Communications in Information and Systems
Volume 20, Number 1, 45–60, 2020

Similarity analysis of protein sequences using a reduced k-mer amino acid model

Jia Wen*, Yuyan Zhang, and Huanxu Wang

Based on the properties of amino acid side chain, the 20 natural amino acids are divided into a simplified feature space, and the original protein sequence could be represented by a reduced amino acid sequence, which contains only four residues. Associating with this reduced protein sequence representation, the k-mer natural vector is defined and utilized to describe the similarity analysis of protein sequences, in which the frequencies and positional information of k-mers appearing in a reduced amino acid sequence are characterized by a feature vector. The similarity analysis of protein sequences can be easily and fast performed without requiring evolutionary models or human intervention. In order to show the utilities of our new method, it is applied on the real protein datasets for similarity analysis, and the obtaining results demonstrate that our new approach can precisely describe the similarities of protein sequences, and also strengthen the computing efficiency, compared with multiple sequence alignment. Therefore, our reduced k-mer amino acid representation model is a very powerful tool for analyzing and annotating protein sequence.

Keywords and phrases: Similarity analysis, protein sequence, a reduced amino acid model, k-mer natural vector, multiple sequence alignment.

1. Introduction

With the advantage of sequencing technology, more and more protein sequences are available now. The rapidly growth of protein sequence data creates many challenges to our bio-scientists, who need to find several credible tools to efficiently analyze a large number of protein sequences, and to disclose the secrets hidden in these data. Similarity analysis of protein sequences is one of the major topics in bioinformatics, which is the work to identify the similarities and dissimilarities of protein sequences. Since similar protein

*Corresponding author.

sequences are expected to have similar structures and biological functions, the similarity analysis of protein sequences has applications in structure and functional site prediction, sub-cellular location prediction, functional classification, evolutionary relationships of protein species, etc (Liu and Wang, 2006; Zhang and Wang, 2010; Liao et al., 2010; Rokas, 2011; Li et al., 2016; Hou et al., 2017; Wu et al., 2018).

Some laboratory testing methods have already been proposed for protein analysis. Although these wet-lab methods can verify the similarities of protein sequences, they are laborious and time consuming. Hence, computational methods accompanying with the advantages of economy are springing up and widely used in protein analysis (Krogh et al., 1994; Altschul et al., 1997; Bhasin and Raghava, 2004; Edgar, 2004; Tamura et al., 2013; Nikhila and Nair, 2018). Most of the computing methods are alignment-based methods that could describe the similarities of protein sequences, but they often associate with longer running time and higher memory burden, and some of them cannot deal with the whole-genome data (Yau et al., 2015; Ping et al., 2017; He et al., 2017). Alignment-free methods based on the numerical characterization of sequence are developed and successfully applied in protein analysis (Yau et al., 2008; Deng et al., 2011; Yu et al., 2011; Yin and Yau, 2015).

Of all alignment-free methods, the k-mer model method is one of the best developed one, in that, the sequence analysis could be performed much faster, and also be applied in the whole-genome comparison. However, the relationships of k-mers appearing in a sequence are more or less neglected by former k-mer model methods (Yu, 2013; Wen et al., 2014a; 2014b; He et al., 2017). To avoid losing the sequence-order information in a protein sequence, Chou (2001) outlined the pseudo amino acid composition (PseAAC) to reflect the sequence order information by a series of correlation factors that form the components of a vector. In addition, utilizing the biochemical properties of 20 amino acids, He et al. (2017) proposed a feature vector to describe the composition of amino acids in a protein sequence. But their cluster method cannot accurately depict the relationships of protein sequences. Although the k-mer model methods and its variants have been widely applied in protein studies, the dimension of numerical vector derived from protein sequence is very high when k is large, and the complexity of biological strings built on a 20-letter of amino acid is much higher than that built from only four letters of DNA or RNA (Ulitsky et al., 2006; Zhang and Yu, 2010).

Based on the properties of amino acid side chains, the 20 natural amino acids are replaced with one of four reduced amino acid residues, which constitute a simplified feature space. Therefore, each protein sequence is simply

represented by a reduced amino acid sequence. Associating with this reduced protein sequence representation, the k-mer natural vector is proposed to numerically characterize protein sequences. Furthermore, our new method is applied in the similarity analysis of proteins, and the obtaining results can fully show its utilities both in accuracy and efficiency.

2. Materials and methods

2.1. A reduced amino acid model

In protein synthesis process, amino acids are linked together into a polypeptide chain on the ribosome, in which the covalent bond between two amino acid residues is formed. There are 20 different amino acids existing in nature. Recently, much effort has been made to simplify protein sequence for better understanding and practical purposes (Li et al., 2008; Zhang and Yu, 2010). In these models, the composition for protein sequence is much simpler than the real one. Motivated by the properties of amino acid side chain, the 20 natural amino acids are divided into four reduced amino acid sets: hydrophobic amino acid, hydrophilic amino acid, acidic amino acid and basic amino acid, which are listed with one-letter code as follows:

1. Hydrophobic amino acid: **A**, V, L, I, P, M, F, G;
2. Hydrophilic amino acid: **Q**, S, T, C, N, Y, W;
3. Acidic amino acid: **D**, E;
4. Basic amino acid: **K**, R, H.

According to above classification for 20 natural amino acids, each group contains several amino acid residues which interact with others in a similar way. Without loss of generality, we assume that the representative amino acid residues for four groups are **A**, **Q**, **D** and **K**, respectively, which are the first amino acids in each group. Thus, a protein primary sequence can be simply represented by a four-letter reduce sequence that is composed of **A**, **Q**, **D** and **K**. If we take the first 20 amino acid residues of NADH Dehydrogenase 5 (ND5) sequence from Human as an example, the original protein sequence of 'MSRSGVAVADESLTAFNDLK' can be simplified by a reduced amino acid sequence of 'AQKQAAAAADDQAAAAQDAK'. It is obvious that the reduced amino acid sequence effectively decreases the complexity of original protein sequence.

2.2. *K*-mer model for a reduced amino acid sequence

Similar to the description for k-mer model of genetic sequence, the k-mer model for a reduced amino acid sequence is introduced as follows: Suppose a reduced amino acid sequence s of length L, '$\mathrm{S}_1\mathrm{S}_2\cdots\mathrm{S}_L$', where $\mathrm{S}_l \in \{\mathbf{A}, \mathbf{Q}, \mathbf{D}, \mathbf{K}\}$, $l = 1, 2, \cdots, L$. A string of consecutive k bases within a reduced amino acid sequence is called a k-mer. The k-mer appearing in the sequence could be enumerated by using a sliding window of length k, shifting one base each time from position 1 to $L - k + 1$, until the entire sequence has been scanned.

For any given k, there are 4^k different possible permutations of k-mers that may appear in a reduced amino acid sequence: s[1], s[2], ..., s[4^k].

2.3. *K*-mer natural vector

For each reduced amino acid sequence **s**, the k-mer natural vector is defined to be the concatenation of the following three vectors, each of which is of length 4^k (Wen et al., 2014a):

1. The k-mer counting vector $\left(n_{s[1]}, n_{s[2]}, \cdots, n_{s[4^k]}\right)$, where $n_{s[i]}$ is the number of k-mer $s\,[i]$ occurring in sequence **s**.
2. The k-mer mean distance vector $\left(\mu_{s[1]}, \mu_{s[2]}, \cdots, \mu_{s[4^k]}\right)$, where $\mu_{s[i]}$ is the arithmetic mean of the distances of the k-mer s[i] to the first base. If a specific k-mer $s\,[i]$ does not exist, $\mu_{s[i]}$ is defined to be zero.
3. The k-mer normalized central moment vector $\left(D_2^{s[1]}, D_2^{s[2]}, \cdots, D_2^{s[4^k]}\right)$, the component of which $(D_2^{s[i]})$ is the variance for the distances of k-mer $s\,[i]$ to the first base, which is defined as follows:

$$D_2^{s[i]} = \sum_{j=1}^{n_{s[i]}} \frac{\left(s\,[i]\,[j] - \mu_{s[i]}\right)^2}{n_{s[i]} \cdot (L - k + 1)},$$

where $n_{s[i]}$ denotes the number of k-mer $s\,[i]$ appearing in the sequence **s,** L is the length of sequence s, $s\,[i]\,[j]$ is the distance of jth k-mer $s\,[i]$ from the first base in sequence s, and $\mu_{s[i]}$ is the arithmetic mean of the distances of the k-mer s[i] to the first base.

If the distribution of each k-mer is different, two reduced amino acid sequences cannot be similar even though they contain the same set of k-mers and the same total distance measurement. Although each subset in the numerical parameters could not sufficient to depict a sequence, the combined

numerical parameters are sufficient to characterize each reduced amino acid sequence. The k-mer natural vector contains the positional information of k-mers, which are often neglected by former k-mer model methods, and we have verified that $3 * 4^k$-dimensional k-mer natural vector $(n_{s[i]},\ \mu_{s[i]}, D_2^{s[i]})$ is sufficient to characterize each sequence (Wen et al., 2014a).

2.4. The choice of k

For k-mer model methods, the parameter k has a great influence on obtaining results and computing efficiency. There is no a criterion to tell us how to choose a suitable k in previous k-mer methods. Therefore, it is important and difficult to select a suitable k for all kinds of proteins. To choose the optimum k^* for k-mer natural vector, we apply our reduced k-mer amino acid model on two real protein datasets from ND5 sequence and beta-globin sequence, which has been widely explored. Comparing with the results obtained by the methods of ClustalW and MUSCLE, we infer that the optimal k^* should be within a range of $[\text{ceil}(\log_4 \min(L)), \text{ceil}(\log_4 \max(L)) + 2]$, where L is the set of lengths for protein sequences considered. The optimal k^* over the range of k is chosen based on the following strategy: if the result of phylogenetic tree for the value k is relatively stable to that of $k + 1$, we choose $k^* = k$; otherwise k^* is equal to the maximum over the range of k chosen.

Once each protein sequence is numerically characterized by a k-mer natural vector, the cosine distance metric can be utilized to calculate the relative distance for each pair of protein sequences, which has been widely used in k-mer model methods (Berry et al., 1999; Stuart et al., 2002; Qi et al., 2004; Wen et al., 2014a; 2014b). Then, the phylogenetic tree can be drawn through the method of Neighbor Joining (NJ) using MEGA 6.06 (Tamura et al., 2013).

3. Results and discussion

Similarity analysis of protein sequences is an important tool to analyze and predict the structure and function of protein sequences. To investigate the performance of our reduced k-mer amino acid model, the similarity analysis of protein sequences is performed on two real datasets of ND5 sequence and beta-globin sequence to illustrate the utilities of our new approach.

3.1. Similarity analysis of 9 ND5 sequences

We first apply our new method in the similarity analysis of 9 ND5 sequences, the length of which are from 602 to 610. The ND5 sequence has been widely

Table 1: The relative distance among 9 DN5 protein sequences based on our new method

	Species	1	2	3	4	5	6	7	8	9
1.	AP_000649 Human	0								
2.	NP_00822 Gorilla	0.250	0							
3.	NP_008196 Common chimpanzee	0.162	0.253	0						
4.	NP_008209 Pigmy chimpanzee	0.137	0.236	0.045	0					
5.	NP_006899 Fin whale	0.458	0.482	0.419	0.434	0				
6.	NP_007066 Blue whale	0.438	0.467	0.396	0.409	0.163	0			
7.	AP_004902 Rat	0.473	0.514	0.497	0.484	0.503	0.489	0		
8.	NP_904338 Mouse	0.468	0.441	0.458	0.452	0.503	0.480	0.392	0	
9.	NP_007105 Opossum	0.440	0.495	0.447	0.439	0.534	0.516	0.488	0.488	0

discussed in the phylogeny and population genetic diversity of species for their high mutation rate. Until now, several methods have been proposed to analyze the ND5 sequences (Yao et al., 2008; Li et al. 2008; Wen and Zhang, 2009; Zhang and Yu, 2010; Liao et al., 2010; He et al., 2011; Yao et al., 2014; Yu et al., 2017). Based on our reduced k-mer amino acid model, to quantify the similarities of protein sequences, the relative distances among 9 ND5 sequences are shown in Table 1 with $k = 6$, and the NJ tree describing the evolutionary relationship of protein species is also shown in Figure 1. Meanwhile, the ClustalW is utilized to demonstrate the utilities of our new method, which is a widely used multiple sequence alignment algorithm to calculate the best matches for selected sequences, and depict the similarities of divergent species. The relative distance matrix and NJ tree for 9 ND5 sequences gotten by ClustalW are shown in Table 2 and Figure 2, respectively.

From Table 1, the pair of common chimpanzee and pigmy chimpanzee is the most similar one with the smallest relative distance. In addition, human, gorilla, common chimpanzee and pigmy chimpanzee are similar to each

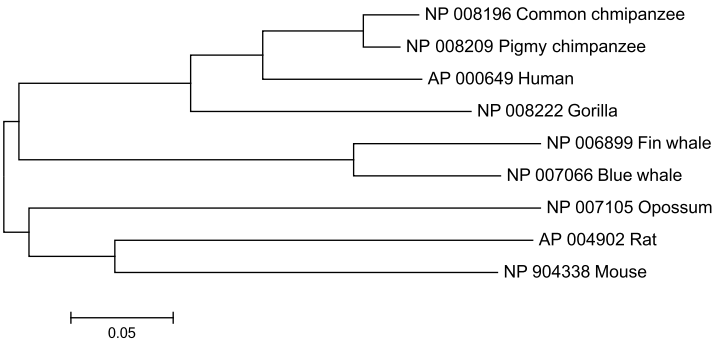

Figure 1: The NJ tree for 9 ND5 sequences based on the reduced k-mer amino acid model. The 9 DN5 sequence are clustering into groups of (human, gorilla, common chimpanzee and pigmy chimpanzee), (fin whale, blue whale) and (mouse, rat, opossum), which are similar to the results of multiple sequence alignment and published papers.

Table 2: The relative distance among 9 DN5 protein sequences based on ClustalW

	Species	1	2	3	4	5	6	7	8	9
1.	AP_000649 Human	0								
2.	NP_00822 Gorilla	0.104	0							
3.	NP_008196 Common chimpanzee	0.067	0.096	0						
4.	NP_008209 Pigmy chimpanzee	0.069	0.093	0.048	0					
5.	NP_006899 Fin whale	0.375	0.390	0.370	0.368	0				
6.	NP_007066 Blue whale	0.377	0.387	0.370	0.368	0.034	0			
7.	AP_004902 Rat	0.456	0.469	0.461	0.453	0.410	0.407	0		
8.	NP_904338 Mouse	0.443	0.453	0.448	0.443	0.422	0.415	0.241	0	
9.	NP_007105 Opossum	0.467	0.496	0.475	0.461	0.488	0.488	0.496	0.472	0

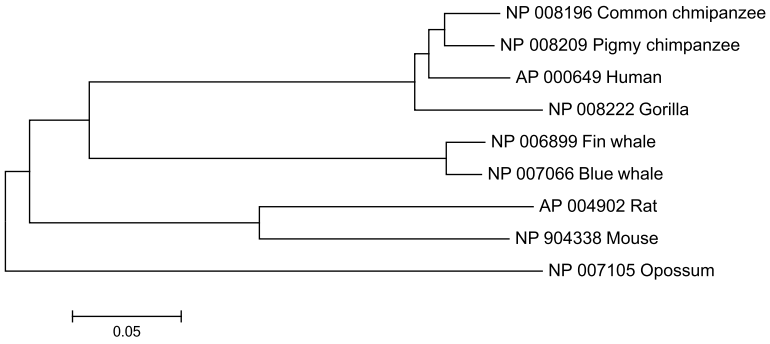

Figure 2: The NJ tree for 9 ND5 sequences based on multiple sequence alignment by ClustalW.

Table 3: The runtime comparisons for the reduced k-mer natural vector model and multiple sequence alignment by ClustalW*

Dataset	Reduced k-mer amino acid model	Multiple sequence alignment by ClustalW
ND5 sequence	1.480 seconds	2.741 seconds
Beta-globin sequence	1.767 seconds	9.052 seconds

*The configuration for our current laptop is Intel Core i5-2450 dual cores 2.50 GHZ with 8.00 Gb memory.

other, as well as the groups of (fin whale, blue whale) and (mouse, rat). However, the opossum looks a little far from others species. These results on the similarities/dissimilarities of 9 ND5 proteins are similar to the results shown in Table 2 obtained by multiple sequence alignment and and published papers (Yao et al., 2008; Li et al. 2008; Wen and Zhang, 2009; Zhang and Yu, 2010; Liao et al., 2010; He et al., 2011; Yao et al., 2014; Yu et al., 2017). Furthermore, comparing Figures 1 and 2, the evolutionary relationships of 9 DN5 proteins are similar to each other. In addition, it is of interest to find the clustering of (rat, mouse, opossum) in Figure 1, because they are all belonging to the Metatheria, which could not be found from the NJ tree shown in Figure 2. Moreover, the running time for our reduced k-mer amino acid model is much shorter than that used by ClustalW (see Table 3).

3.2. Similarity analysis of 88 beta-globin sequences

The beta-globin sequences from 88 species are then analyzed using our new method, which are the most common hemoglobin in adult human and often

utilized to explore the evolutionary relationships of species (Yau et al., 2008; Yu, 2013). This dataset has already been grouped by a new protein cluster method (He et al., 2017), and variance in length is from 140 to 148. The NJ tree of 88 beta-globin sequences was shown in Figure 3 with $k = 5$, based on our reduced k-mer amino acid model.

Look at Figure 3, 88 beta-globin sequences are correctly clustered into 20 different groups: Carnivora, Primates, Insectivora, Perissodactyla, Rodentia, Hyracoidea, Sirenia, Proboscidea, Diprotodontia, Galliformes, Anseriformes, Passeriformes, Columbiformes, Testudines, Perciformes, Cypriniformes, Salmoniformes, Gadiformes, Crocodylia and Anura, which are similar to the results obtained by He et al. (2017). Perciformes, Cypriniformes, Salmoniformes and Gadiformes are all Teleosts, they group together, which conforms to the conclusion in Cladistic analysis (Near et al., 2012). In addition, Galliformes, Anseriformes, Passeriformes and Columbiformes are belonging to the Galloanserae, the main group of modern birds (Sibley et al., 1988). This clustering is supported with morphological data and DNA sequence data (Chubb, 2004; Kriegs et al., 2007). Our phylogenetic tree agrees with the standard biological taxonomy and evolutionary relationship of species. However, these results could not be found from the Figure 2 of He et al. (2017).

To further verify the utilities of our new method, multiple sequence alignment are also performed on the same dataset. The NJ tree of 88 beta-globin sequences gotten from multiple sequence alignment by ClustalW is shown in Figure 4, in which species are colored the same as in Figure 3. Comparing Figures 3 and 4, on the whole, the evolutionary relationships of 88 beta-globin sequences are consistent with each other. In addition, the computing efficiency for our k-mer natural vector model method is higher than that used by ClustalW (see Table 3).

4. Conclusions

Integrating with a simplified amino acid residue representation, k-mer natural vector is defined and utilized to describe the similarity analysis of protein sequences, in which the frequencies and positional information of k-mers appearing in a reduced amino acid sequence are characterized by a feature vector. Our reduced k-mer amino acid model contains the information on relationships of k-mers, overcoming the deficiency of former k-mer model methods. With this new method, the features of k-mers hidden in the sequence can be effectively extracted, and each protein sequence is numerically characterized by a k-mer natural vector. Therefore, similarity analysis

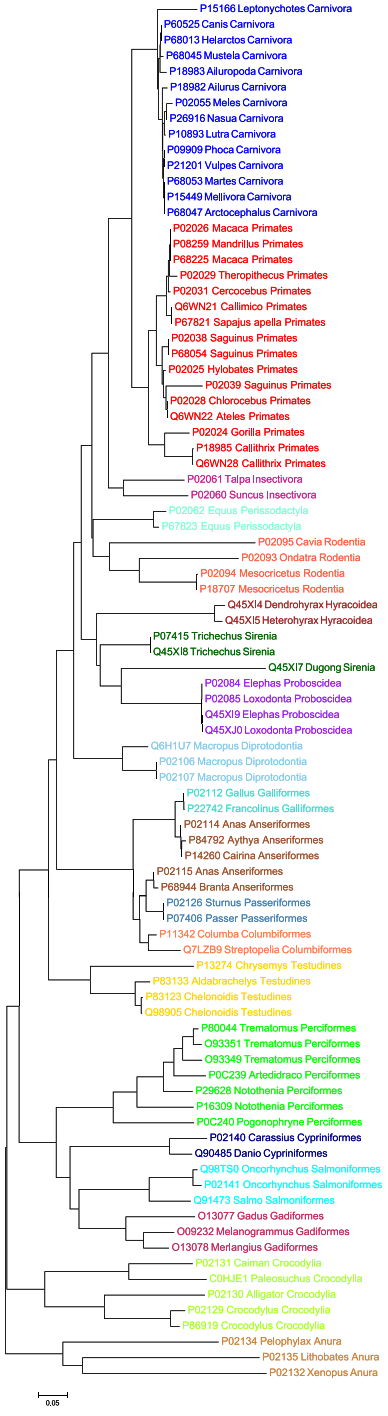

Figure 3: The NJ tree for 88 beta-globin sequences based on the reduced k-mer amino acid model. The 88 beta-globin sequences are correctly clustered into 20 groups: Carnivora, Primates, Insectivora, Perissodactyla, Rodentia, Hyracoidea, Sirenia, Proboscidea, Diprotodontia, Galliformes, Anseriformes, Passeriformes, Columbiformes, Testudines, Perciformes, Cypriniformes, Salmoniformes, Gadiformes, Crocodylia and Anura. This phylogenetic tree agrees with the conclusions in standard biological taxonomy and evolutionary relationship of species.

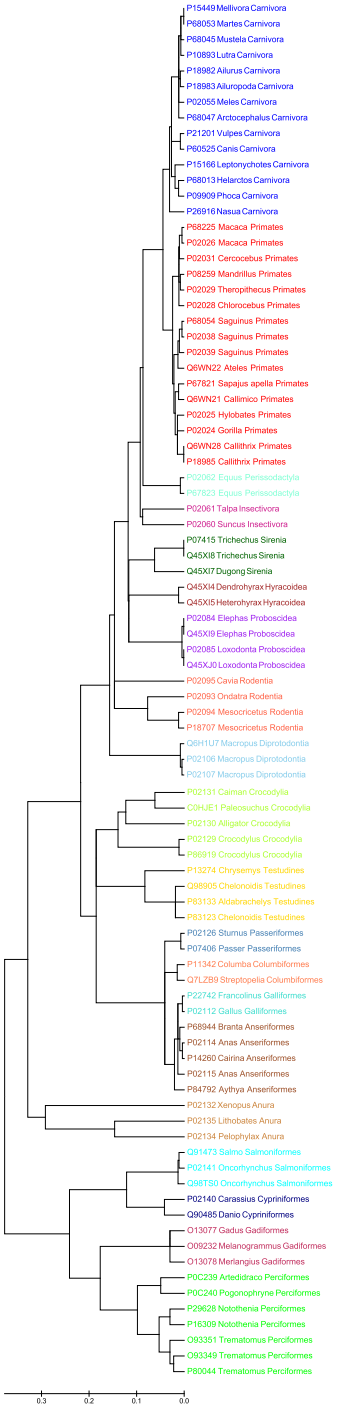

Figure 4: The NJ tree for 88 beta-globin sequences based on multiple sequence alignment by ClustalW. The 88 beta-globin sequences are clustered into 20 groups: Carnivora, Primates, Perissodactyla, Insectivora, Sirenia, Hyracoidea, Proboscidea, Rodentia, Diprotodontia, Crocodylia, Testudines, Passeriformes, Columbiformes, Galliformes, Anseriformes, Anura, Salmoniformes, Cypriniformes, Gadiformes, and Perciformes.

of protein sequences can be easily and fast performed without requiring evolutionary models or human intervention.

We illustrate the utilities of this new method in the similarity analysis of protein sequences on two real datasets, and the obtaining results demonstrate that our reduced k-mer amino acid model can precisely describe the similarities of protein sequences, which is similar to the results of multiple sequence alignment and published papers, and agrees with the conclusions in standard biological taxonomy and evolutionary relationship of species. In addition, we also get the new finding. For example, it is meaningful to find the clustering of (rat, mouse, opossum) from the similarity analysis of ND5 sequences, because they are all belonging to the Metatheria. Moreover, our new method greatly strengthens the computing efficiency, compared with multiple sequence alignment. Therefore, the reduced k-mer amino acid model is a very powerful tool for analyzing and annotating protein sequence.

Conflict of interest statement

The authors declare no competing final interests.

Authors' contributions

JW conceived and designed the project. YZ collected the data. JW and YZ performed the experiments, interpreted the results, and wrote the manuscript. All authors have read and approved the final manuscript.

Acknowledgements

This work is partially supported by Scientific Research Funding of Suihua University (K1501009, 2017-XGYYWF-017) and Natural Scientific Research Funding of Heilongjiang (LH2019A031).

References

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research, 25(17), pp. 3389–3402.

Berry, M.W., Drmac, Z. and Jessup, E.R., 1999. Matrices, vector spaces, and information retrieval. SIAM Review, 41(2), pp. 335–362. MR1684547

Bhasin, M. and Raghava, G.P.S., 2004. GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors. Nucleic Acids Research, 32(suppl_2), pp. W383–W389.

Chou, K.C., 2001. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins: Structure, Function, and Bioinformatics, 43(3), pp. 246–255.

Chubb, A.L., 2004. New nuclear evidence for the oldest divergence among neognath birds: the phylogenetic utility of ZENK (i). Molecular Phylogenetics and Evolution, 30(1), pp. 140–151.

Deng, M., Yu, C., Liang, Q., He, R.L. and Yau, S.S.T., 2011. A novel method of characterizing genetic sequences: genome space with biological distance and applications. PloS One, 6(3), p. e17293.

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research, 32(5), pp. 1792–1797.

He, L., Li, Y., He, R.L. and Yau, S.S.T., 2017. A novel alignment-free vector method to cluster protein sequences. Journal of Theoretical Biology, 427, pp. 41–52. MR3665144

He, P.A., Li, X.F., Yang, J.L. and Wang, J., 2011. A novel descriptor for protein similarity analysis. MATCH: Communications in Mathematical and in Computer Chemistry, 65(2), pp. 445–458. MR2663722

Hou, W., Pan, Q., Peng, Q. and He, M., 2017. A new method to analyze protein sequence similarity using Dynamic Time Warping. Genomics, 109(2), pp. 123–130.

Kriegs, J.O., Matzke, A., Churakov, G., Kuritzin, A., Mayr, G., Brosius, J. and Schmitz, J., 2007. Waves of genomic hitchhikers shed light on the evolution of gamebirds (Aves: Galliformes). BMC Evolutionary Biology, 7(1), p. 190.

Krogh, A., Brown, M., Mian, I.S., Sjölander, K. and Haussler, D., 1994. Hidden Markov models in computational biology: Applications to protein modeling. Journal of Molecular Biology, 235(5), pp. 1501–1531.

Li, C., Wang, J. and Zhang, Y., 2008. Similarity analysis of protein sequences based on the normalized relative-entropy. Combinatorial Chemistry & High Throughput Screening, 11(6), pp. 477–481.

Li, Y., Song, T., Yang, J., Zhang, Y. and Yang, J., 2016. An alignment-free algorithm in comparing the similarity of protein sequences based

on pseudo-Markov transition probabilities among amino acids. PloS One, 11(12), p. e0167430.

Liao, B., Liao, B., Sun, X. and Zeng, Q., 2010. A novel method for similarity analysis and protein sub-cellular localization prediction. Bioinformatics, 26(21), pp. 2678–2683.

Liu, N. and Wang, T., 2006. Protein-based phylogenetic analysis by using hydropathy profile of amino acids. FEBS Letters, 580(22), pp. 5321–5327.

Near, T.J., Eytan, R.I., Dornburg, A., Kuhn, K.L., Moore, J.A., Davis, M.P., Wainwright, P.C., Friedman, M. and Smith, W.L., 2012. Resolution of ray-finned fish phylogeny and timing of diversification. Proceedings of the National Academy of Sciences, 109(34), pp. 13698–13703.

Nikhila, K.S. and Nair, V.V., 2018. Protein Sequence Similarity Analysis Using Computational Techniques. Materials Today: Proceedings, 5(1), pp. 724–731.

Ping, P., Zhu, X. and Wang, L., 2017. Similarities/dissimilarities analysis of protein sequences based on pca-fft. Journal of Biological Systems, 25(01), pp. 29–45. MR3620293

Qi, J., Wang, B. and Hao, B.I., 2004. Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach. Journal of Molecular Evolution, 58(1), pp. 1–11.

Rokas, A., 2011. Phylogenetic analysis of protein sequence data using the Randomized Axelerated Maximum Likelihood (RAXML) Program. Current Protocols in Molecular Biology, 96(1), pp. 19–11.

Sibley, C.G., Ahlquist, J.E. and Monroe Jr, B.L., 1988. A classification of the living birds of the world based on DNA-DNA hybridization studies. The Auk, pp. 409–423.

Stuart, G.W., Moffett, K. and Leader, J.J., 2002. A comprehensive vertebrate phylogeny using vector representations of protein sequences from whole genomes. Molecular Biology and Evolution, 19(4), pp. 554–562.

Tamura, K., Stecher, G., Peterson, D., Filipski, A. and Kumar, S., 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. Molecular Biology and Evolution, 30(12), pp. 2725–2729.

Ulitsky, I., Burstein, D., Tuller, T. and Chor, B., 2006. The average common substring approach to phylogenomic reconstruction. Journal of Computational Biology, 13(2), pp. 336–350. MR2255263

Wen, J., Chan, R.H., Yau, S.C., He, R.L. and Yau, S.S., 2014a. K-mer natural vector and its application to the phylogenetic analysis of genetic sequences. Gene, 546(1), pp. 25–34.

Wen, J., Zhang, Y. and Yau, S.S., 2014b. k-mer Sparse matrix model for genetic sequence and its applications in sequence comparison. Journal of Theoretical Biology, 363, pp. 145–150.

Wen, J. and Zhang, Y., 2009. A 2D graphical representation of protein sequence and its numerical characterization. Chemical Physics Letters, 476(4–6), pp. 281–286.

Wu, C., Gao, R., De Marinis, Y. and Zhang, Y., 2018. A novel model for protein sequence similarity analysis based on spectral radius. Journal of Theoretical Biology, 446, pp. 61–70.

Zhang, S. and Wang, T., 2010. Phylogenetic analysis of protein sequences based on conditional LZ complexity. MATCH Commun Math Comput Chem, 63(3), pp. 701–716. MR2666629

Yao, Y.H., Dai, Q., Li, C., He, P.A., Nan, X.Y. and Zhang, Y.Z., 2008. Analysis of similarity/dissimilarity of protein sequences. Proteins: Structure, Function, and Bioinformatics, 73(4), pp. 864–871. MR3364713

Yao, Y., Yan, S., Xu, H., Han, J., Nan, X., He, P.A. and Dai, Q., 2014. Similarity/dissimilarity analysis of protein sequences based on a new spectrum-like graphical representation. Evolutionary Bioinformatics, 10, pp. EBO-S14713.

Yau, S.S.T., Yu, C. and He, R., 2008. A protein map and its application. DNA and Cell Biology, 27(5), pp. 241–250.

Yau, S.S.T., Mao, W.G., Benson, M. and He, R.L., 2015. Distinguishing proteins from arbitrary amino acid sequences. Scientific Reports, 5, p. 7972.

Yin, C. and Yau, S.S.T., 2015. An improved model for whole genome phylogenetic analysis by Fourier transform. Journal of Theoretical Biology, 382, pp. 99–110. MR3385919

Yu, C., Cheng, S.Y., He, R.L. and Yau, S.S.T., 2011. Protein map: an alignment-free sequence comparison method based on various properties of amino acids. Gene, 486(1), pp. 110–118.

Yu, H.J., 2013. Segmented K-mer and its application on similarity analysis of mitochondrial genome sequences. Gene, 518(2), pp. 419–424.

Yu, L., Zhang, Y., Gutman, I., Shi, Y. and Dehmer, M., 2017. Protein sequence comparison based on physicochemical properties and the position-feature energy matrix. Scientific Reports, 7, p. 46237.

Zhang, Y. and Yu, X., 2010, September. Analysis of protein sequence similarity. In Bio-Inspired Computing: Theories and Applications (BIC-TA), 2010 IEEE Fifth International Conference on (pp. 1255–1258). IEEE.

Jia Wen
School of Information Engineering
Suihua University
Suihua 152061
China
E-mail address: wenjia198021@126.com

Yuyan Zhang
School of Agriculture and hydraulic Engineering
Suihua University
Suihua 152061
China
E-mail address: alice_yuyan@163.com

Huanxu Wang
School of Information Engineering
Suihua University
Suihua 152061
China
E-mail address: 18677319@qq.com

Received September 5, 2019