An alignment-free method for bulk comparison of protein sequences from different species

B. DOĞAN

***Abstract*—The available number of protein sequences rapidly increased with the development of new sequencing techniques. This in turn led to an urgent need for the development of new computational methods utilizing these data for the solution of different biological problems. One of these problems is the comparison of protein sequences from different species to reveal their evolutional relationship. Recently, several alignment-free methods have been proposed for this purpose. Here in this study, we also proposed an alignment-free method for the same purpose. Different from the existing methods, the proposed method not only allows for a pairwise comparison of two protein sequences, but also it allows for a bulk comparison of multiple protein sequences simultaneously. Computational results performed on gold-standard datasets showed that, bulk comparison of multiple sequences is much faster than its pairwise counterpart and the proposed method achieves a performance which is quite competitive with the state-of-the-art alignment-based method, ClustalW.**

Index Terms— **ClustalW, ND5 proteins, Phylogenetic analysis, Protein sequence similarity, Sequence comparison.**

I. INTRODUCTION

DEVELOPMENT OF next-generation sequencing technologies led to a dramatic increase in the number of available DNA and protein sequences. It is quite crucial to effectively extract the biological information provided with this huge number of sequences which has given rise to new research challenges in computational biology and bioinformatics. Similarity analysis of protein sequences from different species is one of these research challenges. With the development of new tools for protein sequence similarity analysis, scientist will be able to elucidate the function of unknown proteins which may shed light on identification of potential drug targets and to gain insights on underlying molecular mechanisms of diseases [1].

BERAT DOĞAN, is with Department of Biomedical Engineering, Inonu University, Malatya, Turkey, (e-mail: berat.dogan@inonu.edu.tr)

https://orcid.org/0000-0003-4810-1970

Manuscript received March 16, 2019; accepted Sep 23, 2019.
DOI: 10.17694/bajece.540873

In literature, methods proposed for the sequence similarity analysis are usually investigated under two different groups: alignment-based methods and alignment-free methods. In alignment-based methods, a score function is used to represent insertion, deletion, and substitution of nucleotides or amino acids in the compared biological sequences. The overall objective of these methods is to align the sequences with the highest scores [2-7]. Although, alignment-based methods are successfully utilized for sequence similarity analysis, they are generally time consuming and memory demanding. Therefore, alignment-free methods are proposed as computationally inexpensive alternatives to alignment-based methods.

On the other hand, the main difficulty of alignment-free methods is the need for an effective method to map a protein sequence into a numerical format (either a vector or a matrix) that could be used in subsequent analyses. In literature, physicochemical properties of amino acids are usually utilized for this purpose. In [8], based on two physicochemical properties of amino acids, authors first converted a protein sequence into a three-letter sequence. Then, based on this three-letter sequence, they obtained a graph without loops and multiple edges and its geometric line adjacency matrix. Next, to characterize a protein sequence numerically, they constructed a generalized PseAAC (pseudo amino acid composition) model. By using the obtained numerical representation of protein sequences, similarity analysis among β-globin proteins of seventeen species and seventy-two spike proteins of coronaviruses were made. They showed that the resulting clusters agreed well with the established taxonomic groups. In [9], authors proposed a novel position-feature-based model for protein sequences by employing physicochemical properties of amino-acids and the measure of graph energy. The proposed method puts emphasis on sequence order information and describes local dynamic distributions of sequences, from which one can get characteristic B-vector. Afterwards, they applied the relative entropy to the sequences representing B-vectors to measure their similarity. They showed that the proposed method competes with the widely utilized alignment-based method, ClustalW [10]. In [11], authors proposed a method to analyze the similarity of proteins by Discrete Fourier Transform (DFT) and Dynamic Time Warping (DTW). They first converted the protein sequences into numerical sequences according to their physicochemical properties. Next, they obtained the power spectra of sequences from DFT and extended the spectra to the

same length to calculate the distance between different sequences by DTW. They tested their method on different datasets and the results were compared with the existing methods. They showed that the proposed method overperformed the existing methods. In [12], based on the three important physicochemical properties of amino acids: the hydropathy index, polar requirement and chemical composition of the side chain, authors proposed a 24-dimensional feature vector describing the composition of amino acids in protein sequences. The results on beta-globin, mammals, and three virus datasets showed that the proposed method is fast and accurate for classifying proteins and inferring the phylogeny of organisms. In [13], based on eight physicochemical properties of amino acids, authors proposed a 40-dimensional feature vector for numerical characterization of each protein sequence. They used the cosine distances of feature vectors to measure the similarity of proteins. Analysis results performed over two real datasets demonstrated that the new scheme is effective in similarity research and phylogenetic analysis. In [14], based on three physicochemical properties of amino acids, authors reduced a protein primary sequence into a six-letter sequence, and then they extracted a set of elements which reflect the global and local sequence-order information. Combining these elements with the frequencies of 20 native amino acids, they obtained a numerical vector to characterize each protein sequence. The utility of the proposed approach was illustrated by phylogenetic analysis and identification of DNA-binding proteins. In [15], based on two physicochemical properties of amino acids, a 2D graphical representation of protein sequences is proposed. The proposed graphical representation of proteins is then used to obtain a numerical vector for each protein which is then used to compute the similarity of different proteins. Experiments performed on ND5 proteins of nine species showed that their method is simple and effective. In [16], based on 12 major physicochemical properties of amino acids and the principal component analysis (PCA) method, authors proposed a simple and intuitive 2D graphical mapping method for protein sequences. Next, they extracted a 20D vector from the graphical mapping which is used to characterize a protein sequence. To validate the proposed method, they first gave a comparison of protein sequences, which consists of nine ND6 proteins. They showed that the similarity/dissimilarity matrix for nine ND6 proteins correctly reveals their evolutionary relationship. Next, they performed the cluster analysis of HA genes of influenza A (H1N1) isolates, results of which is shown to be consistent with the known evolution fact of the H1N1 virus. In [17], authors proposed a new protein map which is based on ten physicochemical properties of amino acids. The proposed method both considers phylogenetic factors arising from amino acid mutations and provides computational efficiency for the huge amount of data. Authors showed that the proposed model is easier and quicker in handling protein sequences than multiple alignment methods and gives protein classification greater evolutionary significance at the amino acid sequence level. In [18], author presented a 2D spectrum-like graphical representation of protein sequences based on the hydrophobicity scale of amino acids. The frequencies of amplitudes of 4-subsequences were adopted to characterize a spectrum-like graph, and a 17D vector was used for numerical representation of a protein sequence. By using protein sequences from the mitochondrion genome of 20 different species, they compared their method to ClustalW to show the utility of their method. In [19], by using nine physicochemical properties of amino acids and PCA, authors proposed a novel graphical representation of protein sequences called ADLD (Alignment Diagonal Line Diagram). Experiments performed on 16 different ND5 proteins and the 29 different spike proteins showed that their method is not only visual, intuitional, and effective in the similarity/dissimilarity analysis of protein sequences but also quite simple, since there are no high dimensional matrices required to be constructed. In some other studies different from the physicochemical properties of amino acids, authors also used codon information [20-21], pseudo-Markov transition probability vector among the 20 amino acids [22], the distributions of each kind of adjacent amino acid (AAA) within sequence [23], K-string dictionary [24], and a set of point masses representing a sequence in a 20D space [25] for alignment-free similarity analysis of protein sequences.

Each of the above studies may have several advantages when compared to one another. However, to the best of our knowledge, none of the above studies allow for a bulk comparison of protein sequences for alignment-free sequence similarity analysis. Here, different from the above studies, in this paper we propose a method which allows for not only pairwise comparison of protein sequences but also a bulk comparison of several protein sequences for similarity analysis. Experiments showed that, bulk comparison of protein sequences with the proposed approach is much faster than the pairwise comparison of the same proteins using the same similarity measure. Moreover, there is no degradation on the performance in terms of the accuracy of the obtained results. It is also shown that, the proposed method achieves a clustering performance which is comparable to the state-of-the-art ClustalW method.

The rest of this paper is organized as follows: in section 2, details of the proposed method is introduced. Section 3 covers a definition of the datasets used in experiments along with the computational results obtained with the proposed and the existing methods. Finally, section 4 concludes the study.

II. METHODOLOGY

In literature, for comparison of different methods a gold-standard dataset which includes sequences of ND5 proteins from nine different species is widely utilized. In this study, we will also use this dataset along with the others for comparison of our method with the existing methods. However, let us first continue with the description of our method by using this gold-standard dataset which we believe will be more informative to the readers.

TABLE I
TWELVE MAJOR PHYSICOCHEMICAL CHARACTERISTIC VALUES OF 20 AMINO ACIDS. (P1: CHEMICAL COMPOSITION OF THE SIDE CHAIN; P2: POLAR REQUIREMENT; P3: HYDROPATHY INDEX; P4: ISOELECTRIC POINT; P5: MOLECULAR VOLUME; P6: POLARITY; P7: AROMATICITY; P8: ALIPHATICITY; P9: HYDROGENATION; P10: HYDROXYTHIOLATION; P11: PK1(-COOH); P12: PK2(-NH3+))

Amino acids	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12
A (Ala)	0	7.0	1.8	6.00	31	8.1	-0.11	0.239	0.33	-0.062	2.34	9.69
C (Cys)	2.75	4.8	2.5	5.07	55	5.5	-0.184	0.22	0.074	0.38	1.71	10.78
D (Asp)	1.38	13.0	-3.5	2.77	54	13.0	-0.285	0.171	-0.371	-0.079	2.09	9.82
E (Glu)	0.92	12.5	-3.5	3.22	83	12.3	-0.067	0.187	-0.254	-0.184	2.19	9.67
F (Phe)	0	5.0	2.8	5.48	132	5.2	0.438	0.234	0.011	0.074	1.83	9.13
G (Gly)	0.74	7.9	-0.4	5.97	3	9.0	-0.073	0.16	0.37	-0.017	2.34	9.60
H (His)	0.58	8.4	-3.2	7.59	96	10.4	0.32	0.205	-0.078	0.056	1.82	9.17
I (Ile)	0	4.9	4.5	6.02	111	5.2	0.001	0.273	0.149	-0.309	2.36	9.68
K (Lys)	0.33	10.1	-3.9	9.74	119	11.3	0.049	0.228	-0.075	-0.371	2.18	8.95
L (Leu)	0	4.9	3.8	5.98	111	4.9	-0.008	0.281	0.129	-0.264	2.36	9.60
M (Met)	0	5.3	1.9	5.74	105	5.7	-0.041	0.253	-0.092	0.077	2.28	9.21
N (Asn)	1.33	10.0	-3.5	5.41	56	11.6	-0.136	0.249	-0.233	0.166	2.02	8.80
P (Pro)	0.39	6.6	-1.6	6.30	32.5	8.0	-0.016	0.165	0.37	-0.036	1.99	10.60
Q (Gln)	0.89	8.6	-3.5	5.65	85	10.5	-0.246	0.26	-0.409	-0.025	2.17	9.13
R (Arg)	0.65	9.1	-4.5	10.76	124	10.5	0.079	0.211	-0.176	-0.167	2.17	9.04
S (Ser)	1.42	7.5	-0.8	5.68	32	9.2	-0.153	0.236	0.022	0.47	2.21	9.15
T (Thr)	0.71	6.6	-0.7	6.16	61	8.6	-0.208	0.213	0.136	0.348	2.63	10.43
V (Val)	0	5.6	4.2	5.96	84	5.9	-0.155	0.255	0.245	0.212	2.32	9.62
W (Trp)	0.13	5.2	-0.9	5.89	170	5.4	0.493	0.183	0.011	0.05	2.38	9.39
Y (Tyr)	0.20	5.4	-1.3	5.66	136	6.2	0.381	0.193	-0.138	0.22	2.20	9.11

The ND5 dataset (Table A1 of appendix) consist of sequences from nine different species; human, gorilla, pigmy chimpanzee (pchimp), common chimpanzee (cchimp), fin whale (fwhale), blue whale (bwhale), mouse, rat and opossum.

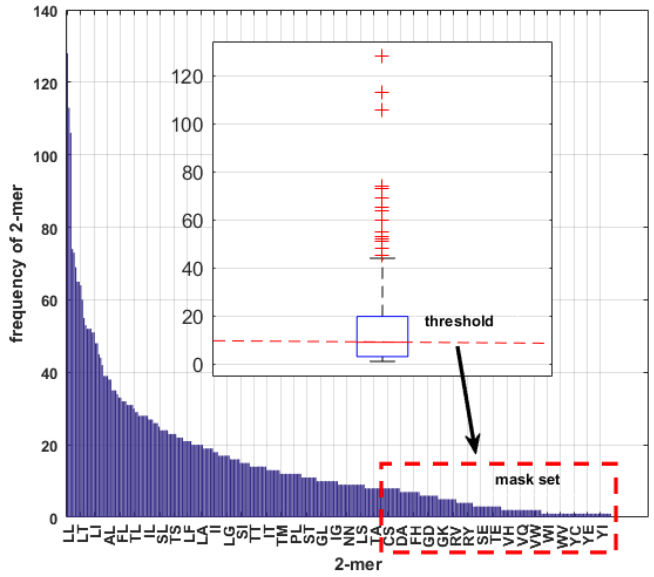

Fig.1. Frequency distribution of 2-mers obtained from the concatenated sequence S_C. 2-mers those have a frequency value below the threshold (median of the distribution) are selected to form the required set of masks.

Let $S = s_1 s_2 \dots s_n$ be a protein sequence, where n represents the sequence length. Then, we could formally describe a sequence in the ND5 dataset as $S_i \in S_{ND5}$, $i = 1,2,\dots,9$. As mentioned before, the proposed method aims for a bulk comparison of the sequences. This is achieved by using several masks representing the sequences from S_{ND5}. For the S_{ND5}, the set of masks are found in the following manner.

Let $S_C = S_1 \, o \, S_2 \, o \dots o \, S_9$ be a sequence obtained from the concatenation of the sequences from S_{ND5}, where o represents the concatenation operator. The obtained S_C includes the necessary information to form the required mask set. To create the set of masks, one should first find the all 2-mers (or dimers) exist in the sequence S_C along with their frequencies. Here, considering the frequency of each 2-mer, we select a threshold to extract the required set of masks from the set of all 2-mers found. The final mask set should ideally represent each sequence from the ND5 dataset. Therefore, selection of the threshold is of critical importance. As shown in Fig.1, the threshold is selected as the median of the frequencies found for each 2-mer. 2-mers those have a frequency below the defined threshold are selected to form the mask set. For the ND5 dataset, the threshold is found to be 9 and the total number of 2-mers in the mask set M is found to be 148, some of which are explicitly shown in Fig.1.

Having formed the required set of masks, we need to define a measure to compute the affinity of sequence $S_i \in S_{ND5}$, $i = 1, 2, \dots, 9$ to mask $m_j \in M$, $j = 1, 2, \dots, 148$. For this purpose, we utilized the twelve physicochemical properties of amino acids listed in Table 1 [16]. By using the PCA, Table 1 is first projected into a two-dimensional space (Table 2) in which each amino acid could be represented as a two-dimensional vector. A vectoral representation of each amino acid is shown in Fig.2A. Now, a given 2-mer, i.e., "CS", could be represented by concatenating two vectors head-to-tail. As shown in Fig.2B, once we concatenate the two vectors, we will have three different points in two-dimensional space. Because the point $p_0 = (0,0)$ is common for the all 2-

mers, we could simply ignore it. Thus, there remains two different points (p_1, p_2) to represent a 2-mer in two-dimensional space. These two points are used to form a four-dimensional feature vector $\vec{f} = [p_1, p_2]$ for a given 2-mer. Considering the 2-mer "CS" the resulting feature vector is $\vec{f}_{CS} = [-28.93, 5.51, -80.96, 5.87]$ as shown in Fig.2B. Each mask similarly could easily be represented as four-dimensional feature vectors $\vec{f}_{m_i} \in R^4,\ j = 1, 2, \ldots, 148$.

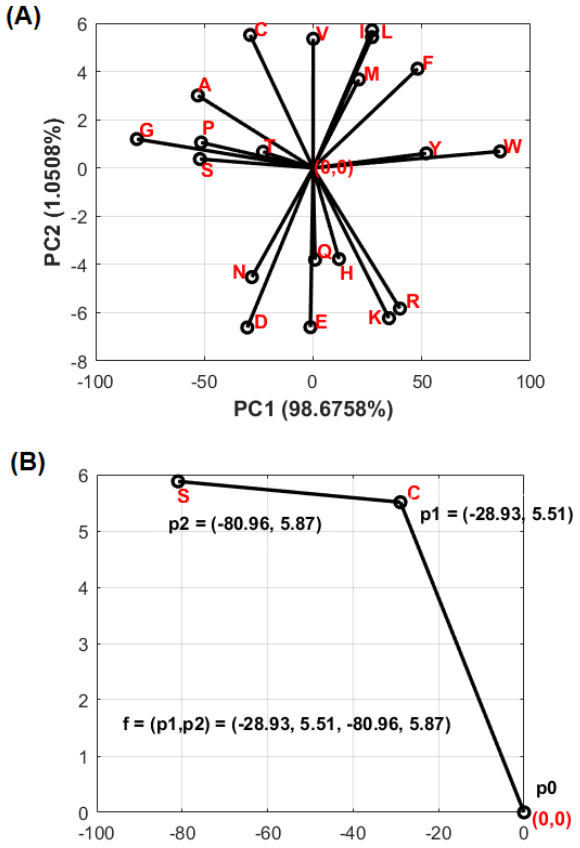

Fig.2. (A) Physicochemical properties of amino acids given in Table 1 are projected into two-dimensional space with the help of PCA. (B) A given 2-mer i.e. "CS" could be represented by a feature vector f ⃗ which is formed by two points (p_1,p_2) obtained after concatenating vectors for amino acids "C" and "S" head-to-tail.

To compute the affinity of a sequence to a mask, we used a sliding window approach. In this approach, a 2-length window is slid through the sequence $S_i,\ i = 1, 2, \ldots, 9$ and the affinity of a sequence to a mask is computed in the following manner. First, sequences of each windowed 2-mer $S^i_{W_1}, S^i_{W_2}, \ldots, S^i_{W_n}$ is found. Here, n represents the total number of windowed 2-mers obtained from S_i after sliding operation. Next, corresponding feature vectors of each windowed 2-mer are found as shown in Fig.2B. Feature vectors for each windowed 2-mer is represented as $\vec{f}^i_{W_1}, \vec{f}^i_{W_2}, \ldots, \vec{f}^i_{W_n}$. Euclidian distances between each feature vector of the windowed 2-mers and the feature vector of a mask m_j, which is $\vec{f}_{m_j}$, are then computed as in Equation (1).

$$d^i_k = dist\left(\vec{f}^i_{W_k}, \vec{f}_{m_j}\right), \quad i = 1, 2, \ldots, 9, \quad k = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, 148 \tag{1}$$

$dist(\vec{f}^i_{W_k}, \vec{f}_{m_i})$ represents the Euclidian distance from $\vec{f}^i_{W_k}$ to $\vec{f}_{m_i}$. Euclidian distances obtained between the windowed 2-mers of sequence S_i and mask pairs are then stored in a vector $\vec{d}_i = (d^i_1, d^i_2, \ldots, d^i_n)$. Affinity of a sequence S_i to mask m_j is then computed as in Equation (2).

$$a_{ij} = \min(\vec{d}_i) \tag{2}$$

Computed affinity values $a_{ij}, i = 1, 2, \ldots, 9,\ j = 1, 2, \ldots, 148$ could be represented as an NxM (for ND5 dataset $N = 9$ and $M = 148$) affinity matrix A. This matrix is then used to compute the similarity between sequences.

TABLE II
FIRST TWO PRINCIPAL COMPONENTS OF 20 AMINO ACIDS OBTAINED AFTER PCA PROJECTION OF 12 MAJOR PHYSICOCHEMICAL CHARACTERISTIC VALUES GIVEN IN TABLE I

Amino acids	PC1 (98.68%)	PC2 (1.05%)
A (Ala)	-52.9779	3.0042
C (Cys)	-28.9349	5.5100
D (Asp)	-30.2566	-6.6135
E (Glu)	-1.2373	-6.6004
F (Phe)	48.0692	4.1135
G (Gly)	-81.0102	1.1981
H (His)	11.9222	-3.7753
I (Ile)	27.0888	5.7064
K (Lys)	34.8912	-6.2279
L (Leu)	27.0930	5.4234
M (Met)	21.0652	3.6697
N (Asn)	-28.1420	-4.5234
P (Pro)	-51.4851	1.0570
Q (Gln)	0.8943	-3.8033
R (Arg)	39.9295	-5.8285
S (Ser)	-52.0304	0.3699
T (Thr)	-23.0132	0.6717
V (Val)	0.0729	5.3547
W (Trp)	86.0342	0.6879
Y (Tyr)	52.0269	0.6058

III. COMPUTATIONAL RESULTS

A. *Datasets used in the experiments*

In this study, along with the ND5 dataset (Table A1 of appendix), three other datasets are used to evaluate the performance of the proposed method. The second dataset consists of thirty-five Coronavirus Spike Proteins which were derived from the NCBI. The information and accession numbers [20] of proteins are listed in Table A2 of appendix. The third dataset consists of twenty-four transferrin and lactoferrin proteins from fish, amphibians and mammals of twenty-four vertebrates. Taxonomic information and accession numbers of these proteins are provided in Table A3 of appendix [20]. The fourth dataset consists of twenty-seven antifreeze protein sequences (AFPs) from spruce budworm (Choristoneura fumiferana, CF), yellow mealworm (Tenebrio

molitor, TM), Hypogastrura harveyi (HH), Dorcus curvidens binodulosus (DCB), Microdera dzhungarica punctipennis (MDP) and Dendroides canadensis (DC) for which the detailed information could be found in [26].

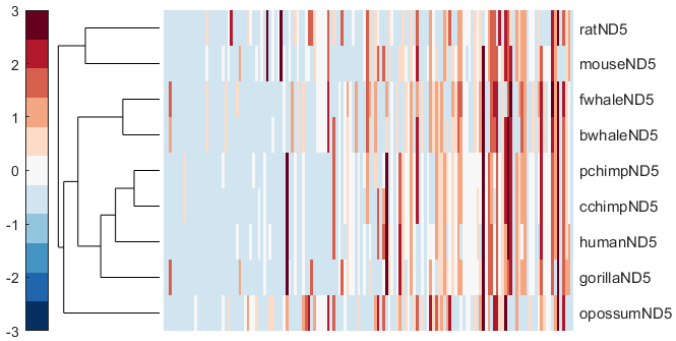

Fig.3. ND5 dataset clustered with the proposed method correctly clusters proteins into their correct groups.

B. *Similarity analysis of ND5 proteins*

Let A_{ND5} be the affinity matrix found for the ND5 proteins. From the previous section we know that A_{ND5} is a $NxM = 9x148$ matrix where N represents the number of sequences and M represents the total number of masks. A_{ND5} could directly be used to cluster the sequences for which the clustering result is shown in Fig.3. From Fig.3 and Fig.4, it is shown that, clusters found by the proposed method is in good agreement with the clusters found by the ClustalW.

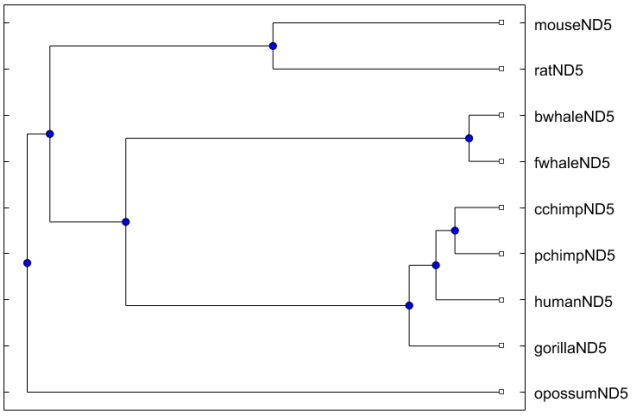

Fig.4. ClustalW result for the ND5 dataset.

To measure the performance of the proposed method, we also provide the correlation coefficients calculated between the distance matrix found by the proposed method and the one found by the ClustalW. The distance matrix for the proposed method is found as in Equation (3).

$$D_{ij} = dist\left(A^i_{ND5}, A^j_{ND5}\right) \quad i = 1, 2, \ldots, 9, \quad j = 1, 2, \ldots, 9 \tag{3}$$

The calculated distance matrix D is shown in Table 3 and the distance matrix found by the ClustalW is provided in Table 4 [20]. In Table 5, Pearson's correlation coefficients calculated between Table 3 and Table 4 are provided along with some other previously published methods. From Table 5, it can be shown that the proposed method is quite competitive and provides relatively higher correlation coefficients when compared to other methods.

C. *Pairwise comparison of protein sequences with the proposed method*

Although the proposed method mainly proposed for bulk comparison of several protein sequences, it also allows for a pairwise comparison of two protein sequences. To achieve this, all we need to do is to concatenate two sequences $S_C = S_1 \, o \, S_2$ that we want to compare. Once we obtained the concatenated sequence S_C, the same procedure given in section 2 is followed. Thereby, one could compute the similarity of two proteins by simply calculating the distances between the affinity vectors $\vec{a}_1$ and $\vec{a}_2$ found for each sequence.

In Fig.5, correlation coefficients obtained by the pairwise comparison are compared to those obtained with the bulk comparison method. From Fig.5, it is shown that both the pairwise and bulk comparisons with the proposed method provide quite similar results. However, bulk comparison of the proteins from ND5 dataset is much faster (21.69 sec.) when compared to the pairwise comparison of all sequences (305.64 sec.) from the same dataset.

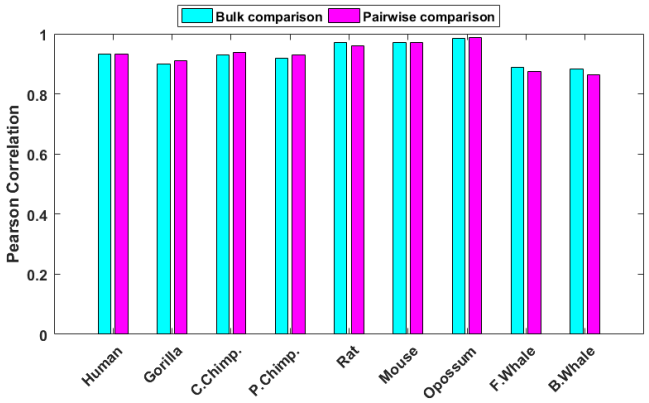

Fig.5. Bulk comparison and pairwise comparison of the sequences from ND5 dataset with the proposed method provides similar results when they are compared to ClustalW.

D. *Similarity analysis of coronavirus spike proteins*

Coronaviruses are important human and animal pathogens which are associated with upper respiratory tract infections in adults and probably also play a role in severe respiratory infections in both children and adults. Coronaviruses are traditionally classified into three groups, where the first group and the second group come from mammalian, and the third group comes from poultry (chicken and turkey). Apart from these three groups, SARS-CoV forming a fourth group, for which the phylogenetic position and origins remain a matter of some controversy. In this study, we utilized the proposed method to classify SARS-CoV spike proteins along with the proteins from other three groups.

In Fig.6, clustering result with the proposed method is shown. From this figure, it is shown that all the four groups of proteins successfully separated by the proposed method. The SARS-CoV spike proteins (group IV) remain closer to the group II. Previous studies [32, 33] also showed that, the closest relatives of SARS-CoVs are the murine, bovine and human coronaviruses from group II which is consistent with the obtained results. On the other hand, ClustalW also correctly separates the proteins into their correct groups (Fig.7).

TABLE III
DISTANCE MATRIX FOUND BY THE PROPOSED METHOD

	Human	Gorilla	C.Chimp.	P.Chimp.	Rat	Mouse	Opossum	F.Whale	B.Whale
Human	0	40.42	29.95	34.30	79.38	71.52	63.52	59.04	59.19
Gorilla		0	44.23	41.94	74.88	72.51	69.43	57.81	60.47
C.Chimp.			0	18.83	78.52	69.27	65.05	58.98	54.91
P.Chimp.				0	76.50	69.23	64.91	55.15	50.78
Rat					0	53.42	71.99	71.68	70.73
Mouse						0	68.74	64.30	63.56
Opossum							0	61.87	59.86
F.Whale								0	24.80
B.Whale									0

TABLE IV
DISTANCE MATRIX FOUND BY THE CLUSTALW [20]

	Human	Gorilla	C.Chimp.	P.Chimp.	Rat	Mouse	Opossum	F.Whale	B.Whale
Human	0	0.104	0.067	0.069	0.456	0.443	0.464	0.375	0.377
Gorilla		0	0.096	0.093	0.469	0.453	0.494	0.39	0.387
C.Chimp.			0	0.048	0.461	0.448	0.472	0.37	0.37
P.Chimp.				0	0.453	0.443	0.459	0.368	0.368
Rat					0	0.241	0.494	0.41	0.407
Mouse						0	0.469	0.422	0.415
Opossum							0	0.486	0.486
F.Whale								0	0.034
B.Whale									0

TABLE V
CORRELATION COEFFICIENTS CALCULATED BETWEEN DISTANCE MATRICES OF SOME STATE-OF-THE-ART METHODS AND CLUSTALW FOR ND5 PROTEINS

	This work	Wu et al.[20]	Yao et al.[18]	Ellakani and Mahran [27]	Zhang et al. [15]	Mu et al. [28]	Liu et al. [29]	Wu et al. [30]	Huang and Hu [31]
Human	0.93	0.96	0.93	-0.09	0.91	0.93	0.94	0.93	0.89
Gorilla	0.90	0.93	0.88	-0.03	0.92	0.93	0.93	0.91	0.93
C.Chimp.	0.93	0.96	0.94	-0.11	0.93	0.91	0.94	0.91	0.95
P.Chimp.	0.92	0.95	0.95	-0.11	0.91	0.93	0.93	0.76	0.91
Rat	0.97	0.96	0.95	0.72	0.92	0.93	0.84	0.63	0.93
Mouse	0.97	0.96	0.98	0.75	0.87	0.97	1.00	0.66	0.86
Opossum	0.99	0.99	0.94	0.99	0.99	0.93	0.89	0.52	0.92
F.Whale	0.89	0.85	0.91	0.16	0.92	0.93	0.89	0.53	0.92
B.Whale	0.88	0.85	0.93	0.15	0.92	0.96	0.87	0.69	0.93

E. Similarity analysis of transferrin and lactoferrin proteins

Iron is an essential element for almost all living organisms as it participates in a wide variety of metabolic processes, including oxygen transport, deoxyribonucleic acid (DNA) synthesis, and electron transport [34]. Iron is transported in the blood by transferrin (TF) proteins. Lactoferrin (LF) is also an iron binding protein which is structurally similar to transferrin. Previous studies have demonstrated the phylogenetic relation between the transferrin and lactoferrin [35, 36]. In this study, we analyzed the similarity of transferrin and lactoferrin proteins with the proposed method. From Fig.8., it is shown that, lactoferrin and transferrin proteins mostly clustered into their corresponding groups. Some of the transferrin proteins (the ones from mammals) remain closer to the lactoferrin proteins which is also consistent with the results reported in [20]. For this dataset, ClustalW achieves a better clustering performance (Fig.9). However, in ClustalW clustering result again the TF proteins from mammals remain closer to the LF proteins which is in good agreement with the results found by the proposed method.

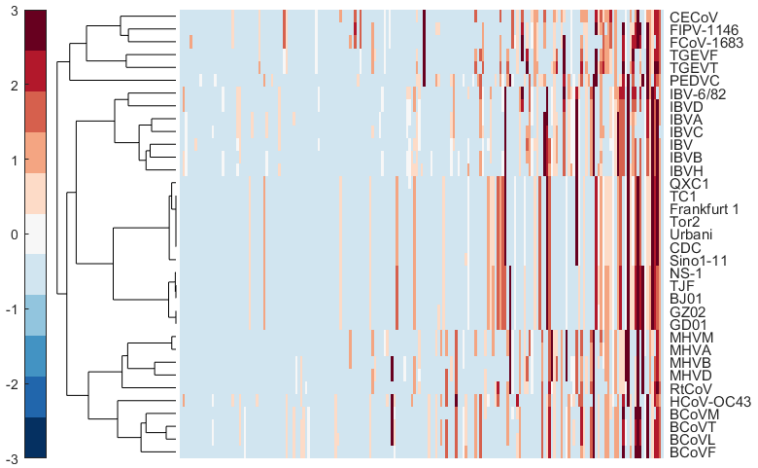

Fig.6. Coronavirus spike proteins clustered with the proposed method. All the four groups of proteins (see Table A2) successfully separated by the proposed method

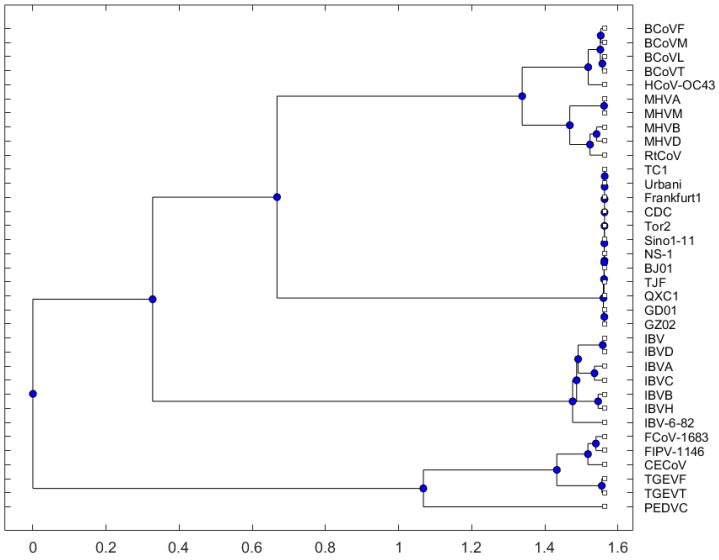

Fig.7. ClustalW clustering result for coronavirus spike proteins is in good agreement with the proposed method.

F. Similarity analysis of antifreeze proteins

Antifreeze proteins (AFPs) are biological antifreezes with unique properties, including thermal hysteresis (TH), ice recrystallization inhibition (IRI), and interaction with membranes and/or membrane proteins. These properties have been utilized in the preservation of biological samples at low temperatures [37]. These unique properties give AFPs to have potential in frozen food industry avoiding the damage in the structure of animal or vegetal foods. In this paper, by utilizing the proposed method, we analyzed the similarity of twenty-seven antifreeze proteins from spruce budworm (Choristoneura fumiferana, CF), yellow mealworm (Tenebrio molitor, TM), Hypogastrura harveyi (HH), Dorcus curvidens binodulosus (DCB), Microdera dzhungarica punctipennis (MDP) and Dendroides canadensis (DC). The proposed method (Fig.10) achieves a better clustering performance in comparison to the ClustalW. In ClustalW result (Fig.11), TM60593179 forms a separate branch which is far away from its own group. On the other hand, other proteins are clustered into their correct groups.

IV. CONCLUSION

Protein sequences similarity analysis is one of the major topics in bioinformatics. It allows researchers to find out evolutional relationships of different species. Within this context, this paper presented a new method for the similarity analysis of proteins from different species. Different from the existing studies, the proposed method not only provides a pairwise comparison of two proteins, but it also allows for a bulk comparison of multiple proteins.

The idea behind the proposed method is quite simple and effective. To achieve a bulk comparison of multiple sequences, we used several masks which are selected from the corresponding sequences that are wanted to be compared. Simply, if a sequence A is similar to a mask C and a sequence B is again similar to mask C, then sequence A and sequence B must also be similar in some degree.

The performance of the proposed method was evaluated on four different datasets. The first dataset is the gold-standard ND5 dataset. Experiments performed on this dataset showed that the proposed method performs quite well, and the

obtained results are superior to most of the other previously published methods. The second dataset consists of thirty-five coronavirus spike proteins. Experiments performed on this dataset showed that, the proposed method successfully clusters the proteins into their correct groups and the obtained results are in good agreement with the ClustalW results. The third dataset consists of twenty-four transferrin and lactoferrin proteins. For this dataset, the results obtained by ClustalW is superior to the proposed method. However, the phylogenetic tree obtained with the proposed method is mostly in good agreement with the one obtained by the ClustalW. The last dataset consists of twenty-seven antifreeze proteins from six different species. For this dataset, the proposed method exhibited a superior performance when compared to the ClustalW.

The concept behind the proposed method could also be utilized to cluster nucleotide sequences which is also a challenging problem in bioinformatics. Future studies will mostly cover the application of this concept in nucleotide sequence clustering problems. However, it could also be utilized in many other areas of pattern recognition.

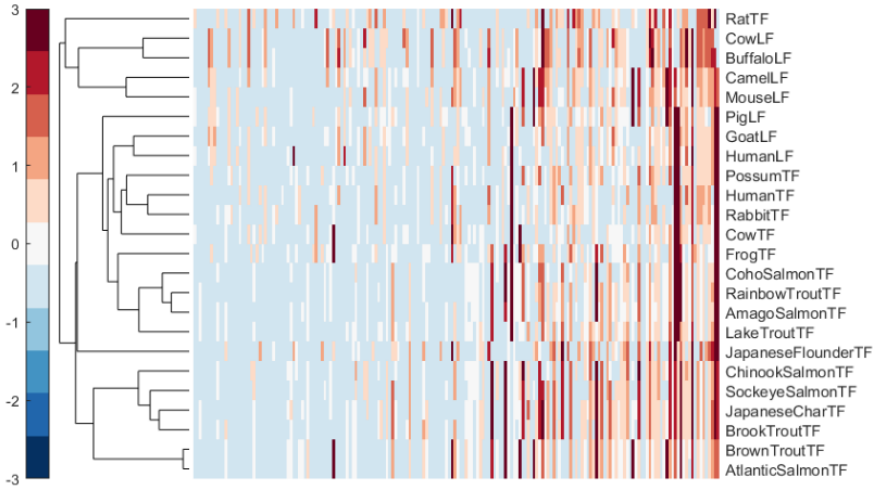

Fig.8. Transferrin and lactoferrin proteins clustered with the proposed method. Some of the LF proteins are grouped with TF proteins. However, they remain close to each other.

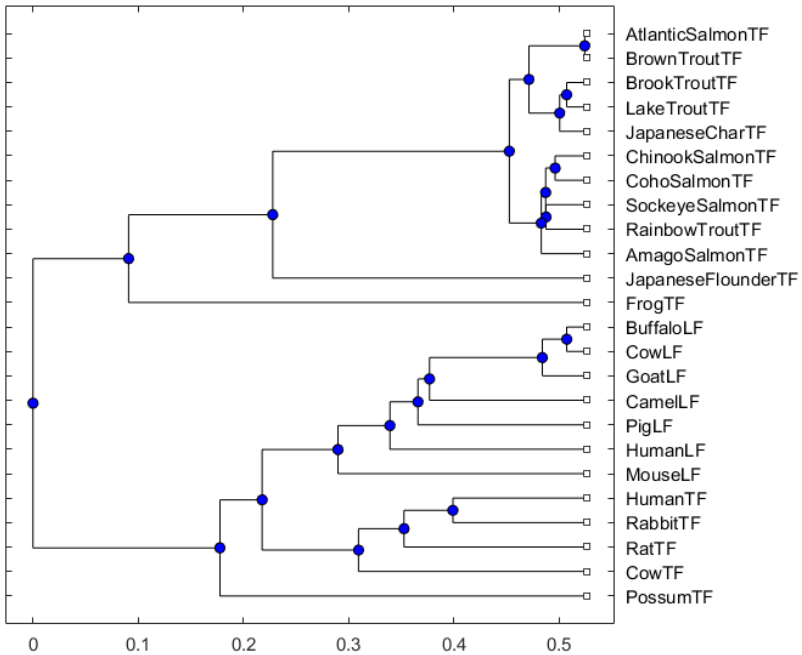

Fig.9. ClustalW clustering result for transferrin and lactoferrin proteins. For this dataset, ClustalW performs better and correctly clusters the TF and LF proteins into their corresponding groups.

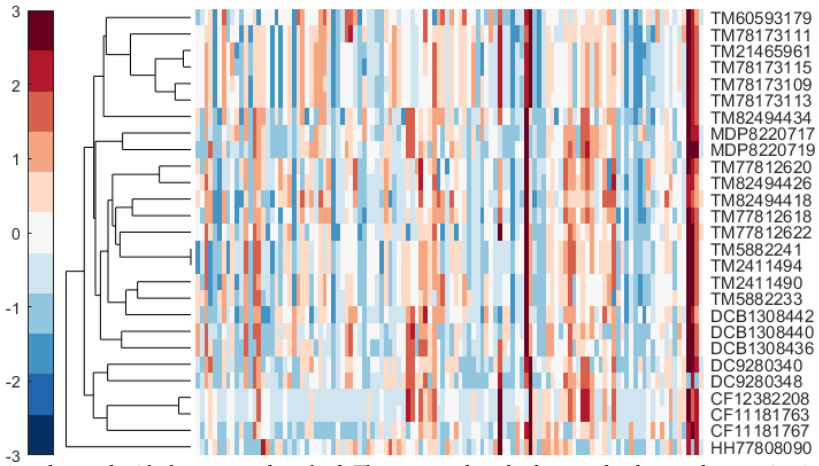

Fig.10. Antifreeze proteins clustered with the proposed method. The proposed method correctly clusters the proteins into their correct groups.

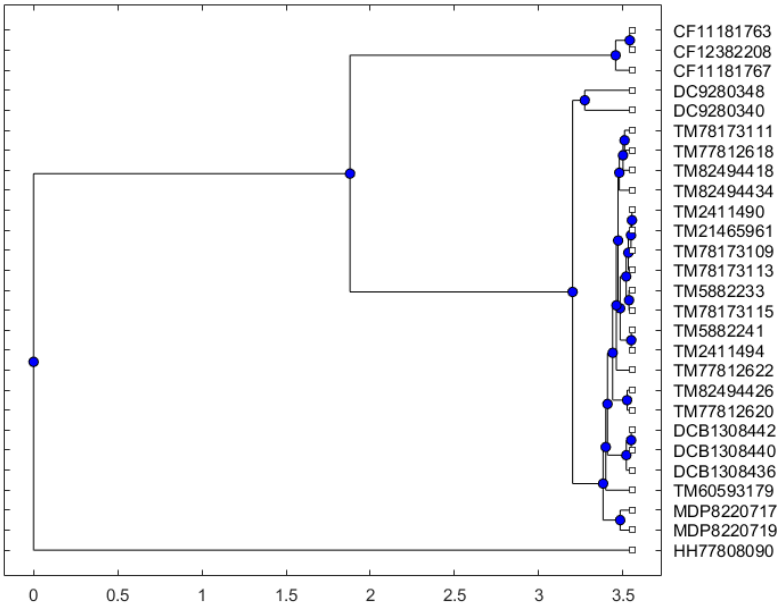

Fig.11. ClustalW clustering result for antifreeze proteins. TM60593179 forms a separate branch which is far away from its own group.

REFERENCES

[1] Z. Jiang and Z. Yanhong, "Using bioinformatics for drug target identification from the genome." American Journal of Pharmacogenomics 5.6 (2005): 387-396.

[2] M.S. Waterman, "Identification of common molecular subsequence." Mol. Biol 147 (1981): 195-197.

[3] S. F. Altschul, et al., "Basic local alignment search tool." Journal of molecular biology 215.3 (1990): 403-410.

[4] J. Yang and L. Zhang, "Run probabilities of seed-like patterns and identifying good transition seeds." Journal of Computational Biology 15.10 (2008): 1295-1313.

[5] A. Chakraborty and B. Sanghamitra, "FOGSAA: Fast optimal global sequence alignment algorithm." Scientific reports 3 (2013): 1746.

[6] O. Gotoh, "An improved algorithm for matching biological sequences." Journal of molecular biology 162.3 (1982): 705-708.

[7] X. Liu, et al., "Number of distinct sequence alignments with k-match and match sections." Computers in biology and medicine 63 (2015): 287-292.

[8] C. Li, et al., "Protein Sequence Comparison and DNA-binding Protein Identification with Generalized PseAAC and Graphical Representation." Combinatorial chemistry & high throughput screening 21.2 (2018): 100-110.

[9] L. Yu, et al., "Protein sequence comparison based on physicochemical properties and the position-feature energy matrix." Scientific Reports 7 (2017): 46237.

[10] J.D. Thompson, G.H. Desmond and J.G. Toby, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice." Nucleic acids research 22.22 (1994): 4673-4680.

[11] W. Hou, et al., "A new method to analyze protein sequence similarity using Dynamic Time Warping." Genomics 109.2 (2017): 123-130.

[12] L. He, et al. "A novel alignment-free vector method to cluster protein sequences." Journal of theoretical biology 427 (2017): 41-52.

[13] Z. Qi, and J. Meng-Zhe, "An Intuitive Graphical Method for Visualizing Protein Sequences Based on Linear Regression and Physicochemical Properties." MATCH-Communications in Mathematical and in Computer Chemistry 75.2 (2016): 463-480.

[14] C. Li, L. Xueqin and L. Yan-Xia., "Numerical characterization of protein sequences based on the generalized Chou's pseudo amino acid composition." Applied Sciences 6.12 (2016): 406.

[15] Y. Zhang, et al., "Novel numerical characterization of protein sequences based on individual amino acid and its application." BioMed research international 2015 (2015).

[16] Z. Qi, et al., "A protein mapping method based on physicochemical properties and dimension reduction." Computers in biology and medicine 57 (2015): 1-7.

[17] C. Yu, et al., "Protein map: an alignment-free sequence comparison method based on various properties of amino acids." Gene 486.1 (2011): 110-118.
[18] Y. Yao, et al., "Similarity/dissimilarity analysis of protein sequences based on a new spectrum-like graphical representation." Evolutionary Bioinformatics 10 (2014): EBO-S14713.
[19] L. Wang, P. Hui and Z. Jinhua, "ADLD: a novel graphical representation of protein sequences and its application." Computational and mathematical methods in medicine 2014 (2014).
[20] C. Wu, et al., "A novel model for protein sequence similarity analysis based on spectral radius." Journal of theoretical biology 446 (2018): 61-70.
[21] N. Jafarzadeh and A. Iranmanesh, "A new measure for pairwise comparison of protein sequences." MATCH: Communications in Mathematical and in Computer Chemistry 74 (2015): 563-574.
[22] Y. Li, et al., "An alignment-free algorithm in comparing the similarity of protein sequences based on pseudo-Markov transition probabilities among amino acids." PloS one 11.12 (2016): e0167430.
[23] H.J. Yu and H. De-Shuang, "Normalized feature vectors: a novel alignment-free sequence comparison method based on the numbers of adjacent amino acids." IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 10.2 (2013): 457-467.
[24] C. Yu, L.He. Rong and SS. Yau, "Protein sequence comparison based on K-string dictionary." Gene 529.2 (2013): 250-256.
[25] A. Czerniecka, et al., "20D-dynamic representation of protein sequences." Genomics 107.1 (2016): 16-23.
[26] Y. Zhang, "A new model of amino acids evolution, evolution index of amino acids and its application in graphical representation of protein sequences." Chemical Physics Letters 497.4-6 (2010): 223-228.
[27] A. El-Lakkani, and H. Mahran, "An efficient numerical method for protein sequences similarity analysis based on a new two-dimensional graphical representation." SAR and QSAR in Environmental Research 26.2 (2015): 125-137.
[28] Z. Mu, et al., "3D–PAF Curve: A Novel Graphical Representation of Protein Sequences for Similarity Analysis." MATCH: Communications in Mathematical and in Computer Chemistry 75.2 (2016): 447-462.
[29] Y. X. Liu, et al, "P–H curve, a graphical representation of protein sequences for similarities analysis." MATCH Communications in Mathematical and in Computer Chemistry 70.1 (2013): 451-466.
[30] ZC. Wu, X. Xuan and C. Kuo-Chen, "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 (2010): 29-34.
[31] G. Huang, and J. Hu., "Similarity/Dissimilarity Analysis of Protein Sequences by a New Graphical Representation." Current Bioinformatics 8.5 (2013): 539-544.
[32] K.V. Holmes, "SARS coronavirus: a new challenge for prevention and therapy." The Journal of clinical investigation 111.11 (2003): 1605-1609.
[33] E.J. Snijder, et al., "Unique and conserved features of genome and proteome of SARS-coronavirus, an early split-off from the coronavirus group 2 lineage." Journal of molecular biology 331.5 (2003): 991-1004.
[34] N. Abbaspour, R. Hurrell and R. Kelishadi, "Review on iron and its importance for human health." Journal of research in medical sciences: the official journal of Isfahan University of Medical Sciences 19.2 (2014): 164.
[35] M.J. Ford, "Molecular evolution of transferrin: evidence for positive selection in salmonids." Molecular biology and evolution 18.4 (2001): 639-647.
[36] G. Chang, and W. Tianming, "Phylogenetic analysis of protein sequences based on distribution of length about common substring." The protein journal 30.3 (2011): 167-172.
[37] H. Kim, et al., "Marine antifreeze proteins: structure, function, and application to cryopreservation as a potential cryoprotectant." Marine drugs 15.2 (2017): 27.

BIOGRAPHIES

BERAT DOĞAN was born in Malatya, Turkey. He received the B.S. degree in electronics engineering from Erciyes University, Turkey in 2006, M.S. degree in biomedical engineering from Istanbul Technical University, Turkey in 2009 and the Ph.D. degree in electronics engineering from Istanbul Technical University, Turkey, in 2015.

From 2008 to 2009 he was a software engineering at Nortel Networks Netas Telecommunication Inc. From 2009 to 2015 he worked as a research assistant at Department of Electronics and Communication Engineering, Istanbul Technical University, Turkey. In 2015, he started to work as an assistant professor at Department of Biomedical Engineering, Inonu University, Turkey. Between 2017-2018 he was a postdoc researcher at Department of Human Genetics, McGill University, Canada. His research interests include, bioinformatics, biomedical signal and image processing and metaheuristics.

APPENDIX

TABLE AI
INFORMATION FOR NINE ND5 PROTEINS

Number	Species	ID (NCBI)	Length
1	Human	AP_000649	603
2	Gorilla	NP_008222	603
3	Common chimpanzee	NP_008196	603
4	Pigmy chimpanzee	NP_008209	603
5	Fin whale	NP_006899	606
6	Blue whale	NP_007066	606
7	Rat	AP_004902	610
8	Mouse	NP_904338	607
9	Opossum	NP_007105	602

TABLE AII
INFORMATION OF THIRTY-FIVE CORONAVIRUS SPIKE PROTEINS

ID (NCBI)	Abbreviation	Name	Group
P10033	FIPV-1146	Feline infectious peritonitis virus strain 79-1146	I
Q66928	FCoV-1683	Feline coronavirus strain 79-1683	I
Q91AV1	PEDVC	Porcine epidemic diarrhea virus strain CV777	I
Q9DY22	TGEVT	Transmissible gastroenteritis virus strain TO14	I
P18450	TGEVF	Porcine transmissible gastroenteritis coronavirus strain FS772/70	I
P36300	CECoV	Canine enteric coronavirus strain INSAVC-1	I
Q9J3E7	MHVM	Murine hepatitis virus strain ML-10	II
Q83331	MHVB	Murine hepatitis virus strain Berkeley	II
P11224	MHVA	Murine hepatitis virus strain A59	II
O55253	MHVD	Murine hepatitis virus strain DVIM	II
Q9IKD1	RtCoV	Rat coronavirus strain 681	II
P25190	BCoVF	Bovine coronavirus strain F15	II
P15777	BCoVM	Bovine coronavirus strain Mebus	II
Q9QAR5	BCoVL	Bovine coronavirus strain LSU-94LSS-051	II
Q91A26	BCoVT	Bovine enteric coronavirus 98TXSF-110-ENT	II
P36334	HCoV-OC43	Human coronavirus strain OC43	II
Q82666	IBV	Infectious bronchitis virus	III
P05135	IBV-6/82	Avian infectious bronchitis virus strain 6/82	III
P12722	IBVD	Avian infectious bronchitis virus strain D274	III
Q64930	IBVC	Infectious bronchitis virus strain CU-T2	III
Q82624	IBVA	Infectious bronchitis virus strain Ark99	III
P11223	IBVB	Avian infectious bronchitis virus strain Beaudette	III
Q98Y27	IBVH	Infectious bronchitis virus strain H52	III
AAP41037	Tor2	SARS coronavirus Tor2	IV
AAP30030	BJ01	SARS coronavirus BJ01	IV
AAR91586	NS-1	SARS coronavirus NS-1	IV
AAP51227	GD01	SARS coronavirus GD01	IV
AAP33697	Frankfurt 1	SARS coronavirus Frankfurt 1	IV
AAP13441	Urbani	SARS coronavirus Urbani	IV
AAQ01597	TC1	SARS coronavirus Taiwan TC1	IV
AAU81608	CDC	SARS Coronavirus CDC 200301157	IV
AAS00003	GZ02	SARS coronavirus GZ02	IV
AAR86788	QXC1	SARS coronavirus ShanghaiQXC1	IV
AAR23250	Sino1-11	SARS coronavirus Sino1-11	IV
AAT76147	TJF	SARS coronavirus TJF	IV

TABLE AIII
THE CONCISE INFORMATION FOR TWENTY-FOUR TRANSFERRIN (TF) AND LACTOFERRIN (LF) PROTEIN SEQUENCES

Sequence name	**Species**	**Accession no**	**Length**
Human TF	Homo sapiens	S95936	698
Rabbit TF	Oryctolagus coniculus	X58533	695
Rat TF	Rattus norvegicus	D38380	698
Cow TF	Bos Taurus	U02564	704
Buffalo LF	Bubahts arnee	AJ005203	708
Cow LF	Bos Taurus	X57084	708
Goat LF	Copra hircus	X78902	708
Camel LF	Camehts dromedaries	AJ131674	708
Pig LF	Sus scrofa	M92089	704
Human LF	H. sapiens	NM 002343	710
Mouse LF	Mus musculus	NM 008522	707
Possum TF	Trichosurus vulpecula	AF092510	711
Frog TF	Xenopus laevis	X54530	702
Japanese flounder TF	Pctralichthys olivaceiis	D88801	685
Atlantic salmon TF	Salmo salar	L20313	690
Brown trout TF	Salmo trutta	D89091	691
Lake trout TF	Salvelimts namaycush	D89090	691
Brook trout TF	Sahelinus fontinalis	D89089	691
Japanese char TF	Sahelinus phius	D89088	691
Chinook salmon TF	Oncorhynchus tshawytscha	AH008271	677
Coho salmon TF	Oncorhynchus kisuich	D89084	691
Sockeye salmon TF	Oncorhynchus nerka	D89085	691
Rainbow trout TF	Oncorhynchus mykiss	D89083	691
Amago salmon TF	Oncorhynchus masou	D89086	691