Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.

Journal of Theoretical Biology 446 (2018) 61–70

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/jtbi

A novel model for protein sequence similarity analysis based on spectral radius

Chuanyan Wu[a], Rui Gao[a,*], Yang De Marinis[b], Yusen Zhang[c]

[a] School of Control Science and Engineering, Shandong University, Jinan 250061, China
[b] Diabetes and Endocrinology, Lund University Diabetes Centre, Malmo 20502, Sweden
[c] School of Mathematics and Statistics, Shandong University, Weihai 264209, China

ARTICLE INFO

Article history:
Received 1 September 2017
Revised 27 February 2018
Accepted 1 March 2018
Available online 7 March 2018

Keywords:
Protein sequence similarity analysis
Functional group
Protein vector
Fluctuation complexity

ABSTRACT

Advances in sequencing technologies led to rapid increase in the number and diversity of biological sequences, which facilitated development in the sequence research. In this paper, we present a new method for analyzing protein sequence similarity. We calculated the spectral radii of 20 amino acids (AAs) and put forward a novel 2-D graphical representation of protein sequences. To characterize protein sequences numerically, three groups of features were extracted and related to statistical, dynamics measurements and fluctuation complexity of the sequences. With the obtained feature vector, two models utilizing Gaussian Kernel similarity and Cosine similarity were built to measure the similarity between sequences. We applied our method to analyze the similarities/dissimilarities of four data sets. Both proposed models received consistent results with improvements when compared to that obtained by the ClustalW analysis. The novel approach we present in this study may therefore benefit protein research in medical and scientific fields.

© 2018 Elsevier Ltd. All rights reserved.

1. Introduction

With the development of sequencing techniques, the discovery of biological sequences increases fast. Effective extraction and analysis of biological information from large data base has drawn much attention in the field of bioinformatics. Sequence similarity and evolution relationship analysis in order to get the function of unknown sequences (Louie et al., 2009) may shed light on identification of potential drug targets and to gain insights on underlying molecular mechanisms of diseases (Jiang and Zhou, 2005).

For protein sequence similarity analysis, there are several commonly applied methods, which can be divided into two groups: alignment-based methods (Chakraborty and Bandyopadhyay, 2012; Gotoh, 1982; Liu et al., 2015) and alignment-free methods. In the alignment-based methods, a sequence alignment scoring matrix and gap penalty parameters are used to represent insertion, deletion or substitution of AAs in the compared sequences.

Nevertheless, due to the fact that alignment-based approaches are generally memory demanding and time consuming, a lot of alignment-free methods (Yu et al., 2017) are applied alternatively, which use numerical characterization of protein sequences by extracting invariants from sequences indirectly. With the representation of a protein sequence, there are mainly two types of alignment-free methods: (1) digital-signal-based representation and (2) graphical representation. The digital-signal-based representation encodes a single amino acid (AA) into a number so that a protein sequence is converted into a digital signal sequence, which is processed by digital signal analysis tools to extract the features of the protein sequence. For example, in a study performed by Hou et al. (2017), protein sequence was converted into numerical sequences with their physicochemical properties to achieve the power spectra by Discrete Fourier Transform (DFT). Furthermore, Dynamic Time Warping (DTW) was used to extend the spectra to the same length in order to calculate the distance between different sequences. Su and Bao (2013) proposed a method based on Discrete Wavelet Transform (DWT) to measure protein sequence similarity. The model employed only the approximation coefficients of DTW so that the feature vector was short enough to bring a great running time promotion.

Graphical representation has been widely explored in bioinformatics research (Czerniecka et al., 2016; Yao et al., 2014a). It represents a protein sequence graphically and then extracts feature vector of the graph. Various approaches on graphical representation were proposed according to the physicochemical properties of the AAs. Some methods converted an AA to a discrete point according to its physicochemical properties (Xu et al., 2014), and some methods mapped an AA to a unique value by principle

* Corresponding author.
E-mail address: gaorui@sdu.edu.cn (R. Gao).

https://doi.org/10.1016/j.jtbi.2018.03.001
0022-5193/© 2018 Elsevier Ltd. All rights reserved.

components of physicochemical properties (Wang et al., 2014). Zhang et al. (2015) put forward a 2-D graphical representation by converting each AA to a point according to the hydrophobicity and hydropathy indexes. Then the cumulative distance of every point was utilized to present the distance of the sequences. In addition, similar approaches have been proposed by several studies (Li et al., 2014; Qi and Jin, 2016). Furthermore, several graphical representations of protein sequences have applied reduced protein models (Li et al., 2009; Randić et al., 2009). Yao et al. (2014c) simplified twenty AAs into four types with preset values according to hydrophobicity. Four consecutive numbers were summed as the amplitudes of vertical axis. Thus, a protein sequence can be characterized by a 17-D vector containing the frequencies of the amplitudes. Based on the idea of cyclic order of 20 AAs, Ellakkani and Mahran (2015) selected twenty concentric evenly spaced circles divided by n radial lines into equal divisions to represent any protein sequence of length n. The mean of each two successive distances between each two successive AAs was calculated. The set of the different mean distances with its frequencies was taken as a mathematical descriptor. The graph-energy-based methods are also effective graphical representation (Sun et al., 2016). Wu et al. (2015) calculated the graph energy and Laplacian energy of 20 AAs by the codons of the AAs, and applied them to a novel 2-D graphical representation of proteins to analyze the similarity/dissimilarity of protein sequences.

Albeit previous achievements, the research on dynamic feature and non-linear feature of protein sequence is relatively few. In this study, we calculated the spectral radii of 20 AAs and applied the obtained spectral radii to a novel 2-D graphical representation. The 2-D graph was characterized mathematically by extracting three groups of features. The static features of the protein sequence included the mean of spectral radii (MSR), distribution of spectral radii (DSR) and distribution of functional groups (DFG). The dynamics features of the protein sequence included distribution of spectral radii transitions (DSRT), distribution of functional groups transitions (DFGT). The non-linear feature was fluctuation complexity (C_f). With the mathematical characterization, two models adopting Gaussian Kernel similarity and Cosine similarity were built to analyze the similarities among nine NADH5 (ND5), thirty-five Coronavirus Spike Proteins (CoVPs), twenty-four transferrin proteins (TFs) and twenty-seven antifreeze protein sequences (AFPs). Our results were consistent with improvements when compared to that achieved by the ClustalW. The simulations showed that the graphical representation represented the sequence visually and comprehensively, and the proposed method was effective for protein sequence similarity analysis.

2. Materials and methods

2.1. Spectral radius of graph

For a graph, there are many measurements, such as graph energy (Qi et al., 2011; Wu et al., 2015; Yu et al., 2017), Laplacian graph ene.g. (Wu et al., 2015), spectral radius, point centrality (Zhou et al., 2016), average degree of nodes (Zhou et al., 2016) etc. They have been applied in sequence analysis successfully. For example, a weighted directed graph was set up for each DNA sequence. The adjacency matrix of the directed graph was used to induce a representative vector for a DNA sequence (Qi et al., 2011).

The spectral radius of a graph is the largest eigenvalue of the adjacency matrix of the graph. It has been widely used as a metric in classification. In the model called PROTNN, a rich set of structural and topological attributes including spectral radius were extracted to classify protein structures (Dhifli and Diallo, 2016). A metric called the spectral radius ratio was defined as the ratio of the spectral radius to the average node degree in order to measure the variation in node degree for complex network graphs (Meghanathan, 2014). All the previous researches indicated that spectral radius was an effective metric. Thus, the spectral radius was adopted to model the protein sequence.

$G = (V, E)$ was set to be a graph possessing n vertices and m edges, with the set of vertices $V = \{v_1, v_2, \ldots, v_n\}$ and the edges set $E = \{e_1, e_2, \ldots, e_m\}$. An adjacency matrix $A(G) = (a_{i,j}) \in R^{m\times n}$ of G was defined, where

$$a_{i,j} = \begin{cases} 1, & if\ (v_i, v_j) \in E, \\ 0, & if\ (v_i, v_j) \notin E. \end{cases} \tag{1}$$

$\lambda_1, \lambda_2, \ldots, \lambda_n$ was set to be the eigenvalues of adjacency matrix $A(G)$. Spectral radius $\rho(A)$ (Yu et al., 2004) was defined as

$$\rho(A) = max\{\lambda_1, \lambda_2, \ldots, \lambda_n\}. \tag{2}$$

2.2. The spectral radii of 20 AAs

The method to get the spectral radii of 20 AAs is briefly described as the following.

- Our method is based on the graphs of 20 AAs introduced by Wu et al. (2015). Four nucleotides i.e., A,G,C,T were mapped to four unit vectors with different directions, respectively. An AA was mapped to a graph by a walking method to connect the codons of the specified AA. The walking method is described briefly as follows. The walker began to walk from (0, 0). If the following nucleotide in the codons was the same with the current one, the walker would not change the walking direction and only add the value of edges by one. Otherwise, the walker would change the walking direction according to the direction of the nucleotide. Thus, the nucleotides were connected together to form a graph. Then the graph was transformed to an undirected graph. In the graph, the value of the edge denoted the walking times. The graphs of 20 AAs are shown in Figure S1 in supplementary materials.
- After getting the graphs of 20 AAs, the adjacency matrices of the graphs were established to calculate the eigenvalues. Thus, the spectral radii of 20 AAs were calculated according to (2) and the result is shown in Table 1.

2.3. The 2-D graphical representation of protein sequence

Suppose a protein sequence is denoted by $S = S_1S_2 \ldots S_N$, where S_i denotes the *ith* AA along the protein sequence. In order to represent the order and types of AAs in the sequence, the x-coordinate value includes two parts. One part is the ordinal number of the ith AA appearing in the sequence. The other part describes the types of the 20 AAs. Type numbers of the AAs were defined according to the alphabetic order to distinguish the 20 AAs with definition shown in Table 1.

For graphical representation of the ith AA, we defined

$$\begin{cases} x_i = i + 0.5 * T_i/20, \\ y_i = sr_i, \end{cases} \tag{3}$$

where T_i denoted the type of the ith AA defined in Table 1, sr_i denoted the spectral radius of the ith AA along the sequence.

Our method was performed on two short fragments of *Saccharomyces cerevisiae* to demonstrate the proposed graphical representation.

Protein I (PI) and protein II (PII) sequences are
PI:WTFESRNDPAKDPVILWLNGGPGCSSLTGL,
PII:WFFESRNDPANDPI ILWLNGGPGCSSFTGL.

The 2-D graphical representation of PI and PII was shown in Fig. 1. Fig. 1 showed that there were four different points in the two sequences intuitively, which was consistent with the result of manual alignment.

Table 1
The spectral radii and type numbers of 20 AAs.

AA	SR	Type	AA	SR	Type	AA	SR	Type	AA	SR	Type
A	6.04	1	G	8.13	6	M	2.24	11	S	5.19	16
C	2.5	2	H	2.93	7	N	3.26	12	T	3.15	17
D	3.02	3	I	5.15	8	P	8.13	13	V	3.61	18
E	2.46	4	K	5.1	9	Q	2.46	14	W	2.24	19
F	3.26	5	L	7.42	10	R	8.32	15	Y	3.26	20

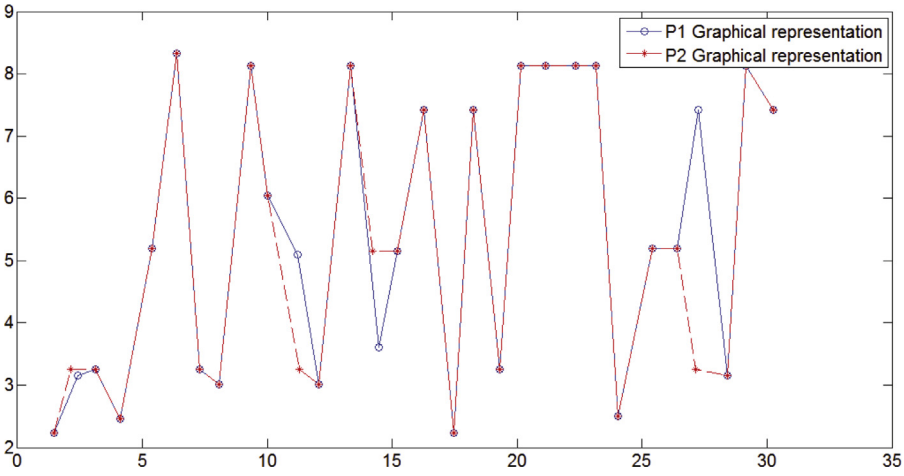


Fig. 1. The 2-D graphical representation of protein I (PI) and protein II (PII) sequences. PI and PII sequences are PI:WTFESRNDPAKDPVILWLNGGPGCSSLTGL, PII:WFFESRNDPANDPIILWLNGGPGCSSFTGL.

2.4. Numerical characterization of protein sequence

2.4.1. Mean spectral radius (MSR)

To extract the mean value of spectral radii $\overline{sr}$ of the protein sequence, we defined

$$\overline{sr} = \frac{1}{N}\sum_{i=1}^{N} sr_i, \tag{4}$$

where sr_i denoted the spectral radius of the ith AA in the protein sequence.

2.4.2. Distribution of spectral radii (DSR)

According to the size and clustering of the spectral radii of 20 AAs, the values of spectral radii were classified into eight intervals as Interval 1 = $\{2 \le sr \le 2.5\}$, Interval 2 = $\{2.5 < sr \le 3.15\}$, Interval 3 = $\{3.15 < sr \le 5\}$, Interval 4 = $\{5 < sr \le 6\}$, Interval 5 = $\{6 < sr \le 7\}$, Interval 6 = $\{7 < sr \le 8\}$, Interval 7 = $\{8 < sr \le 8.3\}$ and Interval 8 = $\{8.3 < sr \le 9\}$. Let $I_i, \quad i = 1, 2, \ldots, 8$, which represented the ith interval. To calculate distribution F of spectral radius intervals, we defined

$$F = [f_{I_1}, f_{I_2}, f_{I_3}, f_{I_4}, f_{I_5}, f_{I_6}, f_{I_7} f_{I_8}], \tag{5}$$

where $f_{I_j} = \frac{1}{N}\sum_{i=1}^{N} h_{I_j}(sr_i)$, $h_{I_j}(sr_i) = \begin{cases} 1, & \text{if } sr_i \in I_j, \\ 0, & \text{otherwise.} \end{cases}$

2.4.3. Distribution of spectral radii transitions (DSRT)

To obtain distribution of spectral radius intervals transitions in a sequence, we defined

$$A = \begin{pmatrix} a_{1,1} & \cdots & a_{1,j} & \cdots & a_{1,8} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{i,1} & \cdots & a_{i,j} & \cdots & a_{i,8} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{8,1} & \cdots & a_{8,j} & \cdots & a_{8,8} \end{pmatrix}_{8\times 8} = (a_{i,j})_{8\times 8},$$

where $a_{i,j} = \frac{1}{N-1}\sum_{l=1}^{N-1} w_{ij}(sr_l)$, $w_{ij}(sr_l) = \begin{cases} 1, & \text{if } sr_l \in I_i \text{ and } sr_{l+1} \in I_j, \\ 0, & \text{otherwise.} \end{cases}$

A was rebuild to a row vector by

$$\tilde{A} = [a_{1,1}, \ldots, a_{8,1}, a_{1,2}, \ldots, a_{8,2}, \ldots, a_{1,8}, \ldots, a_{8,8}]. \tag{6}$$

2.4.4. Distribution of functional groups (DFG)

Physicochemical properties of AAs are largely related to the side chain of AAs. Each property of AAs has its particularity, which depends on the type of the side chain the AAs possess (Hayat and Khan, 2013). By the presence of side chain chemical group, 20 AAs were classified into 10 functional groups: phenyl (F/W/Y), carboxyl (D/E), imidazole (H), primary amine (K), guanidino (R), thiol (C), sulfur (M), amido (Q/N), hydroxyl (S/T) and nonpolar (A/G/I/L/V/P) (Pugalenthi et al., 2008). Let F, D, H, K, R, C, M, Q, S and A represent each group, respectively. Thus, the sequence can be represented by $S = (s_1, s_2, \ldots, s_N), s_i \in \{F, D, H, K, R, C, M, Q, S, A\}$. In order to calculate distributions G of the ten functional groups, we defined

$$G = [g_F, g_D, g_H, g_K, g_R, g_C, g_M, g_Q, g_S, g_A], \tag{7}$$

where $g_k = \frac{1}{N}\sum_{i=1}^{N} q_k(s_i)$, $q_k(s_i) = \begin{cases} 1, & \text{if } s_i = k, \\ 0, & \text{otherwise,} \end{cases}$ $k = \{F, D, H, K, R, C, M, Q, S, A\}$.

2.4.5. Distribution of functional groups transitions (DFGT)

To get distributions of functional groups transitions, we defined

$$B = \begin{pmatrix} b_{1,1} & \cdots & b_{1,j} & \cdots & b_{1,10} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ b_{i,1} & \cdots & b_{i,j} & \cdots & b_{i,10} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ b_{10,1} & \cdots & b_{10,j} & \cdots & b_{10,10} \end{pmatrix}_{10\times 10} = (b_{i,j})_{10\times 10},$$

where $b_{i,j} = \frac{1}{N-1}\sum_{l=1}^{N-1} v_{ij}(s_l)$, $v_{ij}(s_l) = \begin{cases} 1, & \text{if } s_l = k_i \text{ and } s_{l+1} = k_j, \\ 0, & \text{otherwise.} \end{cases}$

Table 2
The distance matrix of the nine ND5 protein sequences calculated by Gaussian Kernel similarity analysis.

	Human	Gorilla	C.Chim.	P.Chim.	Rat	Mouse	Opossum	F.Whale	B.Whale
Human	0	0.48	0.35	0.36	0.87	0.84	0.89	0.75	0.77
Gorilla		0	0.48	0.46	0.86	0.81	0.89	0.77	0.81
C.Chim.			0	0.21	0.85	0.79	0.88	0.73	0.73
P.Chim.				0	0.83	0.76	0.85	0.72	0.72
Rat					0	0.64	0.82	0.85	0.85
Mouse						0	0.78	0.77	0.77
Oppossum							0	0.84	0.83
F.Whale								0	0.20
B.Whale									0

Table 3
The distance matrix of the nine ND5 protein sequences calculated by Cosine similarity analysis.

	Human	Gorilla	C.Chim.	P.Chim.	Rat	Mouse	Opossum	F.Whale	B.Whale
Human	0	0.18	0.13	0.13	0.57	0.54	0.53	0.38	0.42
Gorilla		0	0.18	0.17	0.53	0.48	0.52	0.39	0.43
C.Chim.			0	0.72	0.56	0.47	0.54	0.37	0.37
P.Chim.				0	0.53	0.46	0.49	0.38	0.37
Rat					0	0.30	0.41	0.53	0.53
Mouse						0	0.38	0.43	0.42
Oppossum							0	0.44	0.43
F.Whale								0	0.59
B.Whale									0

Table 4
The distance matrix of the nine ND5 protein sequences calculated by the ClustalW.

	Human	Gorilla	C.Chim.	P.Chim.	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.Chim.			0	0.048	0.461	0.448	0.472	0.37	0.37
P.Chim.				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 5
The correlation coefficients for nine ND5 proteins of Gaussian Kernel similarity method and some state-of-the-art methods as compared with the ClustalW method.

Species	Our Method	Yao et al. (2014c)	Ellakkani and Mahran (2015)	Zhang et al. (2015)	Mu et al. (2016)	Liu et al. (2013)	Wu et al. (2010)	Huang and Hu (2013)	Yao et al. (2014b)
Human	0.96	0.93	−0.09	0.91	0.93	0.94	0.93	0.89	0.89
Gorilla	0.93	0.88	−0.03	0.92	0.93	0.93	0.91	0.93	0.85
C.Chim.	0.96	0.94	−0.11	0.93	0.91	0.94	0.91	0.95	0.86
P.Chim.	0.96	0.95	−0.11	0.91	0.93	0.93	0.76	0.91	0.77
Rat	0.96	0.95	0.72	0.92	0.93	0.84	0.63	0.93	0.87
Mouse	0.96	0.98	0.75	0.87	0.97	1.00	0.66	0.86	0.76
Opossum	0.99	0.94	0.99	0.99	0.93	0.89	0.52	0.92	0.93
F.Whale	0.98	0.91	0.16	0.92	0.93	0.89	0.53	0.92	0.87
B.Whale	0.98	0.93	0.15	0.92	0.96	0.87	0.69	0.93	0.90

B was rebuild to a row vector as

$$\tilde{B} = [b1, 1, \ldots, b_{10,1}, \ldots, b_{1,10}, \ldots, b_{10,10}]. \quad (8)$$

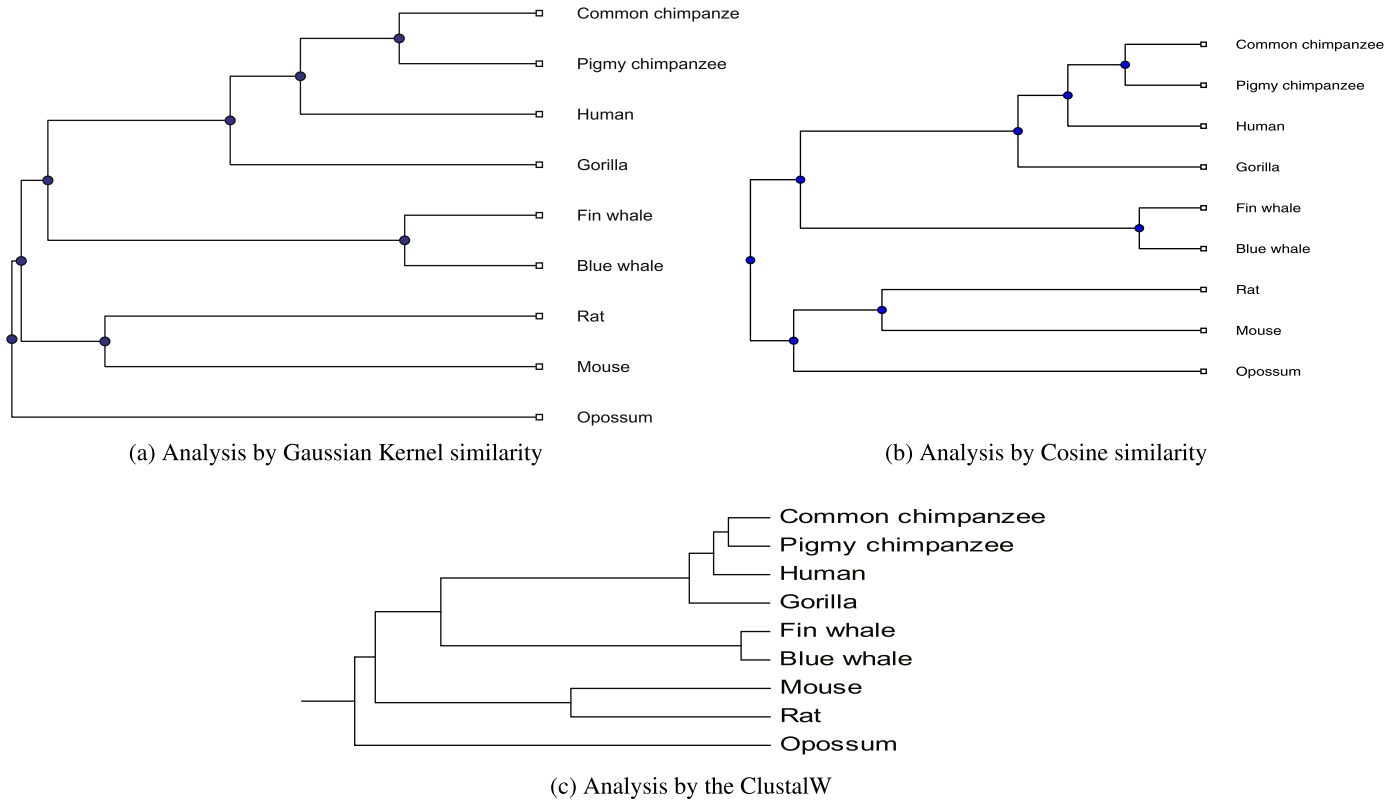
2.4.6. Fluctuation complexity (C_f)

Fluctuation complexity (Grassberger, 1986) can be applied for classification and has been widely used to describe symbol sequences in information sciences (Parrott, 2010), which was defined by Bates and Shepard (Grassberger, 1986).

It is well known that the function of proteins varies according to the type and order of AA residues. Fluctuation complexity considers both the probability of single AA and the transition probability. It reflects the fluctuation in net information gain of the sequence. Thus, fluctuation complexity was adopted to characterize the protein sequence. Fluctuation complexity was defined as

$$C_f = \sum_{i,j=1}^{L} P_{ij} * \left(\log \frac{P_i}{P_j}\right)^2, \quad (9)$$

where L denoted the number of states existing in the sequence which was equal to the type number of AAs as 20 in this paper, P_i calculated by (10) denoted the probability of the ith state in a sequence and P_{ij} calculated by (11) denoted the transition probability of state i followed by the state j in a sequence. P_i and



(a) Analysis by Gaussian Kernel similarity

(b) Analysis by Cosine similarity

(c) Analysis by the ClustalW

Fig. 2. Phylogenetic trees of the nine ND5 constructed by Gaussian Kernel similarity (a), Cosine similarity (b) and the ClustalW (c).

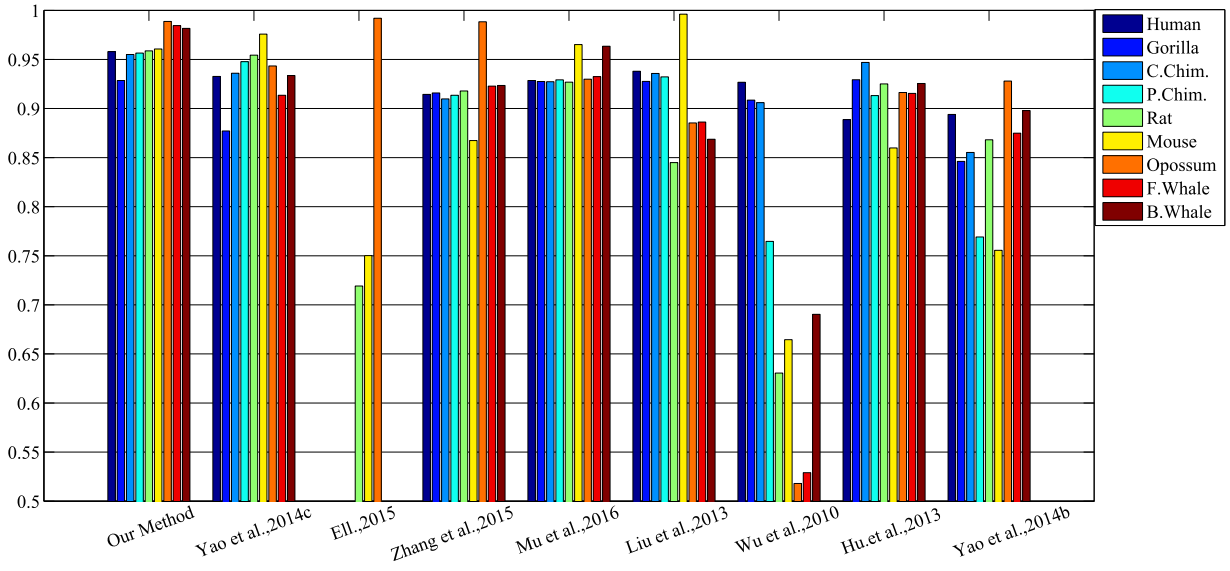


Fig. 3. The correlation coefficients for nine ND5 proteins of Gaussian Kernel similarity method and some state-of-the-art methods as compared with the ClustalW method.

P_{ij} were defined as

$$P_i = \frac{C_i}{N}, i = 1, 2, \ldots, 20, \tag{10}$$

$$P_{ij} = \frac{C_{ij}}{N-1}, i, j = 1, 2, \ldots, 20, \tag{11}$$

where C_i denoted the number of the ith AA in the sequence, C_{ij} denoted the number of the jth AA following the ith AA in the sequence which was calculated by

$$C_{ij} = \sum_{l=1}^{N-1} m_{ij}(s_l), \quad m_{ij}(s_l) = \begin{cases} 1, & \text{if } s_l = AA_i \text{ and } s_{l+1} = AA_j, \\ 0, & \text{otherwise.} \end{cases}$$

2.4.7. The numerical feature vector

The whole numerical feature vector (fv) of the protein sequence was constructed as

$$fv = [\overline{sr}, F, \tilde{A}, G, \tilde{B}, C_f], \tag{12}$$

where $\overline{sr}, F, \tilde{A}, G, \tilde{B}$ and C_f were calculated by 4–(9), respectively.

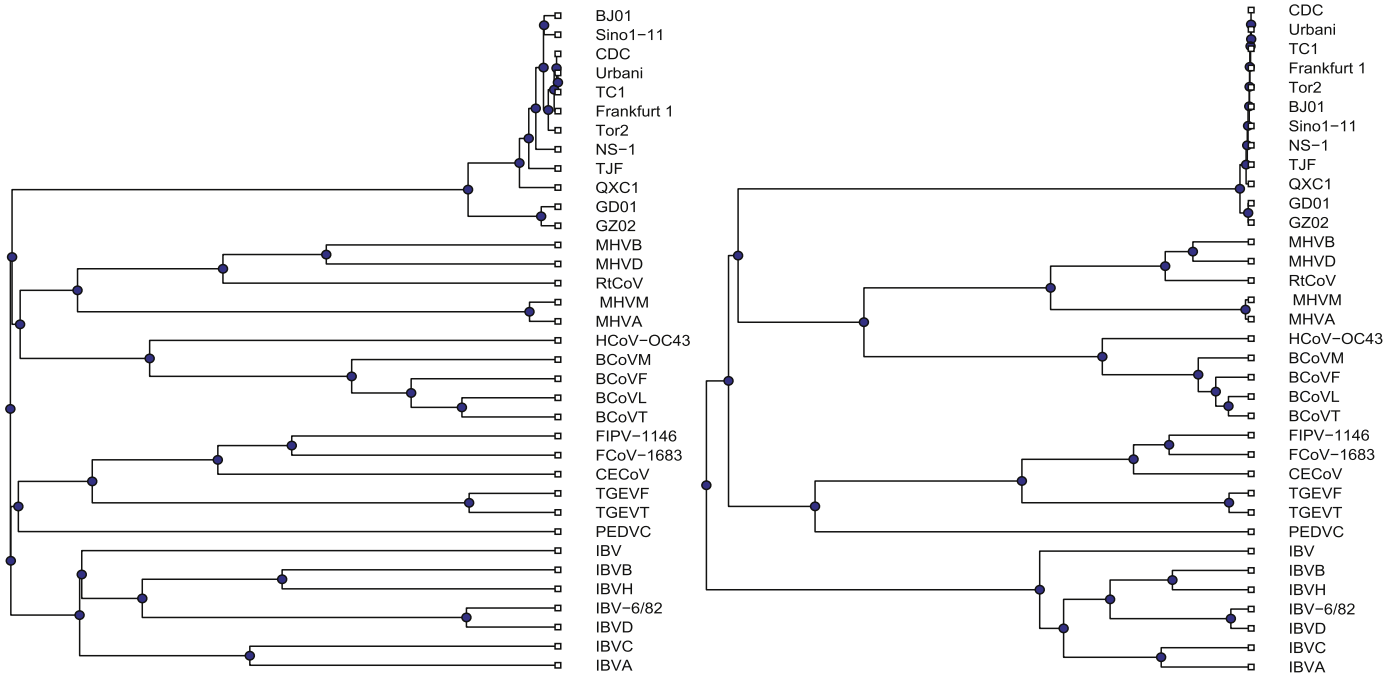


Fig. 4. Phylogenetic trees of thirty-five CoVPs constructed by Gaussian Kernel similarity (a) and Cosine similarity (b).

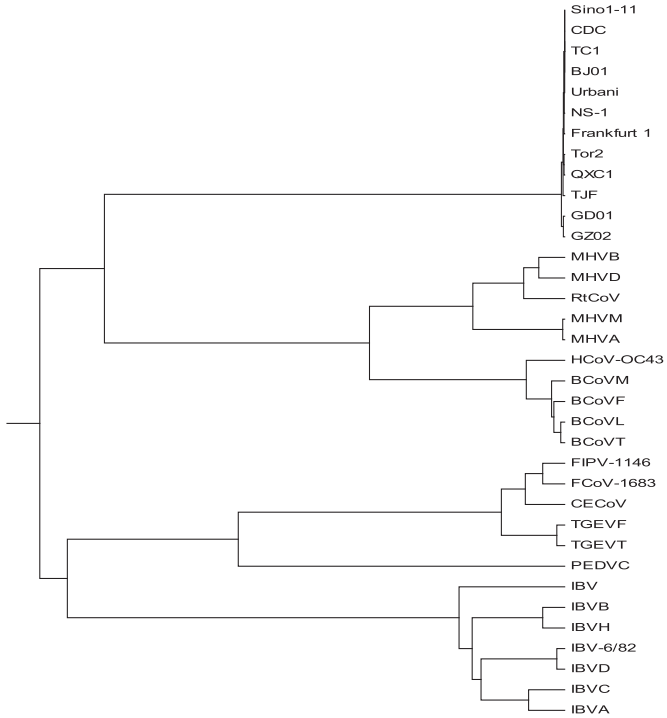


Fig. 5. Phylogenetic tree of thirty-five CoVPs achieved by the ClustalW.

2.5. The models of similarity/dissimilarity analysis

Gaussian Kernel similarity can reflect the degree of the tested point belonged to the cluster with the given centroid and adjustable bandwidth. Cosine similarity measures the direction similarity of two vectors. Thus, we adopted the two popular measurements of distance between two vectors to reflect the similarity/dissimilarity of two protein sequences.

For two protein sequences $P1$ and $P2$, the corresponding feature vectors were $\vec{s} = (f_1^1, f_2^1, \ldots, f_n^1)$ and $\vec{t} = (f_1^2, f_2^2, \ldots, f_n^2)$, where f_i^j denoted the ith feature in the jth protein, $i = 1, 2, \ldots, n$, $j = 1, 2$, n denoted the number of features calculated by (12). The distance based on Gaussian Kernel similarity between $\vec{s}$ and $\vec{t}$ was defined as

$$d_g(s,t) = 1 - exp\left(-\frac{\|s-t\|^2}{2\sigma^2}\right),$$

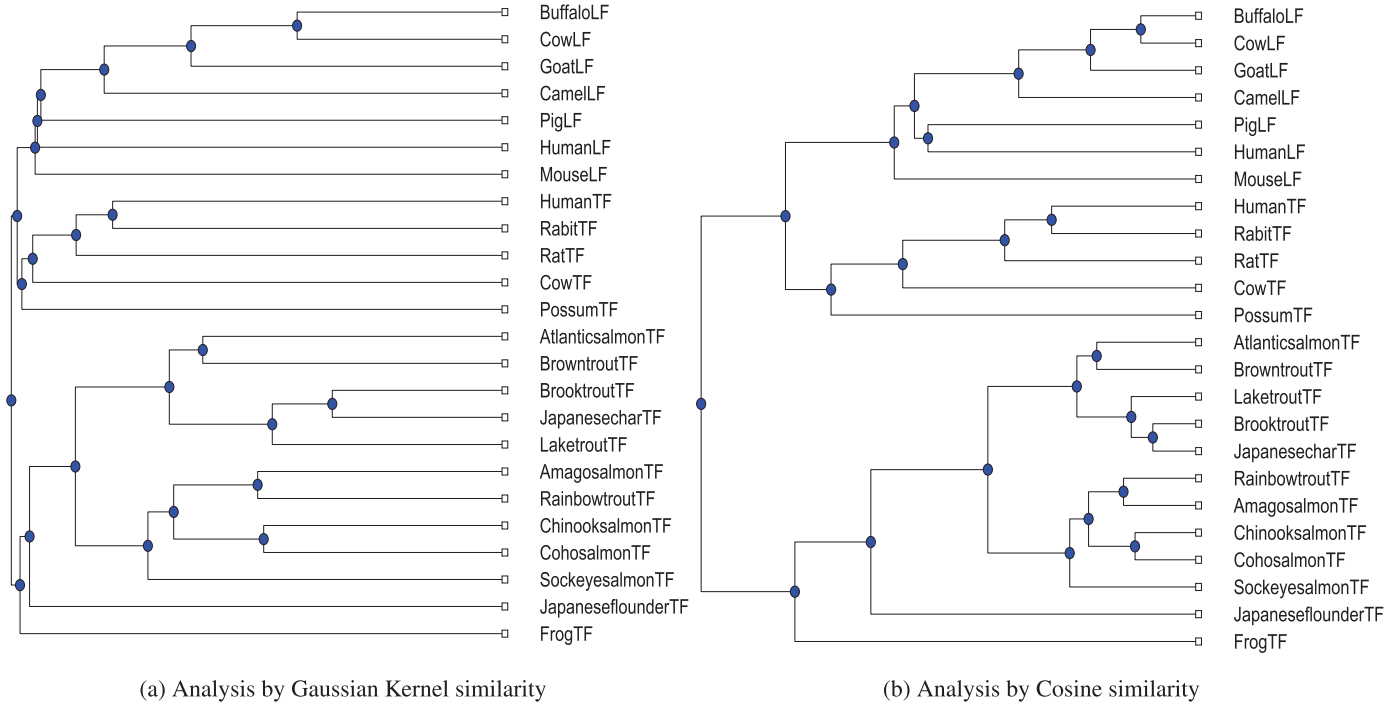
where the parameter σ controlled the bandwidth which was equal to 4 in this paper.

The second distance measurement $d_c(s, t)$ between two vectors $\vec{s}$ and $\vec{t}$ was defined to be one minus the Cosine of the included angle between $\vec{s}$ and $\vec{t}$, which was based on the assumption that two protein sequences were similar if the corresponding feature vectors had similar direction, i.e.,

$$d_c(s,t) = 1 - cos(\vec{s},\vec{t}) = 1 - \left(\frac{\vec{s}\cdot\vec{t}}{|\vec{s}||\vec{t}|}\right).$$

2.6. Materials

Four data sets were curated to evaluate the proposed method. The first data set is nine ND5. Nine ND5 contains the NADH5 of nine species including Human, Gorilla, Pigmy Chimpanzee, Common Chimpanzee, Fin Whale, Blue Whale, Rat, Mouse and Opossum from NCBI (Xu et al., 2016). The accession numbers are listed in Table S1 of supplementary materials. The second data set is thirty-five Coronavirus Spike Proteins. The proteins were derived from the NCBI. Thirty-five Coronavirus Spike Proteins are from species of order Nidovirales, family Coronaviridae and subfamily Coronavirinae. The information and accession numbers (Mu et al., 2016) are listed in Table S2 in supplementary materials. The third data set is twenty-four previously published transferrin proteins from fish, amphibians and mammals of twenty-four vertebrates, whose taxonomic information and accession numbers (Xu et al.,



(a) Analysis by Gaussian Kernel similarity (b) Analysis by Cosine similarity

Fig. 6. Phylogenetic trees of twenty-four TFs constructed by Gaussian Kernel similarity (a) and Cosine similarity (b).

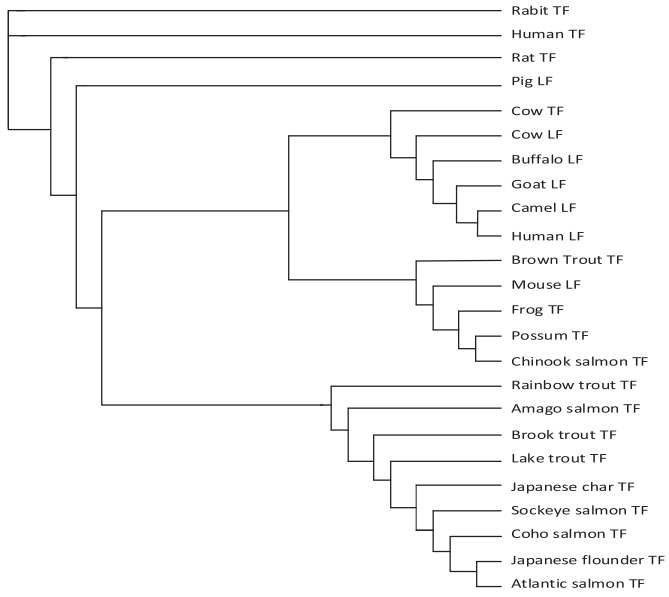


Fig. 7. Phylogenetic tree of twenty-four TFs by the ClustalW.

2016) are provided in Table S3 in supplementary materials. Furthermore, 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) (Zhang, 2010) were analyzed.

3. Results and discussion

3.1. Similarity analysis of nine ND5

Mitochondrial NADH deaminase Subunit 5 (ND5) is widely used in the analysis of phylogeny and population genetic diversity because of its high mutation rate. To illustrate the proposed similarities/dissimilarities models, the similarities of nine ND5 protein sequences across nine species were analyzed.

We calculated the distances by Gaussian Kernel similarity and Cosine similarity, (Tables 2 and 3, respectively), which were then compared to the analysis achieved by the ClustalW (Table 4) in order to validate its effectiveness. The corresponding phylogenetic trees (Fig. 2) showed that Gaussian Kernel similarity analysis is consistent with that of the ClustalW, while Cosine similarity analysis is closely consistent with that of the ClustalW.

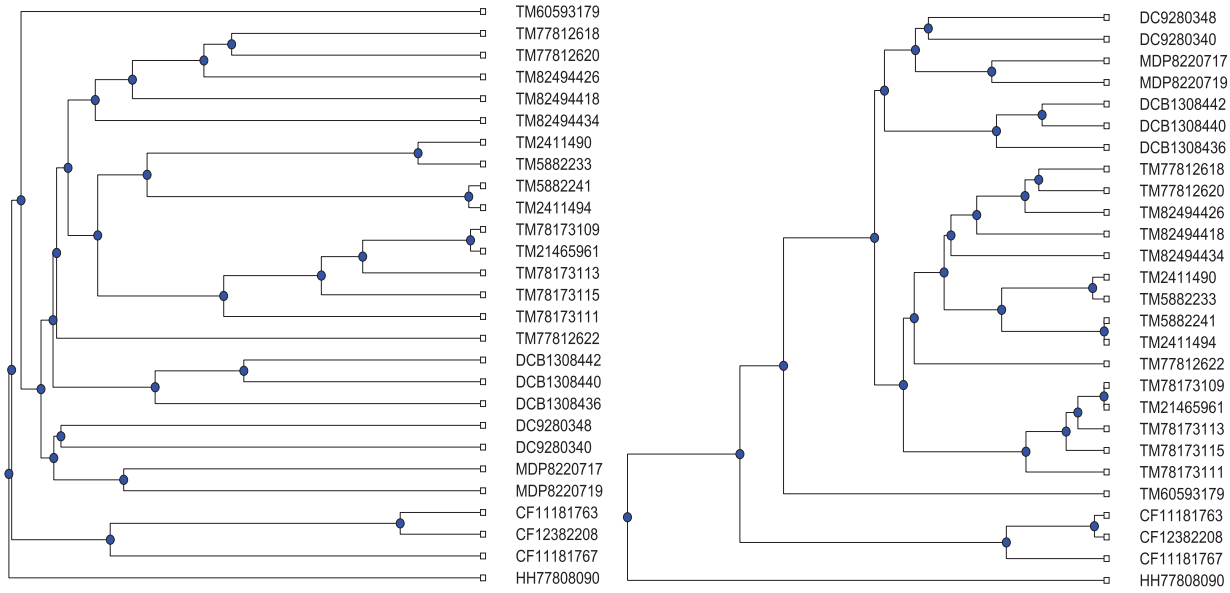
In addition, the correlation coefficients between the distance matrices calculated by Gaussian Kernel similarity and by the ClustalW were calculated. Pearson's correlation coefficient of X and Y is

$$C_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}, \tag{13}$$

where Cov denotes the covariance, σ_X denotes the standard deviation of X, σ_Y denotes the standard deviation of Y. For example, X represents the distances between Human and nine species (listed in the first row in Table 2), which were calculated by Gaussian Kernel similarity. Y represents the distances between Human and nine species (listed in the first row in Table 4), which were calculated by the ClustalW. The correlation coefficient of X and Y were calculated by (13), which is 0.96 as shown in the first column and first row of Table 5. With the same method, the correlation coefficient for other species were calculated, respectively (fist column in Table 5). Similarly, the correlation coefficients between the result by the ClustalW and the results by some state-of-the-art methods were calculated (Table 5). The results were also presented in graphical format (Fig. 3), which showed that the result by the proposed method has relatively higher correlation with that by the ClustalW than other methods. This observation further confirmed the effectiveness of the proposed method.

3.2. Similarity analysis of thirty-five Coronavirus Spike Proteins

Coronaviruses are species of virus which are associated with respiratory, intestinal, liver, and neurological diseases. Generally, coronaviruses were divided into three groups. The first group and



(a) Analysis by Gaussian Kernel similarity (b) Analysis by Cosine similarity

Fig. 8. Phylogenetic trees of 27AFPs constructed by Gaussian Kernel similarity (a) and Cosine similarity (b).

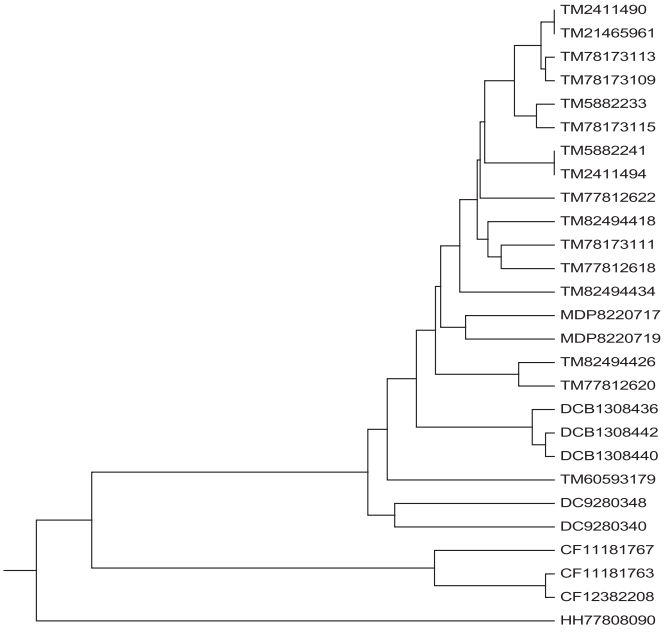


Fig. 9. Phylogenetic tree of twenty-seven AFPs by the ClustalW.

the second group come from mammalian, and the third group comes from poultry (chicken and turkey). To classify the SARS-CoV viruses and associate proteins with the virus virulence, the proposed method was utilized to analyze the coronaviruses spike proteins.

Phylogenetic trees were built by Gaussian Kernel similarity, Cosine similarity and by the ClustalW (Fig. 4a,b and Fig. 5). All the SARS-CoVs clustered in a new strain nearest to group II coronaviruses. This is consistent with the report that SARS-CoV represents a lineage that split off from the group II branch relatively late in coronavirus evolution (Snijder et al., 2003). All the groups of CoVs were also separated correctly by the analysis (Fig. 4). The results from Gaussian Kernel similarity and Cosine similarity analysis were comparable, with only little differences in SARS-CoVs classification. However, SARS-CoVs were not clearly distinguished by the ClustalW (Fig. 5). In conclusion, the results of Gaussian Kernel similarity and Cosine similarity analysis methods were in agreement with the ClustalW with improvement.

3.3. The similarity analysis of twenty-four transferrin proteins

Iron is essential for various metabolic processes such as oxygen transfer, electron transport, DNA synthesis, etc. Transferrin (TF) is the major iron transporting protein in the plasma. Lactoferrin (LF) is an iron binding glycoprotein of the transferrin family (García-Montoya et al., 2012). Previous studies have demonstrated the phylogenetic relation between LFs and TFs (Chang and Wang, 2011; Ford, 2001; Yu et al., 2017). In this study, twenty-four previously published TFs were studied by Gaussian Kernel similarity and Cosine similarity analysis.

The phylogenetic trees of twenty-four TFs were built by Gaussian Kernel similarity and Cosine similarity analysis (Fig. 6a and b). TF proteins and LF proteins were clustered into their corresponding branches. LF proteins were clustered into one branch and they were close to TFs of mammals. The TFs from mammals and salmonoids clustered into their corresponding branches, respectively. Our analysis displayed no misplaced and misclassified species. However, analysis performed by the ClustalW (Fig. 7) could not distinguish LFs from TFs. Thus, our method by Gaussian Kernel similarity and Cosine similarity analysis outperformed the multiple alignment method by the ClustalW.

3.4. Similarity analysis of 27 AF proteins

Antifreeze proteins (AFPs) play a vital role in the antifreeze effect of overwintering organisms. They have a wide range of applications in numerous fields, such as improving crop production and the quality of frozen foods.

Here we generated phylogenetic trees by Gaussian Kernel similarity and Cosine similarity analysis method (Fig. 8); and by the ClustalW analysis (Fig. 9) on the twenty-seven AFPs. Gaussian Kernel similarity and Cosine similarity analysis accurately classified all species (Fig. 8), which outperformed the ClustalW analysis that divided the TM group into three groups (Fig. 9).

4. Conclusions

In this study, we presented for the first time the spectral radii of 20 AAs calculation, and a novel 2-D graphical representation using spectral radii of 20 AAs. This method offers the advantage in easy visibility and inspection of similarity/dissimilarity between proteins, which sets ground for numerical characterizations of proteins. Furthermore, it avoids loss of information and ensures the integrity of the information. The proposed graphical representation method satisfies all requirements of a useful graphical representation proposed in Randić et al. (2010).

To characterize the protein sequence numerically, MSR, DSR, DSRT, DFG, DFGT and C_f of the sequence were extracted as the numerical features. The MSR, DSR and DFG features confirmed the integrity of information in different levels, while the DSRT and DFGT reflected dynamics features of protein sequences. In addition, fluctuation complexity reflected the non-linear feature of protein sequence. These features are based on the distributions of spectral radii and functional groups, and the fluctuation in net information gain of the sequence.

Finally, we employed Gaussian Kernel similarity and Cosine similarity analysis to measure the similarity of protein sequences using the feature vector. The method was performed on the similarity analysis of protein sequences of four data sets: nine ND5, thirty-five CoVPs, twenty-four TFs and twenty-seven AFPs. As the features reflected the protein sequence effectively, both the Gaussian Kernel similarity and Cosine similarity models have obtain satisfying results. Results of nine ND5, thirty-five CoVPs, twenty-four TFs and twenty-seven AFPs were consistent with the ClustalW method with further improvements. When compared to other methods, the analysis presented in this study achieved higher correlation coefficients with the ClustalW for nine ND5, which confirmed the efficiency of the proposed approach. The simulations of nine ND5 and thirty-five CoVPs indicated that the results of Gaussian Kernel similarity were to certain extend more sensitive than that achieved by Cosine similarity analysis.

In summary, we demonstrated that the proposed features and similarity models measured protein sequences efficiently. The obtained analysis was consistent with previously demonstrated evolution patterns. The proposed approach may therefore be applied in identification and classification of unknown species by protein sequences, as well as tracking the source of virus and designing drugs for disease therapy.

Acknowledgments

This work was supported by the Natural Science Foundation of China [Grant Numbers 61473335, 61533011]. We would like to thank Dr. Zhi-Ping Liu from School of Control Science and Engineering at Shandong University for the valuable suggestions.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jtbi.2018.03.001.

References

Chakraborty, A., Bandyopadhyay, S., 2012. Fogsaa: fast optimal global sequence alignment algorithm.. Sci. Rep. 3. 1746–1746.
Chang, G., Wang, T., 2011. Phylogenetic analysis of protein sequences based on distribution of length about common substring. Protein J. 30 (3), 167–172.
Czerniecka, A., Bieli Ska-W, D., W, P., Clark, T., 2016. 20d-dynamic representation of protein sequences. Genomics 107 (1), 16–23. doi:10.1016/j.ygeno.2015.12.003.
Dhifli, W., Diallo, A.B., 2016. Protnn: fast and accurate protein 3d-structure classification in structural and topological space. Biodata Min. 9 (1), 30.
Ellakkani, A., Mahran, H., 2015. An efficient numerical method for protein sequences similarity analysis based on a new two-dimensional graphical representation. Sar Qsar Environ. Res. 26 (2), 125–137.
Ford, M.J., 2001. Molecular evolution of transferrin: evidence for positive selection in salmonids. Mol. Biol. Evol. 18 (4), 639–647.
García-Montoya, I.A., Cendón, T.S., Arévalo-Gallegos, S., Rascón-Cruz, Q., 2012. Lactoferrin a multiple bioactive protein: an overview. BBA-GEN Subjects 1820 (3), 226–236.
Gotoh, O., 1982. An improved algorithm for matching biological sequences. J. Mol. Biol. 162 (3), 705–708.
Grassberger, P., 1986. How to measure self-generated complexity. Physica A 140 (1–2), 319–325.
Hayat, M., Khan, A., 2013. Wrf-tmh: predicting transmembrane helix by fusing composition index and physicochemical properties of amino acids. Amino Acids 44 (5), 1317–1328.
Hou, W., Pan, Q., Peng, Q., He, M., 2017. A new method to analyze protein sequence similarity using dynamic time warping. Genomics 109 (2), 123.
Huang, G., Hu, J., 2013. Similarity/dissimilarity analysis of protein sequences by a new graphical representation.. Curr. Bioinf. 8 (5), 539–544.
Jiang, Z., Zhou, Y., 2005. Using bioinformatics for drug target identification from the genome. Am. J. Perinatol. 5 (6), 387–396.
Li, C., Yu, X., Yang, L., Zheng, X., Wang, Z., 2009. 3-d maps and coupling numbers for protein sequences. Physica A. 388 (9), 1967–1972.
Li, Z., Geng, C., He, P., Yao, Y., 2014. A novel method of 3d graphical representation and similarity analysis for proteins. MATCH Commun. Math. Comput. Chem. 71, 213–226.
Liu, X., Yang, X., Wang, C., Yao, Y., Dai, Q., 2015. Number of distinct sequence alignments with k-match and match sections. Comput. Biol. Med. 63, 287.
Liu, Y., Li, D., Lu, K., Jiao, Y., He, P.-A., 2013. P–h curve, a graphical representation of protein sequences for similarities analysis. MATCH Commun. Math. Comput. Chem. 70 (1), 451–466.
Louie, B., Higdon, R., Kolker, E., 2009. A statistical model of protein sequence similarity and function similarity reveals overly-specific function predictions. Plos One 4 (10), e7546.
Meghanathan, N., 2014. Spectral radius as a measure of variation in node degree for complex network graphs. In: 2014 7th International Conference on u- and e-Service, Science and Technology. Haikou, pp. 30–33. doi:10.1109/UNESST.2014.8.
Mu, Z., Li, G., Wu, H., Qi, X., 2016. 3d–paf curve: A novel graphical representation of protein sequences for similarity analysis. MATCH Commun. Math. Comput. Chem. 75 (2), 447–462.
Parrott, L., 2010. Measuring ecological complexity. Ecol. Indic. 10 (6), 1069–1076.
Pugalenthi, G., Kumar, K.K., Suganthan, P., Gangal, R., 2008. Identification of catalytic residues from protein structure using support vector machine with sequence and structural features. Biochem. Biophys. Res. Commun. 367 (3), 630–634.
Qi, X., Wu, Q., Zhang, Y., Fuller, E., Zhang, C.Q., 2011. A novel model for dna sequence similarity analysis based on graph theory. Evol. Bioinf. Online 7 (7), 149.
Qi, Z., Jin, M., 2016. An intuitive graphical method for visualizing protein sequences based on linear regression and physicochemical properties. MATCH Commun. Math. Comput. Chem. 75 (2), 463–480.
Randić, M., Vračko, M., Novič, M., Plavšić, D., 2009. Spectral representation of reduced protein models. Sar Qsar Environ. Res. 20 (5–6), 415–427.
Randić, M., Zupan, J., Balaban, A.T., Vikić-Topić, D., Plavšić, D., 2010. Graphical representation of proteins. Chem. Rev. 111 (2), 790–862.
Snijder, E.J., Bredenbeek, P.J., Dobbe, J.C., Thiel, V., Ziebuhr, J., Poon, L.L., Guan, Y., Rozanov, M., Spaan, W.J., Gorbalenya, A.E., 2003. Unique and conserved features of genome and proteome of sars-coronavirus, an early split-off from the coronavirus group 2 lineage. J. Mol. Biol. 331 (5), 991–1004.
Su, J., Bao, J., 2013. A wavelet transform based protein sequence similarity model. Appl. Math. Inf. Sci. 7 (3), 1103–1110.
Sun, D., Xu, C., Zhang, Y., 2016. A novel method of 2d graphical representation for proteins and its application. MATCH Commun. Math. Comput. Chem. 75, 431–446.
Wang, L., Peng, H., Zheng, J., 2014. Adld: a novel graphical representation of protein sequences and its application.. Comput. Math. Method M. 2014 (2014), 959753.
Wu, H., Zhang, Y., Chen, W., Mu, Z., 2015. Comparative analysis of protein primary sequences with graph energy. Physica A 437, 249–262.
Wu, Z., Xiao, X., Chou, K., 2010. 2d-mh: a web-server for generating graphic representation of protein sequences based on the physicochemical properties of their constituent amino acids. J. Theor. Biol. 267 (1), 29–34.
Xu, C., Sun, D., Liu, S., Zhang, Y., 2016. Protein sequence analysis by incorporating modified chaos game and physicochemical properties into chou's general pseudo amino acid composition. J. Theor. Biol. 406, 105–115.
Xu, S.C., Li, Z., Zhang, S.P., Hu, J.L., 2014. Primary structure similarity analysis of proteins sequences by a new graphical representation.. Sar Qsar Environ. Res. 25 (10), 791–803.
Yao, Y., Yan, S., Han, J., Dai, Q., He, P.-a., 2014. A novel descriptor of protein sequences and its application. J. Theor. Biol. 347, 109–117.
Yao, Y., Yan, S., Han, J., Dai, Q., He, P.-a., 2014. A novel descriptor of protein sequences and its application. J. Theor. Biol. 347, 109–117.
Yao, Y., Yan, S., Xu, H., Han, J., Nan, X., He, P., Dai, Q., 2014. Similarity/dissimilarity analysis of protein sequences based on a new spectrum-like graphical representation. Evol. Bioinf. Online 10 (1), 87–96.
Yu, A., Lu, M., Tian, F., 2004. On the spectral radius of graphs. Linear Algerbra Appl. 387 (Suppl 2), 41–49.
Yu, L., Zhang, Y., Gutman, I., Shi, Y., Dehmer, M., 2017. Protein sequence comparison based on physicochemical properties and the position-feature energy matrix. Sci. Rep. 7, 46237. doi:10.1038/srep46237.
Zhang, Y., 2010. A new model of amino acids evolution, evolution index of amino

acids and its application in graphical representation of protein sequences. Chem. Phys. Lett. 497 (4), 223–228.
Zhang, Y.P., Sheng, Y.J., Zheng, W., He, P.A., Ruan, J.S., 2015. Novel numerical characterization of protein sequences based on individual amino acid and its application.. Biomed. Res. Int. 2015, 909567.
Zhou, J., Zhong, P., Zhang, T., 2016. A novel method for alignment-free dna sequence similarity analysis based on the characterization of complex networks. Evol. Bioinf. Online 12, 229.