MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **73** (2015) 731-748

ISSN 0340 - 6253

A Measure of Protein Sequence Characteristics Based on the Frequency and the Position Entropy of Existing *K*-words

Zhao-Hui Qi[*1], Meng-Zhe Jin[1], Hong Yang[2]

[1] *College of Information Science and Technology, Shijiazhuang Tiedao University, Shijiazhuang, Hebei, 050043, P. R. China*

[2] *Qingdao Binhai University, Qingdao, Shandong, 266555, P. R. China*

zhqi_wy2013@163.com

(Received November 27, 2014)

Abstract

Based on the frequency and the position distribution entropy of the existing k-words, we construct a modified statistical method for k-words. We call this method as an Existing-k-word method. The method consists of two parts. The first is to extract the existing k-words in proteins but not the all possible 20^k k-words. The other is to design a feature vector consisting of the frequencies and the position distribution entropies of the existing k-words. Then, this proposed method is applied to two datasets, nine ND5 proteins (NADH dehydrogenase subunit 5), and twenty-four transferrin protein sequences. The results illustrate the utility of the proposed method.

1. Introduction

Biomolecular sequence analysis grows enormously in recent years in respond to the explosive increase in the available data driven by the new development phase of sequencing technology. However, the amount of protein sequences confirmed by experiments has far exceeded the number of the proteins whose functions have been identified. The challenging task of researchers is to discover the previously unknown evolutionary relationships between sequences and to predict the functional and structural similarity of proteins. Generally speaking, functional annotation of proteins is highly dependent on sequence similarity to other known proteins. The majority of initial "first-pass" functional annotation on a genomic scale is reasonable and transitive as experiments for all proteins are a great challenge. The similarity analysis of sequences becomes an important methodology for the functional and

structural similarity of proteins. This is also a motivation for this paper in exploring effective and simple method for similarity analysis of sequences.

Considerable efforts have been made to reveal the relationships, most of which are based on the sequence alignment of the earlier researches. Nevertheless, as the number of biology sequences increases continuously, the computational complexity seriously affects the results of comparison by the alignment method, which grows exponentially with the sequence number in genomes and proteomes [1]. With the conservation assumption of homologous segments, the alignment-based comparison becomes misleading when it comes to genetic recombination and, in particular, genetic shuffling [2].

To overcome the drawback, a surge in the interest on the alignment-free sequence analysis has been aroused in the past 20 years. The 2D/3D graphical representation of coding sequences provides the visual inspection for sequence viewing, comparing and sorting [3]. A variety of graphical representations of DNA sequences are used to perform the similarity comparison of DNA sequences [4-11]. Similarly, scientists also begin to consider the graphical representation of protein sequences. Comparing with the graphical representation of DNA sequences, the graphical representation of proteins is difficult in the computational complexity. The number of possible alternative assignments for the 20 amino acids is the biggest difficulty [12]. The early graphical representation of proteins was published in 2004, by Randić [13]. Since then, several 2D/3D protein representations have been brought forward, such as [1,14-28]. And the dimension is not limited to 2D or 3D any longer, e.g. 20D representations in [29,30]. In [22], Randić et al has made a comprehensive review on the graphical representation of protein sequences.

Another alternative method is the numerical characterization method of biological sequences without going through graphical representation. In [31-37] a biological sequence is numerically converted into a vector or a matrix composed of the word frequency, the distance frequency or other statistical variables. Thus, the similarity or dissimilarity distance score can be achieved among their corresponding vectors by distance measures such as the Euclidean distance [38], the Pearson's correlation coefficient [39], the Cosine distance [40], the Manhattan distance [41], etc. In addition, k-word methods are also widely used in the DNA sequence analyses in recent years [2,42,43] since the first report of the k-word method for the sequence comparison in [38]. The most of schemes based on k-words consider the word composition under the background of Markov models, probabilistic models or optimization models [44]. Compared with the classical multiple sequence alignment methods, the k-word frequency provides the fast arithmetic speed. It can be applied to the full gene sequence

comparison. However, these methods have to face some difficulties of protein sequence comparison. That is, for a fixed k, there can be 4^k distinct k-words in a DNA sequence, while 20^k distinct k-words in a protein sequence. A big k will be a challenging in the computing time and space. On the other hand, the excessively low k value will bring the more or less loss of the amino acid adjacency information. Therefore, the length of k-words can not be ignored when they are applied to the protein sequence analysis. In addition, some of the k-word methods underestimate or even ignore the importance of the k-word location information that is closely related to the protein recombination. Recently, researchers have begun to add the word location information in their methods [43-46]. For example, in [43] the first and the last k-word locations are considered in the proposed methods.

In this paper, we construct a modified statistics method for k-words. We call this as an Existing-k-word method. The Existing-k-word method considers only the existing k-words in proteins but not the all possible 20^k k-words. Then, we propose a protein sequence feature extraction method based on the Existing-k-word frequency and the relative distance entropy. All the locations and compositions of Existing-k-words are concerned in the method. After that, the proposed method is applied on nine ND5 (NADH dehydrogenase subunit 5) proteins and twenty-four transferrin sequences for the similarity research and the phylogenetic analysis. By comparing the results with publicly available results from other researches, we draw a conclusion that the proposed method is an effective approach.

2. Methods

The model of k-word frequency is widely used in sequence comparison. Generally, for two sequences, the higher the similarity is, the more identical k-words are. For a DNA sequence, there are 4^k k-word frequencies to be considered. However, researchers have to face the difficulties brought from 20^k k-word frequencies for a protein sequence. There may be a mass of zero values in k-word frequencies. Here, we propose an Existing-k-words method which considers only the existing k-words in proteins but not the all possible 20^k k-words.

The traditional k-word frequency approach only considers the amino acids composition (AAC) in a protein, while the position information is largely ignored. This leads to a loss of some biological information. To correct this deficiency, we use distance entropy to characterize the distribution of Existing-k-words, and use the frequency of Existing-k-words to describe the AAC information. Our proposed protein feature extraction method, which is

based on Existing-k-word frequency and the information entropy, can capture both the quantitative characteristic and the position information of Existing-k-words. Given a sequence, we can build its feature vector composed of the frequency and the relative distance entropy. Thus, more phylogenetic information can be captured in our method.

2.1 Frequency vector of Existing-k-words

Here, we first give the frequency calculation method of Existing-k-words in a protein sequence. Let S denote a protein sequence with a length of n, $S = s_1 s_2 ... s_n$, $1 \le i \le n$. The s_i comes from 20 amino acids {A, R, D, C, Q, E, H, I, G, N, L, K, M, F, P, S, T, W, Y, V}. Suppose an Existing-k-word to be $M = m_1 m_2 \ldots m_k$, where the k is its length and the m_i is an amino acid. The repeat count of the Existing-k-word M is denoted by a $c(M)$ and its corresponding frequency is calculated as follows,

$$f(M) = \frac{c(M)}{n-k+1},$$

where $n-k+1$ equals the total number of Existing-k-words for a fixed k in the sequence.

Consider t different Existing-k-words, M_1, M_2, M_3,…, M_t, which are presented in alphabetical order. The frequency vector of the Existing-k-words can be defined as follows,

$$F = (f_1, f_2, ..., f_t)$$

In this study, we investigate the cases that the length k of Existing-k-words changes within a certain range instead of a fixed length so that more information is reserved in the feature vector. More specifically, let N be the maximum length of Existing-k-words. Counting the number of all Existing-k-words $(1 \le k \le N)$, we can get the frequency vector F of a protein sequence with n amino acids.

$$F = (f_1^1, f_2^1, ..., f_{t_1}^1, \quad f_1^2, f_2^2, ..., f_{t_2}^2, \quad ..., \quad f_1^N, f_2^N, ..., f_{t_N}^N)$$

where f_j^i is the frequency of the j th Existing-k-word in alphabetical order and t_i represents the number of different Existing-k-words. Here, $f_j^i = \frac{c_j^i}{n-i+1}$. And the c_j^i is the count of the j th Existing-k-word. For the sequence with n amino acids, $n-i+1$ equals the total count of the Existing-k-word.

To show the discussion simply, let us consider a short protein segment MGGM and set the maximum length N of Existing-k-word to be 3. According to the above explanation, we can

obtain 3 sub-sets of different frequencies. When $k=1$, we have $f_G^1=\frac{2}{4}$, $f_M^1=\frac{2}{4}$. When $k=2$, we have $f_{GG}^2=\frac{1}{3}$, $f_{GM}^2=\frac{1}{3}$, $f_{MG}^2=\frac{1}{3}$. When $k=3$, we have $f_{GGM}^3=\frac{1}{2}$, $f_{MGG}^3=\frac{1}{2}$. Then we can get the Existing-k-word frequency vector of MGGM by arranging the 7 frequencies in the alphabetical order as follows,

$$F=(f_G^1, f_{GG}^2, f_{GGM}^3, f_{GM}^2, f_M^1, f_{MG}^2, f_{MGG}^3)$$
$$=\left(\frac{2}{4},\frac{1}{3},\frac{1}{2},\frac{1}{3},\frac{2}{4},\frac{1}{3},\frac{1}{2}\right)$$

2.2 Position entropy vector of Existing-k-words

The position distribution of k-words usually contains some important biological information. In [47], the words distribution is used to differentiate various species. Ding et al. [43] have also shown that the k-word position in a sequence can effectively capture a lot of evolutionary information in the sequence.

For the Existing-k-word M (denoted by $M=m_1m_2\ldots m_k$), the $D_M=(d_1,d_2,\ldots,d_{c(M)})$ is its position vector, where the d_i is the position of the ith occurrence of the Existing-k-word M, and the $c(M)$ is the repeat count of the M in a sequence. Here, we define the position of the Existing-k-word M in proteins as the distance between its first amino acid m_1 and the origin when the M appears in the sequence. For the short segment MGGM, we can get the position vector of Existing-1-word G and Existing-2-word GG, $D_G=(2,3)$, $D_{GG}=(2)$.

To calculate the similarity distance between two different sequences conveniently, the position vector is condensed into a characteristic value which can characterize the location feature of a given Existing-k-word in a sequence.

In the Information theory, the entropy is known as a measure of the uncertainty of the probability distribution [48]. Suppose X to be a discrete random vector which has possible values $\{x_1,x_2,\cdots,x_n\}$ with corresponding probabilities $(f(x_1),f(x_2),\ldots,f(x_n))$. The entropy of X is

$$H(X)=-\sum_{i=1}^{n} f(x_i)\log_2 f(x_i).$$

The concept of entropy is extensively applied in many fields as an indicator to describe the disorder of a system.

In this paper, we propose the Existing-k-word position entropy as follows,

$$h(M) = -\sum_{i=1}^{c(M)} p_i \log_2 p_i ,$$

where p_i is the i th normalized probability value obtained from the position vector $D_M = (d_1, d_2, \ldots, d_{c(M)})$. The normalized formula to calculate p_i is the following,

$$p_i = d_i \Big/ \sum_{i=1}^{c(M)} d_i .$$

After working out the entropy of all Existing-k-words ($1 \le k \le N$), we can get a position entropy vector to characterize a protein as follows,

$$H = (h_1^1, h_2^1, ..., h_{t_1}^1, \quad h_1^2, h_2^2, ..., h_{t_2}^2, \quad ..., \quad h_1^N, h_2^N, ..., h_{t_N}^N),$$

where h_j^i is the distance entropy of the j th Existing-i-word in the alphabetical order and t_i represents the number of different Existing-i-words.

Likewise, an example is taken here to illustrate the distance entropy. When $N = 3$, the entropy vector of the segment MGGM is calculated as follows. Position vectors are obtained when k runs from 1 to 3. $D_G^1 = (2,3)$, $D_M^1 = (1,4)$, $D_{GG}^2 = (2)$, $D_{GM}^2 = (3)$, $D_{MG}^2 = (1)$, $D_{GGM}^3 = (2)$, $D_{MGG}^3 = (1)$. Then we can get the position entropy vector to characterize the segment MGGM as follows, $h_G^1 = 0.9710$, $h_M^1 = 0.7219$, $h_{GG}^2 = 0$, $h_{GM}^2 = 0$, $h_{MG}^2 = 0$, $h_{GGM}^3 = 0$, $h_{MGG}^3 = 0$. Then,

$$\begin{aligned} H &= (h_G^1, h_M^1, h_{GG}^2, h_{GM}^2, h_{MG}^2, h_{GGM}^3, h_{MGG}^3) \\ &= (0.9710, 0.7219, 0, 0, 0, 0, 0). \end{aligned}$$

From the results, although amino acids G and M have the same count 2, their distance entropies are obviously different. This is attributed to the different distribution between G and M in the segment.

Then we consider the distance entropy has acceptable differentiation performance when amino acid rearrangement or substitution occurs. We extend the original peptide MGGM to be MGGMMK and set the maximum length N of Existing-k-words to be 2. Considering amino acid rearrangement and substitution, we create two more segments. The three peptides are shown as:

MGGMMR the extended segment

MMMGGR amino acid rearrangement occurred in position 2,3 and 4,5

MGGGMR the 4th amino acid *M* is replaced with *G*

The three segments consist of a sequence group. The Existing-k-words of the group when $N=2$ should be composed of 9 k-words, which is denoted as { G, GG, GM, GR, M, MG, MM, MR, R}. The 9-dimensional entropy vectors of the three segments are shown as follows:

$$H_1 = (0.9710,\ 0,\ 0,\ 0,\ 1.3610,\ 0,\ 0,\ 0,\ 0)$$
$$H_2 = (0.9911,\ 0,\ 0,\ 0,\ 1.4591,\ 0,\ 0.9182,\ 0,\ 0)$$
$$H_3 = (0.9710,\ 0.9710,\ 0,\ 0,\ 0.6500,\ 0,\ 0,\ 0,\ 0)$$

It can be seen that the distance entropy is sensitive to amino acid rearrangement (comparing H_1 to H_2) and amino acid rearrangement substitution (comparing H_1 to H_3). We conclude that the distance entropy has a high degree of differentiation during amino acid rearrangement and substitution.

For a protein sequence, two vectors are obtained by the proposed method. The frequency vector holds the composition information of amino acids while the entropy vector captures their position information in a sequence. Here, the total number of different Existing-k-words $1 \le k \le N$ is defined as K, $K = t_1 + t_2 + \ldots + t_N$, the t_i represents the number of different Existing-i-words, $i = 1, 2, \ldots, N$. And we create a $2K$-dimensional feature vector V to characterize a protein sequence. The feature vector V is a combination of the frequency vector F and the position entropy vector H. We define it as follows,

$$V = (f_1^1, h_1^1, f_2^1, h_2^1, ..., f_K^N, h_K^N).$$

Here, the two successive elements with identical subscripts characterize the same Existing-k-word.

2.3 Distance calculations

The intrinsic structures of sequences are preserved in the proposed feature vectors. The quantitative comparison among sequences is feasible. For different sequences, the length of different Existing-k-words $K_1, K_2, K_3 \cdots$ are usually not identical with each other. Consider the situation where there are two feature vectors with different dimensions presented as follows,

$$V^{seq1} = (f_{M_1}^{seq1}, h_{M_1}^{seq1}, f_{M_2}^{seq1}, h_{M_2}^{seq1}, ..., f_{M_{K1}}^{seq1}, h_{M_{K1}}^{seq1})$$

$$V^{seq2} = (f_{M_1}^{seq2}, h_{M_1}^{seq2}, f_{M_2}^{seq2}, h_{M_2}^{seq2}, ..., f_{M_{K_2}}^{seq2}, h_{M_{K_2}}^{seq2})$$

where the $f_{M_i}^{seq\,j}$ ($h_{M_i}^{seq\,j}$) represents the frequency (entropy) value of the ith Existing-k-word in sequence j. Here, the length of different Existing-k-words in the two sequences are not identical, that is, $K_1 \neq K_2$.

The following three steps are proposed to align the dimension of the vectors.

Step 1. Consider the two Existing-k-word sets, $M_{seq1} = (M_1^{seq1}, M_2^{seq1}, M_3^{seq1}, ..., M_{K_1}^{seq1})$ and $M_{seq2} = (M_1^{seq2}, M_2^{seq2}, M_3^{seq2}, ..., M_{K_2}^{seq2})$. List all Existing-$k$-words of the two sets alphabetically. We only list one time when an Existing-k-word exists in one or both of the two sets. Thus, the merged Existing-k-words in one or both of the two sets are shown as follows,

$$M_{1,2} = (M_1^{1,2}, M_2^{1,2}, M_3^{1,2}, ..., M_{K_{1,2}}^{1,2}).$$

Step 2. Compare every Existing-k-word of the set $M_{1,2}$ with that of the set M_{seq1}. If there is a matched word M, let $f_M^{1*} = f_M^{seq1}$ and $h_M^{1*} = h_M^{seq1}$. But if there is no a matched word in M_{seq1}, then $f_M^{1*} = 0$ and $h_M^{1*} = 0$. Then we can get a new extended vector V_1^* of V^{seq1},

$$V_1^* = (f_1^{1*}, h_1^{1*}, f_2^{1*}, h_2^{1*}, ..., f_{K_{1,2}}^{1*}, h_{K_{1,2}}^{1*})$$

Step 3. The new extended vector V_2^* of V^{seq2} is the follows by the same method,

$$V_2^* = (f_1^{2*}, h_1^{2*}, f_2^{2*}, h_2^{2*}, ..., f_{K_{1,2}}^{2*}, h_{K_{1,2}}^{2*})$$

With the same dimension, the extended feature vectors V_1^* and V_2^* can be used to calculate the similarity distance of two sequences directly. Here, we use the Cosine function of the angle of two vectors to represent the correlation degree of two protein sequences. For the sake of clarity, the two vectors are simplified as $V_1^* = (a_1, a_2, ..., a_{K_{1,2}})$ and $V_2^* = (b_1, b_2, ..., b_{K_{1,2}})$. Thus, the Cosine function $Cos(V_1^*, V_2^*)$ is calculated as follows,

$$Cos(V_1^*, V_2^*) = \frac{\sum_{i=1}^{K_{1,2}} a_i \times b_i}{\sqrt{\sum_{i=1}^{K_{1,2}} a_i^2 \times \sum_{i=1}^{K_{1,2}} b_i^2}}.$$

The Cosine value of an angle is from -1 to 1. To get more biological explanation, we normalize the value of $Cos(V_1^*, V_2^*)$ from 0 to 1. Then the formula to calculate the similarity distance $D(V_1^*, V_2^*)$ of two given protein sequences is

$$D(V_1^*, V_2^*) = \frac{1 - Cos(V_1^*, V_2^*)}{2}.$$

In order to demonstrate the proposed feature extraction method, we consider two short protein sequences (*Protein I* and *Protein II*) and a random protein segment (*Protein R*). The *Protein I* and *Protein II* are segments from the yeast Saccharomyces cerevisiae introduced in Randić et al [14]. *Protein R* is a random segment by *Protein I* and *Protein II*.

Protein I:

WTFESRNDPAKDPVILWLNGGPGCSSLTGL

Protein II:

WFFESRNDPANDPIILWLNGGPGCSSFTGL

Protein R:

SFTESRPDPAADPVILWLSGPPGCKSLTTL

The frequency and the distance entropy of every Existing-k-word are obtained by the proposed method. By the distance formula $D(V_1^*, V_2^*)$, we can get the similarity distances among *Protein I*, *Protein II* and *Protein R* as shown in Table 1 when the length N of Existing-k-words changes from 1 to 7. The result shows that the similarity distances become stable when $N \geq 3$.

Table 1 Similarity distances among *Protein I*, *Protein II* and *Protein R* when the length of Existing-k-words changes from 1 to 7

Length N of Existing-k-words	Distance (PI, PII)	Distance (PI, PR)	Distance (PII, PR)
1	0.03943	0.06961	0.13249
2	0.05002	0.06578	0.13544
3	0.06208	0.06641	0.14551
4	0.06249	0.06727	0.14623
5	0.06298	0.06819	0.14702
6	0.06356	0.06920	0.14784
7	0.06425	0.07029	0.14869

3. Applications

In this section, the proposed feature extraction method is applied to two datasets, nine ND5 proteins and twenty-four transferrin sequences, to prove its correctness and effectiveness. By the proposed method, protein sequences are converted into the corresponding feature vectors composed of the frequency and the distance entropy of Existing-k-words ($1 \leq k \leq N$). To facilitate the calculation, all the feature vectors are extended to the identical dimension according to the three steps. Then, the similarity/dissimilarity matrix can be constructed

directly by calculating the similarity distances based on the feature vectors. In addition, the phylogenetic tree of the nine ND5 proteins and the twenty-four transferrin sequences are constructed to illustrate the correctness and the effectiveness of our method.

3.1 Comparison of nine ND5 proteins

Nine ND5 proteins are selected to test the performance of our method. The Nine proteins follows, the human (Homo sapiens, AP_000649), gorilla (Gorilla gorilla, NP_008222), common chimpanzee (Pan troglodytes, NP_008196), pigmy chimpanzee (Pan paniscus, NP_008209), fin whale (Balenoptera physalus, NP_006899), blue whale (Balenoptera musculus, NP_007066), rat (Rattus norvegicus, AP_004902), mouse (Mus musculus, NP_904338), opossum (Didelphis virginiana, NP_007105). All the proteins are downloaded from the NCBI Genbank.

According to our previous discussion, the extended feature vector V^* can be obtained by calculating the frequencies and entropies of Existing-k-words in protein sequences. By using the similarity distance formula $D(V_1^*, V_2^*)$, we can get the similarity matrix composed of different pair-wise sequence distances. Here, we set the maximum Existing-k-word length N to be 6. Then, we list the similarity matrix by the proposed method, as shown in Table 2.

Table 2 Similarity matrix of nine ND5 proteins (the maximum Existing-k-word length N is 6)

Species	Human	Gorilla	C. Chim.	P. Chim.	F. Whale	B. Whale	Rat	Mouse	Opossum
Human	0	0.02849	0.01987	0.01896	0.05857	0.05727	0.06881	0.07065	0.08032
Gorilla		0	0.02786	0.02387	0.05908	0.05909	0.06897	0.06933	0.07732
C.Chim.			0	0.01563	0.06198	0.06036	0.07055	0.07193	0.08246
P.Chim.				0	0.05679	0.05639	0.06903	0.06727	0.07545
F.Whale					0	0.00759	0.06153	0.05939	0.07169
B.Whale						0	0.06287	0.05933	0.07317
Rat							0	0.04415	0.06625
Mouse								0	0.05981
Opossum									0

From Table 2, we can see that the results are almost consistent with the related works on the nine ND5 proteins, such as the NFV method by Huang and Yu [47], the geometrical center method by Yao et al. [1] and the IFS method by Ma et al. [49]. The distances among four species, the human, gorilla, common chimpanzee and pigmy chimpanzee, are relatively small. The closest pair presented in the table is the fin whale and the blue whale. Besides, the rat and the mouse have a smaller distance. The opossum is the most dissimilar among the nine species.

Based on the similarity matrix of ND5 proteins, we reconstruct their phylogenetic tree to compare with the results by the classical ClustalW method [50]. The evolutionary tree illustrated in Fig.1 is constructed by the similarity matrix of Table 2. As we have noticed, the tree is nearly consistent to the results in Fig.2 by ClustalW. The ClustalW is one of the most widely used multiple sequence alignment methods for homologous proteins and nucleotide sequences because of its excellent performance in phylogenetic analysis. However, when there are a lot of sequences, computational complexity of ClustalW can be a serious challenge. Unlike ClustalW, the k-word method can obtain the sequence similarity based on the simple k-word statistics and some simple calculations [41-43]. With the Existing-k-word concept, we further reduce statistical count comparing to traditional k-word method. The similar results in Fig.1 and Fig.2 tell that the proposed simple method is close to the complicated ClustalW in efficiency. And the computational complexity of the proposed method is reduced.

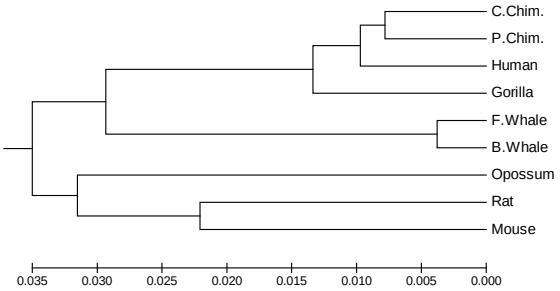

Fig. 1 Phylogenetic tree of nine ND5 proteins based on the extended feature vector with $N = 6$

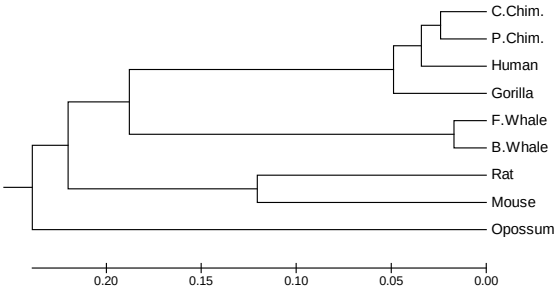

Fig. 2 Phylogenetic tree of nine ND5 proteins by ClustalW

3.2 Phylogenetic analysis on dataset of twenty-four transferrin sequences

To further verify the validity of our method, we select twenty-four transferrin sequences from vertebrate species as our second experiment dataset, which is studied in Ford [51], Chang and Wang [52]. Table 3 provides the taxonomic information and the accession number of every sequence. All the sequences are downloaded from the NCBI.

Table 3 The twenty-four transferrin sequences

No.	Name	Species	Accession	Length	No.	Name	Species	Accession	Length
1	Human TF	Homo sapien	S95936	2347	13	Frog TF	Xenopus laevis	X54530	2280
2	Rabbit TF	Oryctolagus coniculus	X58533	2279	14	Japanese flounder TF	Paralichthys olivaceus	D88801	2296
3	Rat TF	Rattus norvegicus	D38380	2293	15	Atlantic salmon TF	Salmo salar	L20313	2557
4	Cow TF	Bos Taurus	U02564	2338	16	Brown trout TF	Brown trout TF	D89091	2437
5	Buffalo LF	Bubalus arnee	AJ005203	2307	17	Lake trout TF	Salvelinus namaycush	D89090	2421
6	Cow LF	Bos Taurus	X57084	2405	18	Brook trout TF	Salvelinus fontinalis	D89089	2431
7	Goat LF	Capra hircus	X78902	2411	19	Japanese char TF	Salvelinus pluvius	D89088	2437
8	Camel LF	Camelus dromedaries	AJ131674	2337	20	Chinook salmon TF	Oncorhynchus tshawytscha	AH008271	5642
9	Pig LF	Sus scrofa	M92089	2578	21	Coho salmon TF	Oncorhynchus hisutch	D89084	2504
10	Human LF	H. sapiens	NM_002343	2627	22	Sockeye salmon TF	Oncorhynchus nerka	D89085	2153
11	Mouse LF	Mus musculus	NM_008522	2742	23	Rainbow trout TF	Oncorhynchus mykiss	D89083	2634
12	Opossum TF	Trichosurus vulpecula	AF092510	2480	24	Amago salmon TF	Oncorhynchus masou	D89086	2153

In this study, we conduct the phylogenetic tree to study the evolutionary relationships between the twenty-four transferrin sequences. The phylogenetic tree is the tree diagram showing the inferred evolutionary relationships among these twenty-four biological species. In Fig.3, we show the phylogenetic tree by the proposed approach. For comparison, in Fig.4 we show the phylogenetic tree constructed by the traditional alignment-based method ClustalW. With the careful observation on Fig.3 and Fig.4, we can find that the results by the proposed method and by the ClustalW are almost consistent to the results by Ford [51]. Further, it can be seen that in Fig.3 and Fig.4 all the well-separated transferrin (TF) proteins and lactoferrin (LF) proteins are accurately classified into the corresponding taxa.

4. Discussion

4.1 The choice of Existing-k-words

To investigate an appropriate N value (the length choice of Existing-k-words), the thirty-six nonzero elements of the upper triangular similarity matrix are represented in Fig.5 when the length N of words runs from 1 to 6. The Fig.5 is obtained according to the thirty-six pair-

wise distances of the nine ND5 proteins. Fig.4 shows the change tendency of the distance polyline with the length N from 1 to 6. The three broken lines ($N=4$, $N=5$ and $N=6$) are quite close to each other, i.e., when $4 \le N \le 6$, the distances among the proteins by the proposed method begin to reach a stable value. It shows that similarity distances become stable and gradually converge to a fixed value as the maximum length N of Existing-k-words increases.

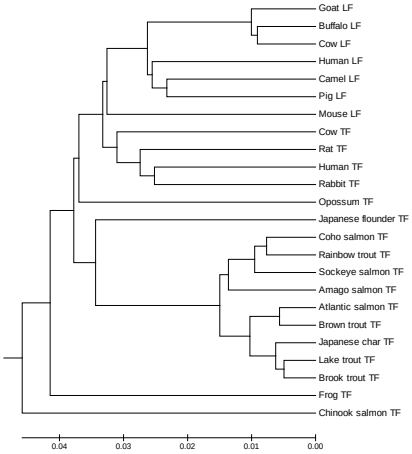

Fig. 3 Phylogenetic tree of twenty-four transferrin sequences by the proposed extended feature vector at N=6

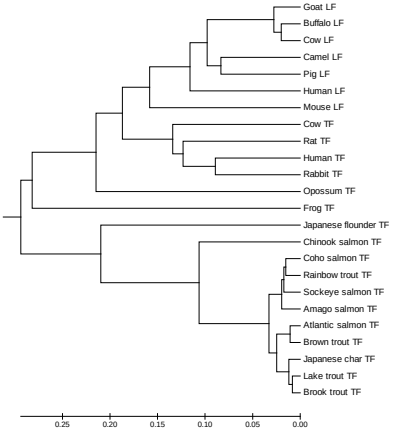

Fig. 4 Phylogenetic tree of twenty-four transferrin sequences by the traditional alignment-based method ClustalW

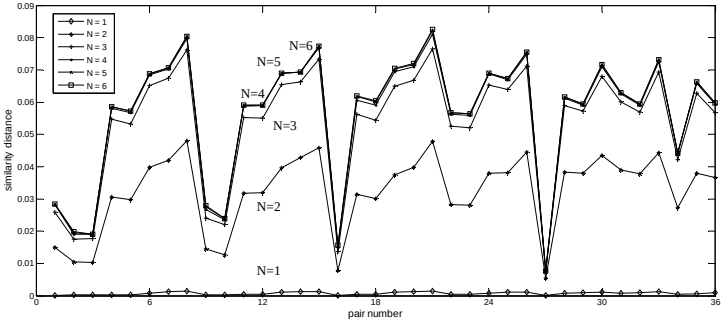

Fig. 5 The twenty-eight pair-wise distances among nine ND5 sequences obtained from the similarity matrixes when N ranges from 1 to 6

4.2 Comparison with the other methods

In order to further validate the proposed feature extraction approach, we compare our results with other results by different methods on the nine ND5 proteins.

More specifically, we first build the alignment of the nine proteins by using MEGA6 software and get the ClustalW distance matrix. Then, we compute the correlation coefficient of the two similarity matrixes from our approach and the ClustalW matrix. In Table 4, we illustrate the results by the related works mentioned in section 3.1 including the NFV method by Huang and Yu [47], the geometrical center method by Yao et al. [1] and the IFS method by Ma et al. [49]. From Table 4, we can find that the similarity comparison for nine ND5 proteins by the proposed method is closer to the results by the ClustalW method than by other studies. The proposed feature extraction method based on the frequency and the distance entropy of Existing-k-words can extract effectively characteristic information from protein sequences.

Table 4 The thirty-six pair-wise distances among the nine ND5 sequences when the length N of words ranges from 1 to 6.

Methods	Existing-*k*-word method & ClustalW	NFV & ClustalW	Geometrical center & ClustalW	IFS & ClustalW
Human	0.9860	0.9297	0.9033	0.9627
Gorilla	0.9846	0.9236	0.7983	0.9531
C.Chim.	0.9890	0.9414	0.9092	0.9671
P.Chim.	0.9890	0.9340	0.9063	0.9679
F.Whale	0.9939	0.9732	0.8003	0.9660
B.Whale	0.9968	0.9712	0.8235	0.8062
Rat	0.9855	0.9510	0.7970	0.8197
Mouse	0.9706	0.9551	0.7724	0.9579
Opossum	0.9547	0.9913	0.6832	0.9483

5. Conclusion

In this paper, we propose a modified k-word method, a measure of protein sequence similarity based on the frequency information and the relative distance entropy information of the Existing k-words in protein sequences. We call this method as the Existing-k-word method. It can characterize a protein by using a lower dimensional vector than the conditional k-word methods. Then, the proposed Existing-k-word method is applied to two separate applications, the similarity comparison of nine ND5 (NADH dehydrogenase subunit 5) proteins, and the evolutional analysis of twenty-four transferrin protein sequences. The results illustrate the utility of the proposed method.

Acknowledgment: This work supported by the National Natural Science Foundation of China (Grant No. 61272254), and by the Natural Science Foundation of Hebei Province, China (Project No. F2012210017), and by the Humanities and Social Sciences Research of Ministry of Education of China (Project name, The Origin, Propagation and Migration of Human Influenza Epidemic (1918-2010) from Space-time Perspective; Project No. 11YJCZH132).

References

[1] Y. H. Yao, Q. Dai, C. Li, P. A. He, X. Y. Nan, Y. Z. Zhang, Analysis of similarity/dissimilarity of protein sequences, *Proteins* **73** (2008) 864–871.

[2] S. Vinga, J. Almeida, Alignment–free sequence comparison – a review, *Bioinformatics* **19** (2003) 513–523.

[3] A. Nandy, M. Harle, S. C. Basak, Mathematical descriptors of DNA sequences: Development and applications, *Arkivoc* **9** (2006) 211–238.

[4] E. Hamori, J. Ruskin, H-curves, a novel method of representation of nucleotide series especially suited for long DNA sequences, *J. Biol. Chem.* **258** (1983) 1318–1327.

[5] A. Nandy, Two–dimensional graphical representation of DNA sequences and intron-exon discrimination in intron-rich sequences, *Comput. Appl. Biosci.* **12** (1996) 55–62.

[6] B. Liao, X. Shan, W. Zhu, R. Li, Phylogenetic tree construction based on 2D graphical representation, *Chem. Phys. Lett.* **422** (2006) 282–288.

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

[8] X. Q. Qi, J. Wen, Z. H. Qi, New 3D graphical representation of DNA sequence based on dual nucleotides, *J. Theor. Biol.* **249** (2007) 681–690.

[9] Z. H. Qi, T. R. Fan, PN-curve: a 3D graphical representation of DNA sequences and their numerical characterization, *Chem. Phys. Lett.* **442** (2007) 434–440.

[10] Z. J. Zhang, DV-Curve: a novel intuitive tool for visualizing and analyzing DNA sequences, *Bioinformatics* **25** (2009) 1112–1117.

[11] D. Bielinska–Waz, Four–component spectral representation of DNA sequences, *J. Math. Chem.* **47** (2010) 41–51.

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

[13] M. Randić, J. Zupan, Highly compact 2D graphical representation of DNA sequences, *SAR QSAR Environ. Res.* **15** (2004) 191–205.

[14] M. Randić, D. Butina, J. Zupan, Novel 2-D graphical representation of proteins, *Chem. Phys. Lett.* **419** (2006) 528–532.

[15] F. Bai, T. Wang, On graphical and numerical representation of protein sequences, *J. Biomol. Struct. Dyn.* **23** (2006) 537–546.

[16] M. Randić, 2-D graphical representation of proteins based on physic–chemical properties of amino acids, *Chem. Phys. Lett.* **444** (2007) 176–180.

[17] J. Feng, T. M. Wang, Characterization of protein primary sequences based on partial ordering, *J. Theor. Biol.* **254** (2008) 752–755.

[18] S. S. T. Yau, C. L. Yu, R. He, A protein map and its application, *DNA Cell Biol.* **27** (2008) 241–250.

[19] C. Li, X. Yu, L. Yang, X. Q. Zheng, Z. F. Wang, 3-D maps and coupling numbers for protein sequences, *Physica A* **388** (2009) 1967–1972.

[20] M. Randić, K. Mehulić, D. Vukičević, T. Pisanski, D. Vikić–Topić, D. Plavšić, Graphical representation of proteins as four–color maps and their numerical characterization, *J. Mol. Graph. Model.* **27** (2009) 637–641.

[21] P. A. He, Y. P. Zhang, Y. H. Yao, Y. F. Tang, X. Y. Nan, The graphical representation of protein sequences based on the physicochemical properties and its applications, *J. Comput. Chem.* **31** (2010) 2136–2142.

[22] M. Randić, J. Zupan, A. T. Balaban, D. Vikić–Topić, D. Plavšić, Graphical representation of proteins, *Chem. Rev.* **111** (2011) 790–862.

[23] A. Ghosh, A. Nandy, Graphical representation and mathematical characterization of protein sequences and applications to viral proteins, *Adv. Protein Chem. Struct. Biol.* **83** (2011) 1–42.

[24] M. Randić, M. Novič, A. R. Choudhury, D. Plavšić, On graphical representation of trans-membrane proteins, *SAR QSAR Environ. Res.* **23** (2012) 327–343.

[25] H. J. Yu, D. S. Huang, Novel 20-D descriptors of protein sequences and it's applications in similarity analysis, *Chem. Phys. Lett.* **531** (2012) 261–266.

[26] Z. H. Qi, J. Feng, X. Q. Qi, L. Li, Application of 2D graphic representation of protein sequence based on Huffman tree method, *Comput. Biol. Med.* **42** (2012) 556–563.

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

[28] L. Huang, H. Tan, B. Liao, HR-curve: a novel 2D graphical representation of protein sequence and its multi–application, *J. Comput. Theor. Nanos*. **10** (2013) 257–264.

[29] M. Novič, M. Randić, Representation of proteins as walks in 20-D space, *SAR QSAR Environ. Res*. **19** (2008) 317–337.

[30] A. Nandy, A. Ghosh, P. Nandy, Numerical characterization of protein sequences and application to voltage-gated sodium channel α subunit phylogeny. *In silico boil.* **9** (2009) 77–87.

[31] J. Qi, B. Wang, B. L. Hao, Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach, *J. Mol. Evol.* **58** (2004) 1–11.

[32] L. Gao, J. Qi, Whole genome molecular phylogeny of large dsDNA viruses using composition vector method, *BMC Evol. Biol.* **7** (2007) #41 (pp. 1–7).

[33] G. Lu, S. Zhang, X. Fang, An improved string composition method for sequence comparison, *BMC Bioinformatics* **9** (2008) #S15 (pp. 1–8).

[34] M. Takahashi, K. Kryukov, N. Saitou, Estimation of bacterial species phylogeny through oligonucleotide frequency distances, *Genomics* **93** (2009) 525–533.

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

[36] Z. H. Qi, M. H. Du, X. Q. Qi, L. J. Zheng, Gene comparison based on the repetition of single–nucleotide structure patterns, *Comput. Biol. Med.* **42** (2012) 975–981.

[37] Q. Dai, Z. F. Yan, Z. X. Shi, X. Q. Liu, Y. H. Yao, P. A. He, Study of LZ-word distribution and its application for sequence comparison, *J. Theor. Biol.* **36** (2013) 52–60.

[38] B. E. Blaisdell, A measure of the similarity of sets of sequences not requiring sequence alignment, *Proc. Natl. Acad. Sci.* **83** (1986) 5155–5159.

[39] G. Fichant, C. Gautier, Statistical method for predicting protein coding regions in nucleic acid sequences, *Comp. Appl. Biosci.* **3** (1987) 287–295.

[40] G. W. Stuart, K. Moffett, S. Baker, Integrated gene and species phylogenies from unaligned whole genome protein sequences, *Bioinformatics* **18** (2002) 100–108.

[41] T. J. Wu, J. P. Burke, D. B. Davison, A measure of DNA sequence dissimilarity based on Mahalanobis distance between frequencies of words, *Biometrics* **53** (1997) 1431–1439.

[42] X. Yang, T. Wang, A novel statistical measure for sequence comparison on the basis of k-word counts, *J. Theor. Biol.* **318** (2013) 91–100.

[43] S. Y. Ding, Y. Li, X. Yang, T. M. Wang, A simple k-word interval method for phylogenetic analysis of DNA sequences, *J. Theor. Biol.* **317** (2013) 192–199.

[44] V. Afreixo, C. A. Bastos, A. J. Pinho, S. P. Garcia, P. J. Ferreira, Genome analysis with inter–nucleotide distances, *Bioinformatics* **25** (2009) 3064–3070.

[45] Y. Gao, L. Luo, Genome–based phylogeny of dsDNA viruses by a novel alignment–free method, *Gene* **492** (2012) 309–314.

[46] Y. Huang, T. Wang, Phylogenetic analysis of DNA sequences with a novel characteristic vector, *J. Math. Chem.* **49** (2011) 1479–1492.

[47] D. S. Huang, H. J. Yu, Normalized feature vectors: A novel alignment–free sequence comparison method based on the numbers of adjacent amino acids, *IEEE Acm. T. Comput. Biol.* **10** (2013) 457–467.

[48] M. David, *Information Theory, Inference, and Learning Algorithms*, Cambridge Univ. Press, Cambridge, 2003.

[49] T. Ma, Y. Liu, Q. Dai, Y. Yao, P. A. He, A graphical representation of protein based on a novel iterated function system, *Physica A* **403** (2014) 21–28.

[50] M. A. Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, D. G. Higgins, Clustal W and Clustal X version 2.0, *Bioinformatics* **23** (2007) 2947–2948.

[51] M. J. Ford, Molecular evolution of transferrin: evidence for positive selection in salmonids, *Mol. Biol. Evol.* **18** (2001) 639–647.

[52] G. Chang, T. Wang, Phylogenetic analysis of protein sequences based on distribution of length about common substring, *Protein J.* **30** (2011) 167–172.