A Novel Method for Similarity Analysis of Protein Sequences

Longlong Liu[1, a], Tingting Zhao[1,b] and Maojuan Liu[1,c *]

[1]School of Mathematical Sciences, Ocean University of China, Qingdao 266100, P.R. China

[a]xinxijishu@ouc.edu.cn, [b]sxzhaott@126.com, [c]18706480795@163.com

Keywords: protein sequence, similarity analysis, feature vector, cluster analysis, mitochondria NADH dehydrogenase

Abstract. A new method for similarity analysis of protein sequences is presented in this paper. On the basis of positions, proportion difference and various physicochemical properties of 20 kinds of amino acid in different protein sequences, representative information was extracted from protein sequence and converted into a numeric vector, thus further similarities of protein sequences were analyzed by studying the similarities between vectors. To facilitate the comparison between protein sequences of different length, every protein sequence is first mapped to a fixed-length vector, of which the vector information is relative position of amino acids. Then percentage of 20 kinds of amino acids in the sequence and 3 physicochemical properties are combined to constitute physicochemical information vector. Finally, a one-dimensional feature vector with 80 elements(feature vector) representing a protein sequence is synthesized. The shortest distance method was applied for cluster analysis on feature vectors so as to analyze similarities in protein sequences. In the numerical experiment part of the article, similarity analysis was conducted for 9 different species of the mitochondrial NADH dehydrogenase. The result of numerical experiment is consistent with the biological fact, which validates the effectiveness of model to a certain extent.

Introduction

In the field of biology, there are a variety of mathematical methods such as PSI-BLAST[1], Hidden Markov Model (HMM)[2], etc. used to analyze the protein sequences and thus to search the classification structure and function. In recent years, graphical methods for protein analysis have been proposed[3]. One class of methods is to represent 20 amino acids that compose protein by alphabet Ω={A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y}, so that any protein can be regarded as a sequence composed of 20 letters. Randic[4], P. He[5] and Yao[6] proposed methods for similarity analysis of protein sequences based on 20 kinds of amino acids which considered either the location information or the physicochemical properties of protein sequence.Nandy[7], Randic[8] and Raychaudhury[9] have put forward a new method, namely compressed matrix invariant method to investigate the mapping relationship between biological sequences. However, due to the nature of the complexity of the protein, such methods to reveal the nature of the difference between protein sequences are not enough. A new model for study on similarity of protein sequences was proposed in this paper based on position information of 20 amino acids in the protein sequence combined with their physicochemical properties.

Methods

Construction of amino acid position information vector

Mathematically speaking, a protein sequence may be regarded as a string of 20 amino acids on the alphabet Ω={A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y}. Thus, a protein sequence of 20 amino acids in this order that can be mapped (marked as f) into an adjoin matrix $G=\left(g_{ij}\right)_{20\times N}$, of which the rows represent 20 amino acids(the order is same to Ω), the columns represent the positions of amino acids in the sequence , N represents the length of protein sequence (number of amino acid residues). The element g_{ij} of matrix G is defined as follow:

(1) If the ith alphabet of Ω appears in the jth position of the sequence , then

$$g_{ij} = j$$

(2) If the ith alphabet of Ω disappears in the jth position of the sequence, then

$$g_{ij} = 0$$

For example, in a given protein sequence YCFESRNDPAKDPVILWLNGGPGCSSLTGA, the obtained adjoin matrix G according to the above mapping f is:

$$G = \begin{bmatrix} 0 & 0 & \mathbf{L} & 30 \\ 0 & 2 & \mathbf{L} & 0 \\ \mathbf{M} & \mathbf{M} & \mathbf{O} & \mathbf{M} \\ 1 & 0 & \mathbf{L} & 0 \end{bmatrix}_{20\times 30}$$

According to the definition of adjoin matrix, it is sparse matrix, while such matrix representation of protein sequence is unique. Therefore, the conversion successfully maps a protein sequence into a matrix. On the contrary, the matrix can be converted into the original protein sequence.

In order to extract the sequence information contained in adjoin matrix, the Euclidean distance between each two rows of adjoin matrix is obtained, so that we can get distance matrix $D=(d_{ij})_{20\times 20}$. Each element d_{ij} of D refers to the Euclidean distance between the i type and j type of amino acids, namely

$$d_{ij} = \sqrt{\sum_{k=1}^{N} \left(g_{ik} - g_{jk}\right)^2}, i, j = 1, 2, \ldots, 20 \qquad (1)$$

Where N refers to the length of protein sequence. As can be seen, D is a real symmetric matrix. The elements of matrix D are normalized, then the elements of each row are averaged , $a_i = \sum_{j=1}^{20} d_{ij} / 20, i = 1, 2, \ldots, 20$ i.e; where a_i represents the average of distances of the amino acid i with 20 kinds of amino acids, so a one-dimensional column vector with 20 elements $a = (a_1, a_2, ..., a_{20})'$ can be obtained, which is the position information vector of protein sequence.

Construction of the physicochemical information vector

Physicochemical properties of amino acids are the major basis of protein to fold into spatial structure. Since the value pK_a of amino acid refers to the relative difficulty for amino acid to release its dissolvable proton. The two functional groups $a-COOH$ and $a-NH_3^+$ are weak acidic groups that release protons via ionization in aqueous solution, while the capacity of releasing and receiving protons is the fundamental chemical property of protein which is also involved in the catalytic activity of the enzyme and the structure composition of the protein. pI is the isoelectric point of protein which can reflect the general information of protein composed of amino acid. Hence, the 3 physicochemical properties are selected for mathematical description of the protein sequence.

The corresponding pK_a(COOH), pK_a(NH_3^+) and pI (25°C)values are respectively given in reference [5]. However, these physicochemical values are same in different protein sequences under same conditions, which mean the differences comparison between protein sequences cannot only rely on these indicators. Thus, the proportions of 20 amino acids in protein sequence as well as the above 3 physicochemical indicators should be taken into consideration.

First of all, make a statistics for the percentage of 20 amino acids in protein sequence, namely:

$$b_i = \frac{n_i}{N}, i = 1, 2, \mathbf{L}, 20 \qquad (2)$$

Where n_i refers to the present times of the amino acid i in the protein sequence, N indicates the length of the protein sequence. Therefore, the percentage vector $b = (b_1, b_2, ..., b_{20})$ of 20 kinds

of amino acids is obtained. Then respectively multiply the three physicochemical indicators of 20 kinds of amino acids with corresponding percentages, i.e. $b_i*(y_{i1},y_{i2},y_{i3})$, where $y_{ij}, j=1,2,3$ refers to the physicochemical indicator of i amino acid. Thus a one-dimensional vector with 60 elements $g=(g_1,g_2,...,g_{60})$ composed of 3 physicochemical indicators of 20 amino acids (namely physicochemical information vector) is obtained. The physicochemical information vector is normalized later.

Synthesis of the feature vectors of protein sequences

First of all, transpose the one-dimensional position information vector with 20 elements a obtained in 2.1 and get $a'=(a_1,a_2,...,a_{20})$, then combine a' with the one-dimensional physicochemical information vector with 60 elements $g=(g_1,g_2,...,g_{60})$ and get a one-dimensional feature vector with 80 elements $P=(a',g)=(a_1,...a_{20},g_1,...,g_{60})=(p_1,p_2,\mathbf{L},p_{80})$, namely the 20 amino acids location information and physicochemical characteristics vector in protein sequence. Here we have completed the conversion from a protein sequence to the value vector. A one-dimensional feature vector with 80 elements can be obtained for protein sequence with any length via above methods. Only by converting protein sequences of different length without loss of information values into fixed-length vector can further comparative analysis be conducted.

Algorithm

Specific algorithm is summarized as follow:

a. To map the protein sequence containing 20 kinds of amino acid residues into a unique adjoin matrix $G_{20\times N}$.

Calculate the Euclidean distances between rows of adjoin matrix and obtain the distance matrix $D_{20\times 20}$ and then normalized the distance matrix $D_{20\times 20}$.

Calculate the average of row of $D_{20\times 20}$ and obtain a one-dimensional column vector with 20 elements $a=(a_1,a_2,...,a_{20})'$, i.e. the position information vector.

b. Conduct statistics for the percentage of 20 amino acids.

Respectively multiply the three physicochemical indicators of 20 kinds of amino acids with corresponding percentages, a one-dimensional physicochemical information vector with 60 elements $g=(g_1,g_2,...,g_{60})$ is obtained.

c. Transpose the vector $a=(a_1,a_2,...,a_{20})'$ obtained in step a and combine a' with g, then, get a one-dimensional feature vector with 80 elements $P=(p_1,p_2,\mathbf{L},p_{80})$.

d. Calculate the Euclidean distances between every 2 different protein sequences based on the feature vectors of protein sequences.

e. With the shortest Euclidean distance for hierarchical clustering, further analyze protein sequences similarity.

Result

The 9 different species of mitochondrial NADH dehydrogenase (ND5) sequences were selected from NCBI website (data from paper [10]).

The proposed method was applied to the similarity analysis of 9 kinds of ND5 protein, which means first obtain a one-dimensional feature vectors with 80 elements of 9 species of protein sequences. In order to reduce the impact of numerical difference on numerical experiment, the feature vector data were normalized. Eq.(3) is adopted to calculate the Euclidean distance of pairwise species, so that a 9×9 real symmetric matrix was obtained as shown in Table 1.

Table 1 Euclidean distances of nine species'ND5 protein sequences corresponding to the feature vectors

Species	Human(1)	Gorilla(2)	P.Chimpanzee(3)	C.Chimpanzee(4)	F.Whale(5)	B.Whale(6)	Rat(7)	Mouse(8)	Opossum(9)
Human(1)	0.000								
Gorilla(2)	0.186	0.000							
P.Chimpanzee(3)	0.174	0.225	0.000						
C.Chimpanzee(4)	0.138	0.209	0.121	0.000					
F. Whale(5)	0.280	0.286	0.280	0.317	0.000				
B.Whale(6)	0.296	0.327	0.291	0.330	0.097	0.000			
Rat(7)	0.635	0.637	0.577	0.635	0.546	0.543	0		
Mouse(8)	0.708	0.717	0.672	0.718	0.604	0.595	0.238	0	
Opossum(9)	0.942	0.913	0.889	0.949	0.792	0.808	0.464	0.509	0

The shortest-distance method was used in this paper for hierarchical cluster analysis. The result was provided in Tab2.

Table 2 The inter-class distance $D(1)$ after the first merger based on the shortest distance

	1	2	3	4	5,6	7	8	9
1	0							
2	0.186	0						
3	0.174	0.225	0					
4	0.138	0.209	0.121	0				
5，6	0.280	0.286	0.280	0.317	0			
7	0.635	0.637	0.577	0.635	0.543	0		
8	0.708	0.717	0.672	0.718	0.595	0.238	0	
9	0.942	0.913	0.889	0.949	0.792	0.464	0.509	0

we can see that the ND5 protein sequence of Fin Whale(5) and Blue Whale(6) are the most similar. ND5 protein sequence in Pigmy Chimpanzee(3),Common Chimpanzee(4)，Human(1) and Gorilla(2) are very similar, while Table.3 further demonstrates the ND5 protein {Common Chimpanzee(4),Human(1)} are more similar, while ND5 protein in Rat(7) and Mouse(8) are more similar. On the other hand, we found that the difference of ND5 protein between specie (9) and other species are greatest. In addition, the results obtained from numerical experiments are consistent with the fact of biological evolution[11,12], which also demonstrates the effectiveness of this method to some extent.

Conclusion

First, a one-dimensional feature vector with 80 elements was built in this paper, including spatial information and physicochemical information of protein sequence; then, the analysis of sequences similarity was conducted via hierarchical clustering for feature vectors of protein sequences; finally, in the numerical experiment part, similarity analysis was carried out for nine species of mitochondrial NADH dehydrogenase. The numerical results coincide with the fact that biological evolution. Compared with the previous methods, the proposed method in this paper has a small amount of calculation but with high accuracy, which is an effective new method.

Acknowledgment

The authors would like to thank all of the researchers who made publicly available the data used in this study. The authors would like to thank the University Basic Research Foundation

(No:201362031) and the National Natural Science Foundation of China (No: 61303145) for the support to this work .

References

[1] Altschul SF, Madden TL, Schafer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped blast and psi-blast: A new generation of protein data, Nucleic Acids Research, 25(17) (1997) 3389-3402.
[2] Krogh A, Brown M, Mian I, Sjolander K, Haussler D. Hidden markov models in computational biology: Applications to protein modeling, Journal of Molecular Biology, 235 (1994) 1501-1531.
[3] Randic M. 2-D Graphical represeatation of proteins based on virtual genetic code, SARQSAR Environ Res, 15 (2004) 147-157.
[4] Randic M , J. Zupan, A.T. Balaban, D. Vikic-Topic, D. Plavsic, Graphical representation of proteins, Chem. Rev. 111 (2011) 790–862.
[5] P.A. He, J.Z. Wei, Y.H. Yao, Z.X. Tie, A novel graphical representation of proteins and its application, Physica A ,2012, 391:93–99.
[6] Maggiora GM, Shanmugasundaram V. Molecular similarity measures, Methods Mol Biol. 672 (2011) 39-100.
[7] Nandy A, Basak SC, Simple numerical descriptor for quantifying effect of toxic substances on DNA sequences, J Chen InfComputSci. 40 (2000) 915.
[8] Randic M.,Mehulic, K.,Vukicevic, D.,Graphical representation of proteins as four-color maps and their numerical characterization, J.Mol.Graph.Model. 27 (2009) 637–641.
[9] Rayehaudhury C, Nandy A. Index scheme and similarity measures for macromolecular sequences, J ChemInfComputSci, 39 (1999) 243.
[10] Maggiora GM, Shanmugasundaram V. Molecular similarity measures, Methods Mol Biol. 672 (2011) 39-100.
[11] Yao YH, Dai Q, Li C etc. Analysis of similarity/dissimilarity of protein sequences, Proteins, 73(4) (2008) 864-871.
[12] Shiyuan Wang, FengchunTian, YuQiu , XiaoLiu, Bilateral similarity function: A novel and universal method for similarity analysis of biological sequences, Journal of Theoretical Biology 265 (2010) 194–201.