International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 11 Issue 3 ǁ March 2023 ǁ PP. 22-25

Research on Protein Sequence Alignment

Zhichong Ma[1]

[1]*Department of Communication and Art Design, University of Shanghai for Science and Technology, Shanghai, China*

Corresponding Author: Zhichong Ma

Abstract

Protein is a very important biological macromolecule for organisms. When its primary structure is mutated, it may cause disease. The emergence of protein sequence alignment can compare before and after the lesion to understand the pathogen. Not only that, it can also predict the function, evolution and structure of protein sequence through alignment. Based on this, this article summarizes and analyzes the more classic protein sequence alignment models on the market, which has certain reference value and far-reaching significance for the future development of protein sequence alignment.

Keywords: *protein, protein sequence, sequence alignment*

Date of Submission: 20-02-2023 Date of acceptance: 03-03-2023

I. INTRODUCTION

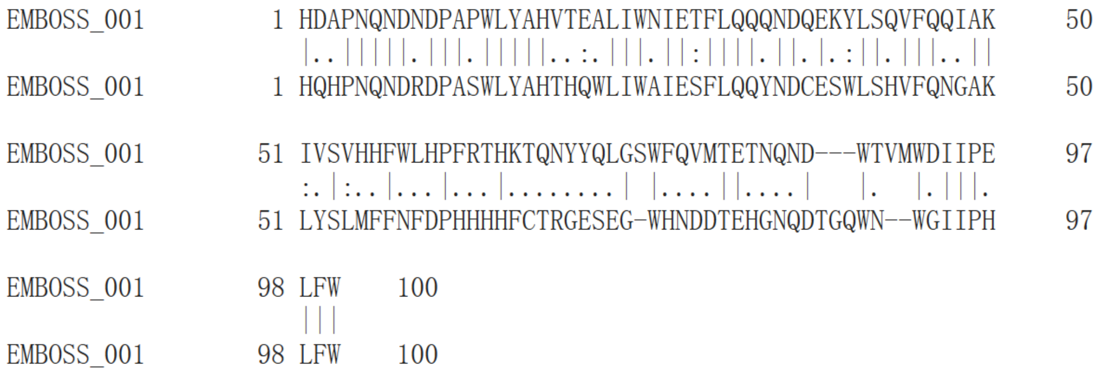
Protein is a complex and important biological polymer, which maintains the growth, differentiation and repair of biological cells. According to the structural level, it can be divided into primary, secondary, secondary and quaternary structures[1]. The primary structure of protein is composed of the composition and arrangement order of amino acids, which determines the high-level structure of the other three proteins, and then affects the characteristics and functions of the protein.

However, many diseases occur because of changes in the primary structure. Sicklemia is a biological change caused by the replacement of the 6th GLU on the HBS β-peptide chain by VAL. It will cause significant changes in red blood cell function, which will cause hemolytic anemia and other symptoms[2]. AD (Alzheimer's disease) is an extremely serious neurodegenerative disease, and its effects may last for years or even decades. According to reports, AD patients currently account for about 60-80% of all dementia patients, and it has become a serious social problem and research hotspot[3], so there is an urgent need to find a good diagnosis and treatment method. Current studies have found that clearing Aβ protein and regulating Tau protein play an important role in the treatment of this disease, although many inherent details have not been fully revealed[4]. For another example, pancreatic cancer, liver cancer, lung cancer and other malignant tumors will have abnormal protein expression in the early stage of the disease. Protein sequence alignment is the core of bioinformatics. It is the most effective means to compare various biological sequences. It is the core data structure of evolutionary analysis. It helps to understand the structural characteristics of new DNA and protein sequences. It can compare similarities and different regions between sequences to obtain protein functional similarity, structural motifs and evolutionary relationships[5]. It is possible to find pathogens and find treatments. Therefore, protein sequence ratio is of great significance to biology and medicine[6].

Protein sequence alignment can be divided into the following two categories according to the number of protein sequence alignments, namely protein paired sequence alignment and protein multiple sequence alignment. Here, this article will briefly summarize and analyze the widely used protein pairwise sequence alignment and protein multiple sequence alignment models on the market, so as to provide some ideas for the direction of protein sequence alignment.

II. PROTEIN PAIR SEQUENCE ALIGNMENT

Protein pair sequence alignment is to select two partial fragments of the sequence to be compared (which can be extended to all) and provide the greatest similarity[7]. It is a simple and fast method to find the similarity and homology between two sequences, as shown in Figure 1.The following is the classic protein paired sequence alignment model, as shown in Table 1, to expand: Altschul SF et al.[8] proposed the basic local alignment search tool BLAST, now the most common and extremely rapid protein paired sequence alignment method is to compare the input sequence with the sequence in the database, and quickly output the relationship between the two sequences, and not only can the protein and protein sequence alignment, but also nucleic acid sequence to protein library, nucleic acid sequence to nucleic acid library, protein sequence to nucleic acid library and nucleic acid sequence to nucleic acid library and other methods. Pearson WR proposed FASTA[9], which only considers the identity of amino acids, so it is fast and selectable. The amino acids of two sequences are scored by the PAM20 matrix. It is very effective to compare the sequence similarity of the loop-segmented sequences whose length can be transformed. It allows spaces or gaps to be inserted in the comparison process. Edgar RC[10] proposed UBLAST and USEARCH, respectively for local and global protein sequence alignment, is a clustering method, it is to quickly find several good, high similarity sequences, not all homologous sequences, so the speed is faster than BLAST, higher throughput, very suitable for NGS (next generation sequencing, also known as high-throughput sequencing technology, large-scale parallel or deep DNA sequencing technology[11]) classification. Pujari, J. J. et al.[12] proposed a semi-global sequence alignment technique based on NCSGA. This model can detect the optimal number of gaps and their positions of the DNA or protein sequences to be aligned and optimize them with genetic algorithms to obtain the best sequence alignment score. This semi-global sequence alignment method is very accurate. Talibart, H et al.[13] proposed using the Potts model to represent proteins and establishing a program PPalign based on an integer linear programming formula to calculate the best sequence alignment in the processable time, which is significantly better aligned than the uncoupled HHalign and PPalign.



```
EMBOSS_001          1 HDAPNQNDNDPAPWLYAHVTEALIWNIETFLQQQNDQEKYLSQVFQQIAK     50
                      |..|||||.|||.|||||..:.|||.||:||||.||.|.:||.|||..||
EMBOSS_001          1 HQHPNQNDRDPASWLYAHTHQWLIWAIESFLQQYNDCESWLSHVFQNGAK     50

EMBOSS_001         51 IVSVHHFWLHPFRTHKTQNYYQLGSWFQVMTETNQND---WTVMWDIIPE     97
                      :.|:..|...|...|........| |....||....|   |.  |.|||.
EMBOSS_001         51 LYSLMFFNFDPHHHHFCTRGESEG-WHNDDTEHGNQDTGQWN--WGIIPH     97

EMBOSS_001         98 LFW    100
                      |||
EMBOSS_001         98 LFW    100
```

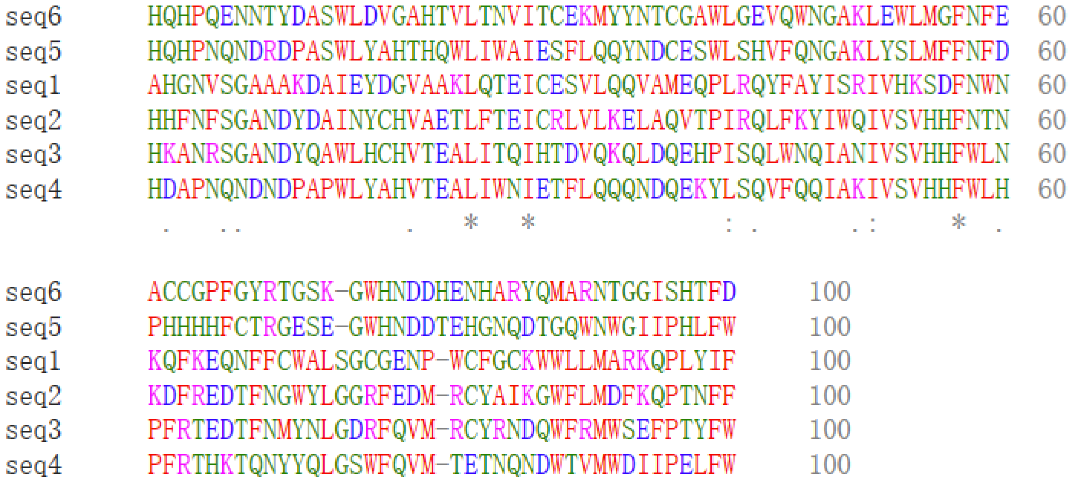
Figure1:Protein Pair-wise Sequence Alignment

Table1: Relevant Studies on Protein Pair-wise Sequence Alignment

Author	Model	Target
Altschul SF et al. [8]	BLAST	Protein local comparison search
Pearson WR et al [9]	FASTA	Initn score calculations are especially useful for sequences that share sequence similarity areas variable-length loops
Edgar RC et al [10]	UBLAST and USEARCH	Enable critical local and global searches for large sequence databases at extremely fast speeds
Pujari, J.J et al [12]	Protein Pair-wise Sequence Alignment Based NCSGA	It automatically measures the alignment of a sequence of bases to determine the best score for a DNA or protein sequence
Talibart, H et al [13]	PPalign	Optimized alignment using the Potts model with direct coupling information

III. PROTEIN MULTIPLE SEQUENCE ALIGNMENT

Protein multi-sequence alignment is to arrange 3 or more protein sequences together, and the amino acids in the same column are similar or the same to obtain the evolutionary connection of the query sequence, as shown in Figure 2.The following is the classical protein multiple sequence alignment model, as shown in Table 2, expanded: Robert C Edgar[14] proposed MUSCLE protein multiple sequence alignment, in turn, through the kmer count, logarithmic expected score function and tree to limit the partition method, accurate and fast output results. Sievers F[15] proposed Clustal Omega, which can accurately and quickly align most protein sequences. Used on a small test set, the accuracy can be very high, while on a larger data set, the quality is high and time-consuming, and sequences and information can be added during protein comparison. Alawneh, L[16] proposed a CPU-GPU hybrid parallel multiple sequence alignment method, which overcomes the scalability problems of other technologies. It combines the functions of multi-core CPUs and contemporary GPUs, and can output results extremely fast and effectively. O. Selvitopi[17] proposed the multi-sequence alignment software PASTIS. It combines a sparse matrix with a fully distributed sequence dictionary to get a better output effect. And without changing the basic sparse matrix, the original paranoia is replaced by amino acid sequences, and then expanded to millions of protein sequences. Zafalon, G[18] proposed a protein multi-sequence alignment method that mixes progressive and genetic algorithms. It uses the genetic algorithm part to generate multiple sequence alignments, and then uses the progressive part to perform partial reordering. This eliminates the propagation error problem of the progressive method and the local optimal problem of the genetic algorithm part, and greatly improves the quality of the comparison results. Rehman, H. A [19] improved the bird swarm algorithm BSA, and introduced multiple sequence alignment, thus proposed a new protein sequence alignment method EBSAA, which compared with genetic algorithm GA and particle swarm alignment algorithm PSAA, the comparison effect is very good.



```
seq6      HQHPQENNTYDASWLDVGAHTVLTNVITCEKMYYNTCGAWLGEVQWNGAKLEWLMGFNFE   60
seq5      HQHPNQNDRDPASWLYAHTHQWLIWAIESFLQQYNDCESWLSHVFQNGAKLYSLMFFNFD   60
seq1      AHGNVSGAAAKDAIEYDGVAAKLQTEICESVLQQVAMEQPLRQYFAYISRIVHKSDFNWN   60
seq2      HHFNFSGANDYDAINYCHVAETLFTEICRLVLKELAQVTPIRQLFKYIWQIVSVHHFNTN   60
seq3      HKANRSGANDYQAWLHCHVTEALITQIHTDVQKQLDQEHPISQLWNQIANIVSVHHFWLN   60
seq4      HDAPNQNDNDPAPWLYAHVTEALIWNIETFLQQQNDQEKYLSQVFQQIAKIVSVHHFWLH   60
           .   ..           .   *   *             : .      .:     *  .

seq6      ACCGPFGYRTGSK-GWHNDDHENHARYQMARNTGGISHTFD      100
seq5      PHHHHFCTRGESE-GWHNDDTEHGNQDTGQWNWGIIPHLFW      100
seq1      KQFKEQNFFCWALSGCGENP-WCFGCKWWLLMARKQPLYIF      100
seq2      KDFREDTFNGWYLGGRFEDM-RCYAIKGWFLMDFKQPTNFF      100
seq3      PFRTEDTFNMYNLGDRFQVM-RCYRNDQWFRMWSEFPTYFW      100
seq4      PFRTHKTQNYYQLGSWFQVM-TETNQNDWTVMWDIIPELFW      100
```

Figure2:Protein Multiple sequence alignment

Table2: Relevant Studies on Protein Multiple sequence alignment

Author	Model	Target
Robert C Edgar et al [14]	MUSCLE	Fast distance estimation using kmer counts, asymptotic alignment using the new contour function of log-expected fractions, and refinement using tree-dependent restricted partitions
Sievers F et al [15]	Clustal Omega	Fast aligns almost any number of protein sequences and provides accurate alignment
Alawneh, L et al [16]	CPU-GPU hybrid parallel Multiple sequence alignment	The scalable multipair protein sequence alignment is accelerated using a hybrid CPU-GPU approach
O. Selvitopi et al[17]	PASTIS	Distributed many-to-many protein sequence alignment using sparse matrices
Zafalon, G et al [18]	A Hybrid Approach based on Progressive and Genetic Algorithms	A progressive approach is used to locally rearrange multiple sequence alignments generated by genetic algorithm-based tools
Rehman, H.A et al[19]	EBSAA	Enhanced flock alignment algorithm to determine the optimal alignment between different sequences

IV. CONCLUSION

This paper expounds the importance of protein sequence alignment for biology and medicine, and divides protein sequence alignment into two types: protein paired sequence alignment and protein multiple sequence alignment, and summarizes their respective definitions, functions and some classic and widely used models on the market, and lists the different scope of application and related principles of many protein sequence alignment models, aiming to point out the direction for the development of protein sequence alignment.

REFERENCES

[1]. Bordoloi, Hemashree. (2021). Analytical Model to Predict Protein Structure using Soft-Computing Approach. Bioscience Biotechnology Research Communications. 14. 10.21786/bbrc/14.6.67.

[2]. Pecker L H, Lanzkron S. Sickle cell disease. Ann Intern Med, 2021, 174: ITC1–ITC16.

[3]. Alzheimer's Association. 2020 Alzheimer's disease facts and figures[J]. Alzheimer's & dementia: the journal of the Alzheimer's association, 2020, 16(3):391-460.

[4]. Bai B, Wang X, Li Y, et al. Deep multilayer brain proteomics identifies molecular networks in Alzheimer's disease progression. Neuron, 2020, 105:975–991.

[5]. Almanza-Ruiz, S. H., Chavoya, A. & Duran-Limon, H. A. Parallel protein multiple sequence alignment approaches: a systematic literature review. J Supercomput 79, 1201–1234 (2023). https://doi.org/10.1007/s11227-022-04697-9.

[6]. Wang, Y., Wu, H. & Cai, Y. A benchmark study of sequence alignment methods for protein clustering. BMC Bioinformatics 19 (Suppl 19), 529 (2018). shttps://doi.org/10.1186/s12859-018-2524-4.

[7]. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol. 1981 Mar 25;147 (1): 195-7. doi: 10.1016/0022-2836(81)90087-5.

[8]. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct 5;215(3): 403-10. doi: 10.1016/S0022-2836(05) 80360-2.

[9]. Pearson WR. Rapid and sensitive sequence comparison with FASTP and FASTA. Methods Enzymol. 1990;183:63-98. doi: 10.1016/0076-6879(90)83007-v.

[10]. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010 Oct 1;26(19): 2460-1. doi: 10.1093/bioinformatics/btq461. Epub 2010 Aug 12.

[11]. Behjati S, Tarpey PS. What is next generation sequencing?Arch Dis Child Educ Pract Ed. 2013 Dec;98(6):236-8. doi: 10.1136/archdischild-2013-304340. Epub 2013 Aug 28.

[12]. Pujari, J.J., Canadam, K.P. (2022). Semi global pairwise sequence alignment using new chromosome structure genetic algorithm. Ingénierie des Systèmes d'Information, Vol. 27, No. 1, pp. 67-74. https://doi.org/10.18280/isi.270108

[13]. Talibart, H., Coste, F. PPalign: optimal alignment of Potts models representing proteins with direct coupling information. BMC Bioinformatics 22, 317 (2021). https://doi.org/10.1186/s12859-021-04222-4

[14]. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004 Mar 19;32(5):1792-7. doi: 10.1093/nar/gkh340

[15]. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011;7:539.

[16]. Alawneh, L., Shehab, M.A., Al-Ayyoub, M. et al. A scalable multiple pairwise protein sequence alignment acceleration using hybrid CPU–GPU approach. Cluster Comput 23, 2677–2688 (2020). https://doi.org/10.1007/s10586-019-03035-8

[17]. O. Selvitopi, S. Ekanayake, G. Guidi, G. A. Pavlopoulos, A. Azad and A. Buluç, "Distributed Many-to-Many Protein Sequence Alignment using Sparse Matrices," SC20: International Conference for High Performance Computing, Networking, Storage and Analysis, Atlanta, GA, USA, 2020, pp. 1-14, doi: 10.1109/SC41405.2020.00079.

[18]. Zafalon, G.;Gomes, V.;Amorim, A. and Valêncio, C. (2021). A Hybrid Approach using Progressive and Genetic Algorithms for Improvements in Multiple Sequence Alignments. In Proceedings of the 23rd International Conference on Enterprise Information Systems-Volume 2: ICEIS, ISBN 78989-758-509-8;ISSN 2184-4992, pages 384-391. DOI: 10.5220/0010495303840391

[19]. Rehman, H.A., Zafar, K., Khan, A., & Imtiaz, A. (2021). Multiple sequence alignment using enhanced bird swarm alignment algorithm. J. Intell. Fuzzy Syst., 41, 1097-1114.LI H, ZHAO D, ZENG J. KPGT: Knowledge-Guided Pre-training of Graph Transformer for Molecular Property Prediction [J]. Knowledge Discovery and Data Mining, 2022.