossum species has large distance among other species which indicated comparatively large evolutionary between them. Corresponding to these as input, phylogenetic tree has been constructed as shown in Figure 1. The length of the branches of tree represents the lineages but we are focused to find the close relatedness among different species. On comparison of our approach with other ones, it has been found that there exists consistency with the result of evolution and biological history.

Analysis of similarity between eight different proteins of ND6

In order to examine our proposed method a sequence genes fromND6 (NADH dehydrogenase sub-unit 6 proteins) has been considered. The accession number of all the species are: Human(YP_003024037.1), Chimpanzee

(NP_008197), Wallaroo (NP_007405), Gorilla (NP_008223), Harbour Seal (NP_006939), Rat (AP_004903), Mouse (NP_904339), and Grey Seal *(NP_007080)* , the naming convention of these genes are as shown in Table 4. Then, we calculated the distance matrix of these set of gene sequences as shown in Table 5. As per our observation from this distance matrix$\mathcal{H}$, $\mathcal{C}$ and $\mathcal{G}$ are closely evolutionary related. The distance between $\mathcal{HS}$and GS are also very small, that is to say, they are very much similar with each other as compare to other species in this family. The corresponding phylogenetic tree has been constructed and shown in Figure 2. Phylogenetic tree obtained using this distance matrix found accurate based on their biological and revolutionary relationship. However, it is not quite sensible to say that$\mathcal{W}$ are much revolutionary close to $\mathcal{H}$,$\mathcal{C}$ and $\mathcal{G}$. This may because of loss of some physio-chemical properties as well as biological information.

In addition to the above two family , we carried out other protein families like $G10$ and $F11$ of the xylanases containing glycoside hydrolase families 10 and 11 respectively in our experiment to examine the usefulness of our method. Specifically, the $F10$ data set contains ten xylanases with NCBI accession IDs$O59859$, $P56588$, $P33559$, $Q00177$, $P07986$, $P07528$,$P40943$, $P23556$,$P45703$ and $Q60041$ respectively. The G11 data set also consists of ten xylanases with NCB : IDs$P33557$, $P55328$, $P55331$, $P45705$, $P26220$, $P55334$,$Q06562$, $P55332$, $P55333$ and $P17137$ respectively. Similar to ND5 and ND6, Euclidean distance of $G10$ and$F11$ sequences are computed as shown in Table 6 and Table 7. From Table 6, the NCBI- ID $P55332$, $P55333$,$P45705$ and $P17137$ are more similar as compare to others in the same family as they are more evolutionary. The corresponding phylogenetic tree is also generated by considering the distance matrix shown in Figure 3 and Figure 4 for $G10$ and $F11$ respectively. The phylogentic trees shows more consistent biological revolutionary relationship among all the species of $G10$ and $F11$ family.

The proposed method compare with other exiting methods.

The ClustalW platform is considered to be one of the most useful sequence alignment method for protein and DNA sequence analysis [57]. We have utilized the ClustalW multiple sequence alignment results and our proposed method results in form of distance matrices. In order to examine for the linear correlation among all proposed method and ClustalW method, the parametric based correlation analysis has been used. The greater the correlation coefficient between two sequence represents the stronger linear correlation. For ND5 data set, the results have been listed in Table 8. On comparing the results with those in Table 3, it has been found that the biological and evolutionary relationship listed above is in accordance to known phylogeny relationship.

The Correlation Coefficient is defined as the strength of linear relationship between two vectors. It is defined as ratio of covariance of variables to their standard deviations. We Used this parametric based correlation analysis to test the linear correlation. To find the relationship among our methods and ClustalW method, correlation coefficient has been calculated between these two methods. For calculating the correlation coefficient, rows from Table 3 and Table 8 has been taken into consideration. On taking the first row of similarity of Table 3 and similarity matrix of ClustalW(Table 8), correlation coefficient has been found to be 0.91367. Similarly for all other rows, correlation coefficient has been found, that has been listed in Table 9 and in Figure 5.

Let P and Q are to variable defined for positive integer. $P$and Q are said to be in linear correlation if the coefficient of correlation r satisfies $r_{0.05}(n-2) < |r| < r_{0.01}(n-2)$. and are in strong linear correlation if $|r| > r_{0.01}(n-2)$. on the basis of this, for considering ND5 dataset where $n = 9$and $0.666 < |r| <= 0.798$. Here, the variable P and Q are said to be in linear correlation and when $|r| > 0.798$, P and$Q$ are said to be in strong correlation. As a result, all species are in linearly correlated and except F-Whale. Considering our data sample size $n = 9$, which is too small, it implies that we may have high correlation coefficients. In order to validate our results, we examined significance analysis to check the strength of correlation between two sets. This analysis has been conducted for correlation coefficients greater than 0.7 through $t-$test. The value of alpha considered here for the significance analysis id 0.05 and corresponding $t-$value is 2.365. In Table 2 , we have considered only those t-values whose corresponding r values are greater than 0.7. On the basis of our computed results it can be said that r values do not occur by chance as all $t-$ values are greater than 2.365. All the nine$t-$ values satisfy $t > 2.365$ in our method, while there are only3,4, 7, 6, 5 $t-$values in other methods [29, 62, 42, 4, 11] respectively.

Similarly considering ND6 dataset, the results for ClustalW has been listed in Table 11. On comparing this data with data of Table 3 , it has been found that the biological and evolutionary relationship found by method listed above is in accordance to known phylogeny relationship. To find the relationship among our method and ClustalW method, correlation coefficient has been calculated between these two methods. For calculating the correlation coefficient, corresponding rows from Table 3 and Table 11 has been taken into account. The rows of similarity/ dissimilarity of Table 3 and similarity/dissimilarity matrix of ClustalW(Table 11), correlation coefficient has been found and listed in Table 12 and in Figure 6.

Two variable P and Q are said to be in linear correlation if the coefficient of correlation r satisfies$r_{0.05}(n-2) < |r| < r_{0.01}(n-2)$. and are in strong linear correlation if $|r| > r_{0.01}(n-2)$. Following the same, for ND6 dataset, $n = 8$, which implies when $0.707 < |r| <= 0.834$, so the variable P and Q are said to be in linear correlation and$|r| > 0.834$, so the variable P and Q are in strong correlation. On the basis of our results, all species of ND6are linearly correlated except $H-$Seal, $G-$Sealand $\mathcal{R}$. Comparing our results with ClustalW, we found out a strong correlation between two species. As the sample size $n = 8$, which is too small, it implies that we may have high correlation coefficients. The correlation coefficient is greater than 0.708through $t-$test in our proposed method. The value of αconsidered here for the significance analysis and id 0.05 and corresponding $t-$ value is 2.45. In Table 12 , we have considered only those $t-$ values whose corresponding $r$values are greater than 0.707. On the basis of our computed results it can be said that r values do not occur by chance as all $t-$values are greater than 2.45. Similarly, when we consider all seven$t-$ values, which are satisfy $t > 2.45$ in our method, while there are only two $t-$values in other methods [34], [13] respectively.

The Time Complexity of the proposed method is $O(n^2)$. It is known that the multiple sequence alignment is an NP-hard Problem. Taking into consideration the space required in our method has been also reduced as it does not stores coordinates of the amino acids for the values of x and y coordinates equal to the sequence length.

Conclusions

The result declared in the previous sections, shows that the characterization of amino acid based on their physio-chemical properties could be considered as a significant scheme for similarity analysis of protein sequences. On the other hand, a 80 dimension feature vector has been devised on the basis of distribution of physio-chemical properties of amino acid. This association is a great advantage for understanding the similarity among gene sequences of different families. The result obtained for nine species of ND5 and eight species of ND6 proved that our method is simple, convenient, intuitive and computationally less intensive. We also observed that the phylogenetic tree obtained by this method shows much biological and revolutionary relationship among the species. Further, our method tested on $G10$ and $F11$ data set of ten species each which shows appropriate phylogeny. We believe that the novel features and the result reported in this article will be useful for biologist in the similar problems related to DNA and RNA sequences.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgements

Not Applicable.

References

1. Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology* , 215(3):403–410, 1990.
2. Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology* , 215(3):403–410, 1990.

3. Christian B Anfinsen. Principles that govern the folding of protein chains. *Science* , 181(4096):223–230, 1973.
4. Christian B Anfinsen. Principles that govern the folding of protein chains. *Science* , 181(4096):223–230, 1973.
5. Konstantin Berlin, Sergey Koren, Chen-Shan Chin, James P Drake, Jane M Landolin, and Adam M Phillippy. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing.*Nature biotechnology* , 33(6):623, 2015.
6. Guillaume Bernard, Cheong Xin Chan, Yao-ban Chan, Xin-Yi Chua, Yingnan Cong, James M Hogan, Stefan R Maetschke, and Mark A Ragan. Alignment-free inference of hierarchical and reticulate phylogenomic relationships. *Briefings in bioinformatics* , 20(2):426–435, 2017.
7. B Edwin Blaisdell. A measure of the similarity of sets of sequences not requiring sequence alignment. *Proceedings of the National Academy of Sciences* , 83(14):5155–5159, 1986.
8. Oliver Bonham-Carter, Joe Steele, and Dhundy Bastola. Alignment-free genetic sequence comparisons: a review of recent approaches by word analysis. *Briefings in bioinformatics* , 15(6):890–905, 2013.
9. Vibha Bafna Bora, Ashwin G Kothari, and Avinash G Keskar. Robust automatic pectoral muscle segmentation from mammograms using texture gradient and euclidean distance regression. *Journal of digital imaging* , 29(1):115–125, 2016.
10. Tolga Can and Y-F Wang. Ctss: a robust and efficient method for protein structure alignment based on local geometrical and biological features. In *Computational systems bioinformatics. CSB2003. Proceedings of the 2003 IEEE bioinformatics conference. CSB2003* , pages 169–179. IEEE, 2003.
11. olga Can and Y-F Wang. Ctss: a robust and efficient method for protein structure alignment based on local geometrical and biological features. In *Computational systems bioinformatics. CSB2003. Proceedings of the 2003 IEEE bioinformatics conference. CSB2003* , pages 169–179. IEEE, 2003.
12. Wah Chiu, Matthew L Baker, Wen Jiang, and Z Hong Zhou. Deriving folds of macromolecular complexes through electron cryomicroscopy and bioinformatics approaches. *Current opinion in structural biology* , 12(2):263–269, 2002.
13. Wah Chiu, Matthew L Baker, Wen Jiang, and Z Hong Zhou. Deriving folds of macromolecular complexes through electron cryomicroscopy and bioinformatics approaches. *Current opinion in structural biology* , 12(2):263–269, 2002.
14. Kuo-Chen Chou and Hong-Bin Shen. Recent progress in protein subcellular location prediction. *Analytical biochemistry* , 370(1):1, 2007.
15. Kuo-Chen Chou. Prediction of human immunodeficiency virus protease cleavage sites in proteins. *Analytical biochemistry* , 233(1):1–14, 1996.
16. Qi Dai, Yan Li, Xiaoqing Liu, Yuhua Yao, Yunjie Cao, and Pingan He. Comparison study on statistical features of predicted secondary structures for protein structural class prediction: From content to position. *BMC bioinformatics* , 14(1):152, 2013.
17. Jayanta Kumar Das, Provas Das, Korak Kumar Ray, Pabitra Pal Choudhury, and Siddhartha Sankar Jana. Mathematical characterization of protein sequences using patterns as chemical group combinations of amino acids. *PloS one* , 11(12):e0167651, 2016.
18. Burk A Dehority. *Rumen microbiology* , volume 372. Nottingham University Press Nottingham, 2003.
19. Shuyan Ding, Shengli Zhang, Yang Li, and Tianming Wang. A novel protein structural classes prediction method based on predicted secondary structure. *Biochimie* , 94(5):1166–1171, 2012.
20. Ali El-Lakkani and Seham El-Sherif. Similarity analysis of protein sequences based on 2d and 3d amino acid adjacency matrices.*Chemical Physics Letters* , 590:192–195, 2013.
21. Moheb I Abo el Maaty, Mervat M Abo-Elkhier, and Marwa A Abd Elwahaab. 3d graphical representation of protein sequences and their statistical characterization. *Physica A: Statistical Mechanics and Its Applications* , 389(21):4668–4676, 2010.
22. Joseph Felsenstein. Phylip (phylogeny inference package) version 3.6. distributed by the author. *http://www. evolution. gs. washington. edu/phylip. html* , 2004.
23. Antara Ghosh and Soma Barman. Application of euclidean distance measurement and principal component analysis for gene identification.*Gene* , 583(2):112–120, 2016.

24. Charles Miller Grinstead and James Laurie Snell. *Introduction to probability* . American Math. Soc., 2012.
25. MK Gupta, R Niyogi, and M Misra. An alignment-free method to find similarity among protein sequences via the general form of chouâ\euros pseudo amino acid composition. *SAR and QSAR in Environmental Research* , 24(7):597–609, 2013.
26. Eugene Hamori and John Ruskin. H curves, a novel method of representation of nucleotide series especially suited for long dna sequences. *Journal of Biological Chemistry* , 258(2):1318–1327, 1983.
27. Bernhard Haubold. Alignment-free phylogenetics and population genetics. *Briefings in bioinformatics* , 15(3):407–418, 2013.
28. Ping-an He, Jinzhou Wei, Yuhua Yao, and Zhixin Tie. A novel graphical representation of proteins and its application. *Physica A: Statistical Mechanics and its Applications* , 391(1-2):93–99, 2012.
29. Ping-an He, Jinzhou Wei, Yuhua Yao, and Zhixin Tie. A novel graphical representation of proteins and its application. *Physica A: Statistical Mechanics and its Applications* , 391(1-2):93–99, 2012.
30. Zhisong He, Jian Zhang, Xiao-He Shi, Le-Le Hu, Xiangyin Kong, Yu-Dong Cai, and Kuo-Chen Chou. Predicting drug-target interaction networks based on functional groups and biological features. *PloS one* , 5(3):e9603, 2010.
31. Tao Huang, Shen Niu, Zhongping Xu, Yun Huang, Xiangyin Kong, Yu-Dong Cai, and Kuo-Chen Chou. Predicting transcriptional activity of multiple site p53 mutants based on hybrid properties. *PLoS One* , 6(8):e22940, 2011.
32. Xiaoqiu Huang and Jinhui Zhang. Methods for comparing a dna sequence with a protein sequence. *Bioinformatics* , 12(6):497–506, 1996.
33. Le-Le Hu, Tao Huang, Yu-Dong Cai, and Kuo-Chen Chou. Prediction of body fluids where proteins are secreted into based on protein interaction network. *PLoS One* , 6(7):e22989, 2011.
34. Chris A Kaiser, Monty Krieger, Harvey Lodish, and Arnold Berk.*Molecular cell biology.* WH Freeman, 2007.
35. Katsuko Komatsu, Shu Zhu, Hirotoshi Fushimi, Tran Kim Qui, Shaoqing Cai, and Shigetoshi Kadota. Phylogenetic analysis based on 18s rrna gene and matk gene sequences of panax vietnamensis and five related species. *Planta medica* , 67(05):461–465, 2001.
36. Liang Kong, Lichao Zhang, and Jinfeng Lv. Accurate prediction of protein structural classes by incorporating predicted secondary structure information into the general form of chou's pseudo amino acid composition. *Journal of Theoretical Biology* , 344:12–18, 2014.
37. Sudhir Kumar, Glen Stecher, and Koichiro Tamura. Mega7: molecular evolutionary genetics analysis version 7.0 for bigger datasets.*Molecular biology and evolution* , 33(7):1870–1874, 2016.
38. Tian Liu and Cangzhi Jia. A high-accuracy protein structural class prediction algorithm using predicted secondary structural information.*Journal of theoretical biology* , 267(3):272–275, 2010.
39. Bi-Qing Li, Tao Huang, Lei Liu, Yu-Dong Cai, and Kuo-Chen Chou. Identification of colorectal cancer related genes with mrmr and shortest path in protein-protein interaction network. *PloS one* , 7(4):e33393, 2012.
40. Bi-Qing Li, Le-Le Hu, Shen Niu, Yu-Dong Cai, and Kuo-Chen Chou. Predict and analyze s-nitrosylation modification sites with the mrmr and ifs approaches. *Journal of Proteomics* , 75(5):1654–1665, 2012.
41. Chun Li, Lili Xing, Xin Wang, et al. 2-d graphical representation of protein sequences and its application to coronavirus phylogeny.*BMB Rep* , 41(3):217–222, 2008.
42. Chun Li, Lili Xing, Xin Wang, et al. 2-d graphical representation of protein sequences and its application to coronavirus phylogeny.*BMB Rep* , 41(3):217–222, 2008.
43. Yushuang Li, Tian Song, Jiasheng Yang, Yi Zhang, and Jialiang Yang. An alignment-free algorithm in comparing the similarity of protein sequences based on pseudo-markov transition probabilities among amino acids. *PloS one* , 11(12):e0167430, 2016.
44. Zengchao Mu, Jing Wu, and Yusen Zhang. A novel method for similarity/dissimilarity analysis of protein sequences. *Physica A: Statistical Mechanics and its Applications* , 392(24):6361–6366, 2013.
45. Hasan H Otu and Khalid Sayood. A new sequence distance measure for phylogenetic tree construction.

Bioinformatics , 19(16):2122–2130, 2003.

46. William R Pearson. [5] rapid and sensitive sequence comparison with fastp and fasta. 1990.
47. Michal J Pietal, Janusz M Bujnicki, and Lukasz P Kozlowski. Gdfuzz3d: a method for protein 3d structure reconstruction from contact maps, based on a non-euclidean distance function. *Bioinformatics* , 31(21):3499–3505, 2015.
48. Luca Pinello, Giosue Lo Bosco, and Guo-Cheng Yuan. Applications of alignment-free methods in epigenomics. *Briefings in Bioinformatics* , 15(3):419–430, 2013.
49. Dan Ralescu and Gregory Adams. The fuzzy integral. *Journal of Mathematical Analysis and Applications* , 75(2):562–570, 1980.
50. Jie Ren, Kai Song, Minghua Deng, Gesine Reinert, Charles H Cannon, and Fengzhu Sun. Inference of markovian properties of molecular sequences from ngs data and applications to comparative genomics.*Bioinformatics* , 32(7):993–1000, 2015.
51. Ranjeet Kumar Rout, Pabitra Pal Choudhury, Santi Prasad Maity, BS Daya Sagar, and Sk Sarif Hassan. Fractal and mathematical morphology in intricate comparison between tertiary protein structures.*Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* , 6(2):192–203, 2018.
52. Ajay Kumar Saw, Binod Chandra Tripathy, and Soumyadeep Nandi. Alignment-free similarity analysis for protein sequences based on fuzzy integral. *Scientific reports* , 9(1):2775, 2019.
53. Ariya Shajii, Deniz Yorukoglu, Yun William Yu, and Bonnie Berger. Fast genotyping of known snps through approximate k-mer matching.*Bioinformatics* , 32(17):i538–i544, 2016.
54. Ping Wang, Lele Hu, Guiyou Liu, Nan Jiang, Xiaoyun Chen, Jianyong Xu, Wen Zheng, Li Li, Ming Tan, Zugen Chen, et al. Prediction of antimicrobial peptides based on sequence alignment and feature selection methods. *PloS one* , 6(4):e18476, 2011.
55. Leyi Wei, Minghong Liao, Xing Gao, and Quan Zou. An improved protein structural classes prediction method by incorporating both sequence and structure information. *IEEE transactions on nanobioscience* , 14(4):339–349, 2014.
56. Leyi Wei, Minghong Liao, Xing Gao, and Quan Zou. Enhanced protein fold prediction method through a novel feature extraction technique.*IEEE transactions on nanobioscience* , 14(6):649–659, 2015.
57. Zhi-Cheng Wu, Xuan Xiao, and Kuo-Chen Chou. 2d-mh: A web-server for generating graphic representation of protein sequences based on the physicochemical properties of their constituent amino acids.*Journal of theoretical biology* , 267(1):29–34, 2010.
58. Jian-Yi Yang, Zhen-Ling Peng, and Xin Chen. Prediction of protein structural classes for low-homology sequences based on predicted secondary structure. *BMC bioinformatics* , 11(1):S9, 2010.
59. Yu-Hua Yao, Qi Dai, Chun Li, Ping-An He, Xu-Ying Nan, and Yao-Zhou Zhang. Analysis of similarity/dissimilarity of protein sequences.*Proteins: Structure, Function, and Bioinformatics* , 73(4):864–871, 2008.
60. Chenglong Yu, Rong L He, and Stephen S-T Yau. Protein sequence comparison based on k-string dictionary. *Gene* , 529(2):250–256, 2013.
61. Hong-Jie Yu and De-Shuang Huang. Novel 20-d descriptors of protein sequences and it's applications in similarity analysis. *Chemical Physics Letters* , 531:261–266, 2012.
62. Hong-Jie Yu and De-Shuang Huang. Novel 20-d descriptors of protein sequences and it's applications in similarity analysis. *Chemical Physics Letters* , 531:261–266, 2012.
63. Lichao Zhang, Xiqiang Zhao, and Liang Kong. Predict protein structural class for low-similarity sequences by evolutionary difference information into the general form of chou s pseudo amino acid composition. *Journal of theoretical biology* , 355:105–110, 2014.
64. Shengli Zhang, Shuyan Ding, and Tianming Wang. High-accuracy prediction of protein structural class for low-similarity sequences based on predicted secondary structure. *Biochimie* , 93(4):710–714, 2011.
65. Shengli Zhang, Yunyun Liang, and Xiguo Yuan. Improving the prediction accuracy of protein structural class: Approached with alternating word frequency and normalized lempel–ziv complexity. *journal of theoretical biology* , 341:71–77, 2014.
66. Shengli Zhang, Feng Ye, and Xiguo Yuan. Using principal component analysis and support vector ma-

chine to predict protein structural class for low-similarity sequences via pssm. *Journal of Biomolecular Structure and Dynamics* , 29(6):1138–1146, 2012.

67. Zhaojun Zhang and Wei Wang. Rna-skim: a rapid method for rna-seq quantification at transcript level. *Bioinformatics* , 30(12):i283–i292, 2014.

68. Andrzej Zielezinski, Susana Vinga, Jonas Almeida, and Wojciech M Karlowski. Alignment-free sequence comparison: benefits, applications, and tools. *Genome biology* , 18(1):186, 2017.

Figure Legends

Figure 1. Phylogenetic-tree of different species of ND5 family generated by the proposed method (cluster dendrogram using UPGMA distance method).

Figure 2. Phylogentic-tree of eight different species of ND6 generated by our proposed method (cluster dendrogram using UPGMA distance method).

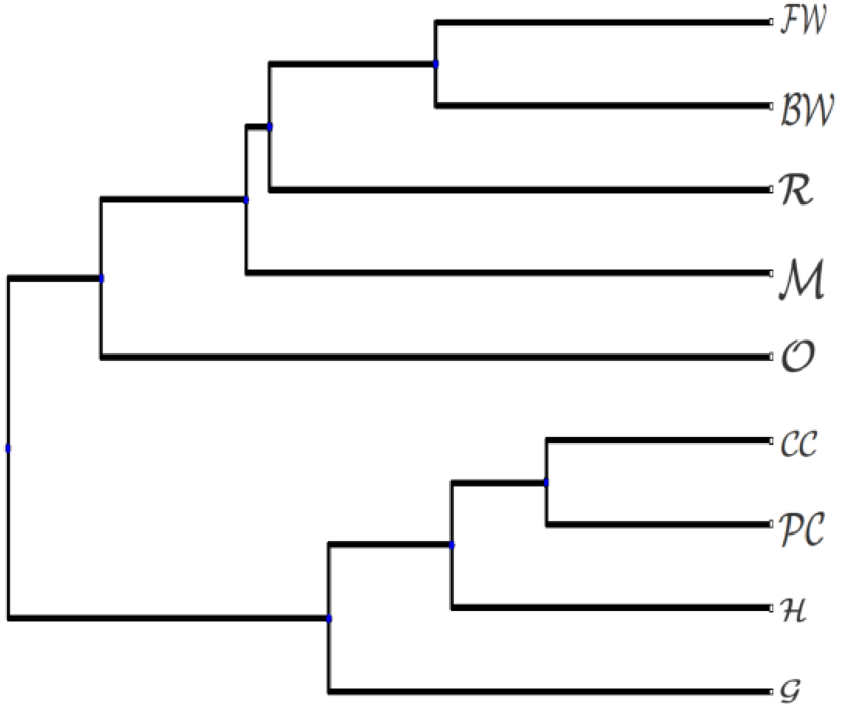
Figure 3. Shows the phylogentic tree of 10 different species of G10 family generated by our proposed method (cluster dendrogram using UPGMA distance method).

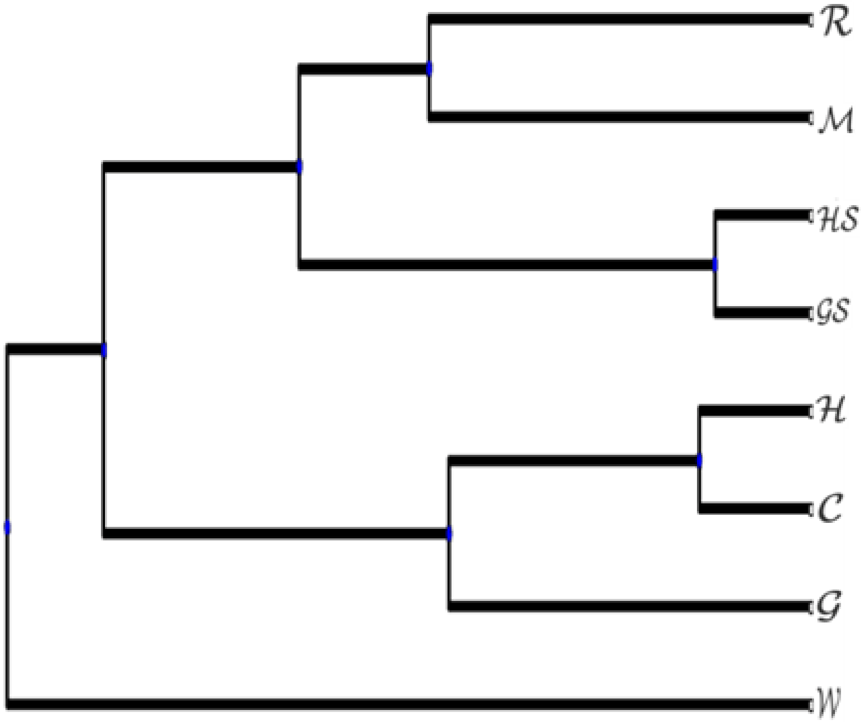
Figure 4 . Shows the phylogentic tree of 10 different species of F11 family constructed by our proposed method (cluster dendrogram using UPGMA distance method.

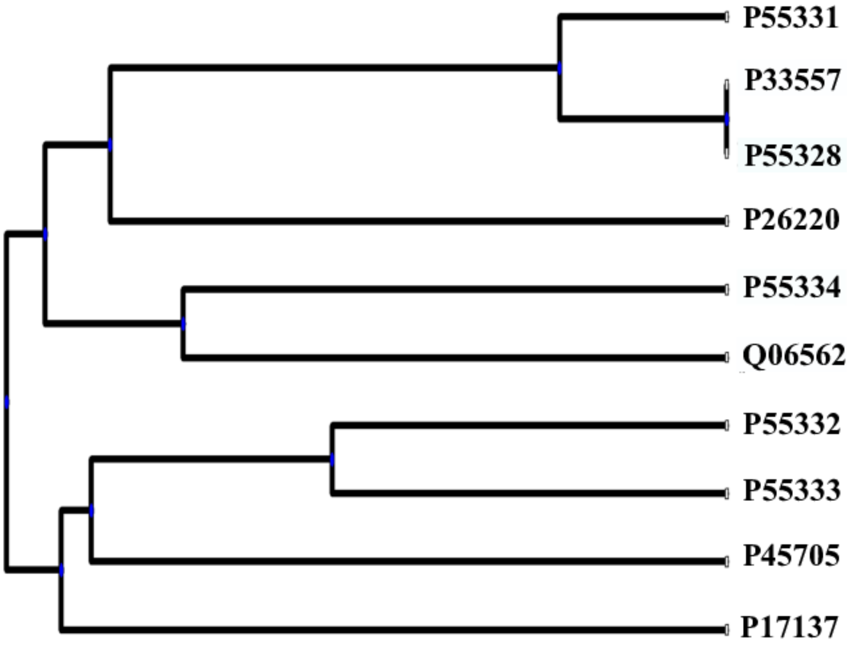
Figure 5. The correlation coefficients for nine ND5 proteins of our method(cluster dendrogram using UPGMA distance method) and other methods in [29, 62, 42, 4, 11]referring to ClustalW.

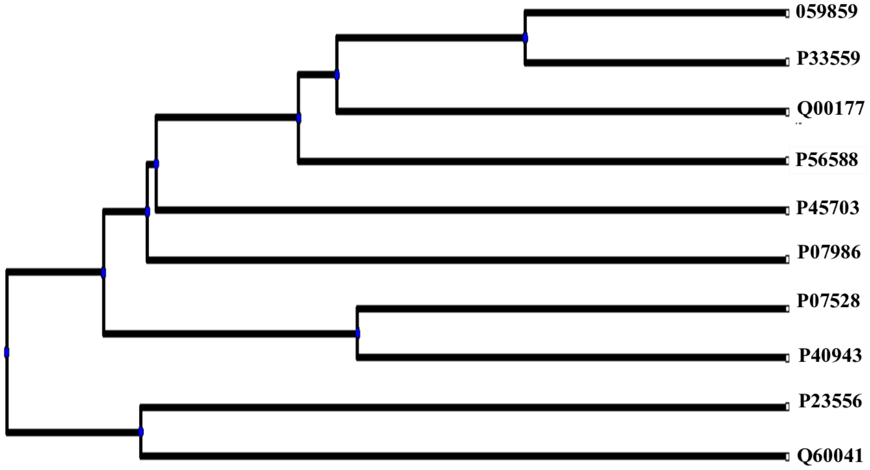
Figure 6 . The correlation coefficients for nine ND5 proteins of our method (cluster dendrogram using UPGMA distance method) and other methods in Ref [13] and Ref [34] referring to ClustalW.

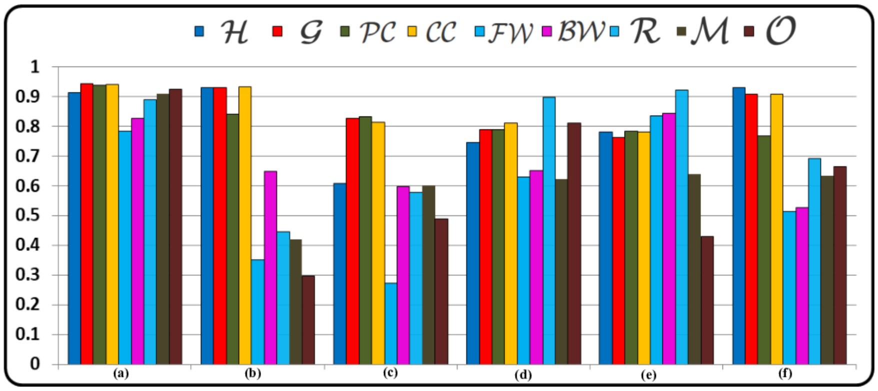
Figures

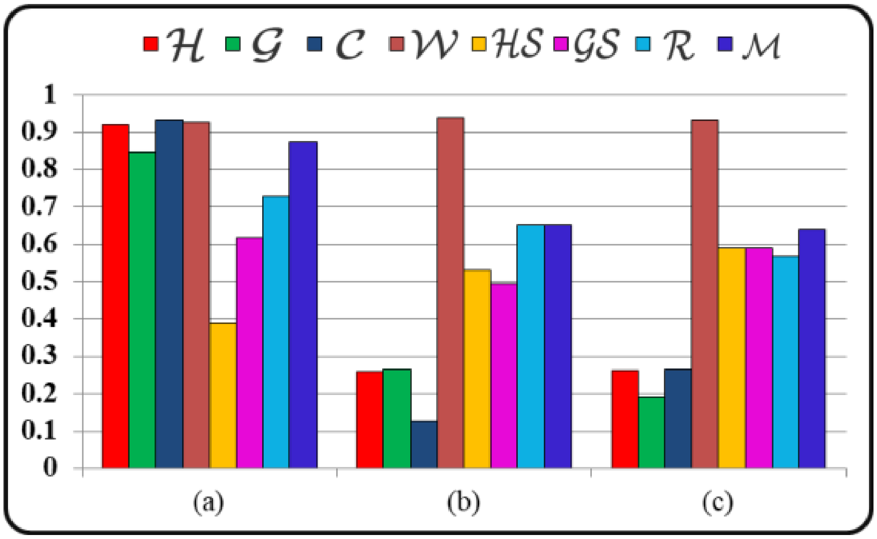



$\mathcal{R}$
$\mathcal{M}$
$\mathcal{HS}$
$\mathcal{GS}$
$\mathcal{H}$
$\mathcal{C}$
$\mathcal{G}$
$\mathcal{W}$


P55331
P33557
P55328
P26220
P55334
Q06562
P55332
P55333
P45705
P17137


059859
P33559
Q00177
P56588
P45703
P07986
P07528
P40943
P23556
Q60041


H
G
PC
CC
FW
BW
R
M
O
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
(a)
(b)
(c)
(d)
(e)
(f)


H
G
C
W
HS
GS
R
M
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
(a)
(b)
(c)

Tables

Table 1. Classification of amino acid based on their chemical properties.

Group Name based on physio-chemical Nature	**Amino Acids**
Acidic	D (Aspartate), E (Glutamate)
Basic	H (Histidine), R (Arginine), K (Lysine)
Acidic amide	N (Asparagine), Q (Glutamine)
Aromatic side chain	F (Phenylalanine), Y (Tyrosine), W (Tryptophan)
Aliphatic side chain	V (Valine), A (Alanine), I (Isoleucine), L (Leucine), G (Glycine)
Hydroxyl containing	T (Threonine), S (Serine)
Cyclic	P (Proline)
Sulfur containing	C (Cysteine), M (Methionine)

Table 2. Shows the ND5 protein sequences from nine different species specific ID from NCBI.

Species Name	**NCBI-ID**	**Length of the Sequence**
Homo Sapiens (Human($\mathcal{H}$))	AP_000649	603
Gorilla (Gorilla ($\mathcal{G}$))	NP_008222	603
Paniscus (Pigmy-chimpanzee (PC))	NP_008209	603
Patroglodytes (Common-chimpanzee (CC))	NP_008196	603
Balenopteraphysalus (Fin-whale (FW))	NP_006899	606
Balenopteramusculus (Blue-whale (BW))	NP_007066	606
Tattusnorvegicus (Rat ($\mathcal{R}$))	NP_004902	610
Musculus (Mouse ($\mathcal{M}$))	NP_904338	607
Didelphis-virginiana (Opossum ($\mathcal{O}$))	NP_007105	602

Table 3. Shows the Euclidean distance for nine different species of ND5 dataset.

	$\mathcal{H}$	$\mathcal{G}$	PC	CC	FW	BW	$\mathcal{R}$	$\mathcal{M}$	$\mathcal{O}$
$\mathcal{H}$	0	0.23064	0.17773	0.13416	0.32402	0.29557	0.33018	0.32907	0.37786
$\mathcal{G}$		0	0.21192	0.20465	0.4552	0.43538	0.42633	0.42179	0.46362
PC			0	0.10982	0.39658	0.32733	0.37827	0.3617	0.38343
CC				0	0.3623	0.3121	0.36537	0.33012	0.36661
FW					0	0.16366	0.23612	0.24082	0.36521
BW						0	0.2531	0.27765	0.3167
$\mathcal{R}$							0	0.24998	0.2949
$\mathcal{M}$								0	0.3298
$\mathcal{O}$									0

Table 4. Shows the ND6 proteins from eight different Species with length from NCBI.

Species Name	**ID(NCBI)**	**Length**
Human(Homo Sapiens) ($\mathcal{H}$)	YP_003024037	174
Gorilla ($\mathcal{G}$)	NP_008223	174
Chimpanzee ($\mathcal{C}$)	NP_008197	174
Wallaroo ($\mathcal{W}$)	NP_007405	167
Harbor Seal ((HS))	NP_006939	175
Gray Seal (GS)	NP_007080	175
Rat ($\mathcal{R}$)	AP_004903	172

Mouse ($\mathcal{M}$)	NP_904339	172

Table 5. Shows the Euclidean distance for eight different species of ND6 dataset.

	$\mathcal{H}$	$\mathcal{C}$	$\mathcal{W}$	$\mathcal{G}$	HS	$\mathcal{R}$	$\mathcal{M}$	GS
$\mathcal{H}$	0	0.17345	1.43722	0.56252	1.10975	1.0833	1.3178	1.05992
$\mathcal{C}$		0	1.48613	0.55556	1.13161	1.12388	1.34884	1.08749
$\mathcal{W}$			0	1.34584	1.09474	1.13112	1.0698	1.12556
$\mathcal{G}$				0	0.88894	0.94691	1.16781	0.84555
HS					0	0.62522	0.9186	0.14937
$\mathcal{R}$						0	0.59008	0.64615
$\mathcal{M}$							0	0.97191
GS								0

Table 6. Shows the Euclidean distance of 10 different species of G10 family.

	O59859	**P56588**	**P33559**	**Q00177**	**P07986**	**P07528**	**P40943**	**P23556**	**P45703**	**Q60041**
O59859	0	0.39325	0.23584	0.3872	0.58341	0.59811	0.56592	0.70144	0.59211	0.68837
P56588		0	0.41756	0.50865	0.65723	0.64931	0.6726	0.67524	0.56879	0.58614
P33559			0	0.42422	0.58154	0.63333	0.63459	0.73321	0.58734	0.68134
Q00177				0	0.45625	0.5463	0.62323	0.65885	0.52477	0.76296
P07986					0	0.63273	0.63388	0.70408	0.60316	0.89509
P07528						0	0.38694	0.68325	0.54357	0.73348
P40943							0	0.64219	0.65628	0.75863
P23556								0	0.70227	0.58209
P45703									0	0.64184
Q60041										0

Table 7. Shows the Euclidean distance of 10 different species of F11 family.

	P33557	**P55328**	**P55331**	**P45705**	**P26220**	**P55334**	**Q06562**	**P55332**	**P55333**	**P17137**
P33557	0	0	0.18332	0.88675	0.68962	0.7375	0.81858	0.72524	0.86874	0.76299
P55328		0	0.18332	0.88675	0.68962	0.7375	0.81858	0.72524	0.86874	0.76299
P55331			0	0.85406	0.64525	0.67866	0.77247	0.67194	0.82831	0.702
P45705				0	0.89796	0.87705	0.98011	0.70644	0.68477	0.75726
P26220					0	0.6595	0.7466	0.69372	0.82564	0.71154
P55334						0	0.59538	0.64301	0.76547	0.66001
Q06562							0	0.68119	0.71404	0.9222
P55332								0	0.43159	0.65881
P55333									0	0.76945
P17137										0

Table 8. Shows the Euclidean distance of ten different species of ND5 family using ClustalW.

	$\mathcal{G}$	PC	CC	FW	BW	$\mathcal{R}$	$\mathcal{M}$	$\mathcal{O}$
$\mathcal{H}$	10.7	7.1	6.9	41.0	41.3	50.2	48.9	50.4
$\mathcal{G}$		9.7	9.9	42.7	42.4	51.4	49.9	54.0
PC			5.1	40.1	40.1	50.2	48.9	50.1

CC	40.4	40.4	50.8	49.6	51.4
FW		3.5	45.3	46.8	52.7
BW			45.0	45.9	52.7
$\mathcal{R}$				25.9	54.0
$\mathcal{M}$					50.8

Table 9. Shows the correlation coefficient of our proposed methods with other methods. Here correlation coefficients have been computed using ClustalW.

	Proposed	**Ref[29]**	**Ref[62]**	**Ref[42]**	**Ref[4]**	**Ref[11]**
$\mathcal{H}$	0.91367	0.9306	0.609	0.747	0.7819	0.9294
$\mathcal{G}$	0.94417	0.9293	0.8278	0.7902	0.763	0.9101
PC	0.93886	0.8403	0.8332	0.7887	0.7856	0.7673
CC	0.94091	0.9344	0.8148	0.811	0.7808	0.9096
FW	0.78355	0.3508	0.2729	0.631	0.836	0.5143
BW	0.82812	0.6486	0.5964	0.6523	0.843	0.5274
$\mathcal{R}$	0.88939	0.4453	0.5777	0.897	0.9213	0.6919
$\mathcal{M}$	0.90924	0.4192	0.6027	0.6229	0.6391	0.6352
$\mathcal{O}$	0.92575	0.2975	0.4899	0.8109	0.4299	0.6663

Table 10. Shows the significant values (t-value computed using ClustalW) among proposed methods of ND5 dataset.

	Proposed	**Ref[29]**	**Ref[62]**	**Ref[42]**	**Ref[4]**	**Ref[11]**
$\mathcal{H}$	5.94736	6.72645	-	2.97278	3.31842	6.66252
$\mathcal{G}$	7.58225	6.65726	3.90385	3.4114	3.12301	5.81072
PC	7.21463	4.10097	3.98653	3.39424	3.35928	3.16564
CC	7.35085	6.93996	3.71848	3.66758	3.30644	5.79221
FW	3.33652	-	-	-	4.03085	-
BW	3.90865	-	-	-	4.14633	-
$\mathcal{R}$	5.14734	-	-	5.36895	6.26852	2.5352
$\mathcal{M}$	5.77896	-	-	-	-	-
$\mathcal{O}$	6.47731	-	-	3.66626	-	-

Table 11. Shows the Euclidean distance of eight different species of ND6 using ClustalW.

	$\mathcal{H}$	$\mathcal{G}$	$\mathcal{C}$	$\mathcal{W}$	HS	GS	$\mathcal{R}$	$\mathcal{M}$
$\mathcal{H}$	0	3.5	4.1	92.9	59.4	60.6	79.9	72.4
$\mathcal{G}$		0	4.8	92.9	60.6	61.9	83	75.3
$\mathcal{C}$			0	94.8	61.9	63.1	81.4	75.3
$\mathcal{W}$				0	88.3	84.8	106.1	95.8
HS					0	2.9	69.6	64.3
GS						0	66.9	64.3
$\mathcal{R}$							0	23
$\mathcal{M}$								0

Table 12. Shows the correlation coefficient (computed using ClustalW) among proposed methods of ND6 dataset.

	Proposed	**Ref [34]**	**Ref[13]**
$\mathcal{H}$	0.92169	0.2579	0.2614
$\mathcal{G}$	0.84737	0.2647	0.191
$\mathcal{C}$	0.93198	0.1253	0.2645
$\mathcal{W}$	0.92769	0.9397	0.9325
HS	0.39047	0.5307	0.5896
GS	0.61872	0.4928	0.5895
$\mathcal{R}$	0.72954	0.6525	0.5682
$\mathcal{M}$	0.87448	0.6508	0.6387

Table 13. Shows the significant values (t-values computed using ClustalW) among proposed methods.

	Proposed	**Ref [34]**	**Ref [13]**
$\mathcal{H}$	6.28613	-	-
$\mathcal{G}$	4.22207	-	-
$\mathcal{C}$	6.80199	-	-
$\mathcal{W}$	6.57407	7.26963	6.83103
HS	-	-	-
GS	-	-	-
$\mathcal{R}$	2.82216	-	-
$\mathcal{M}$	4.76982	-	-