A Review about Different Multiple Sequence Alignment (MSA) Software

Daniel Park

Korea International School

ABSTRACT

This paper presents an overview of the Multiple Sequence Alignment (MSA) methods to investigate the methodology behind them and evaluate their advantages and shortcomings. We introduce the background behind sequence alignment and study some of the multiple sequence alignment software: Clustal and MAFFT in particular. Clustal, first developed in 1988 by Des Higgins, is an extensively utilized MSA computer program during the 1990s. MAFFT was published in 2002 and is the most commonly used MSA program recently. Both programs use progressive MSA, finding the region of similarity by using pairwise alignment. We compare both of the software and appraise their effectiveness and a few potential improvements. For instance, both programs require an extensive amount of memory and use GPU(Graphics Processing Unit). Lastly, we conclude the paper with future approaches to mitigate the drawbacks and lead to the overall development of the study of bioinformatics.

Introduction

Sequence alignment is a useful technique in bioinformatics. It arranges sequences of DNA, RNA, or protein to identify regions of similarities and eventually identify the evolutionary relationship between sequences. Therefore, with sequence alignment, it is possible to predict the function of the species.

There are various types of sequence alignment. Global alignment performs end-to-end alignment of the query sequence with the reference sequence. It is most useful when the sequences have similar nucleotide alignment and have an approximately equal length. Therefore, it is useful when comparing closely related species. The Needleman-Wunsch algorithm can be used to score the alignment. While global alignment focuses on matching the start and the end of the sequence, local alignment focuses on matching the similar regions between the query sequence and the reference sequence. It matches a contiguous subsection of one sequence with a contiguous subsection of another. Local alignment is most useful when comparing sequences that are different but have similar regions. The Smith-Waterman algorithm can be used to score local alignment. It is similar to the Needleman-Wunsch algorithm but has different rules and structures.

The history of sequence alignment started with Frederick Sanger, a British biochemist. Before DNA sequencing, proteins and RNA were sequenced first. Sanger determined the first sequenced protein, insulin, in the 1950s. He later developed the new DNA sequencing method called the “Sanger Sequencing” in 1977. It became the most widely used DNA sequencing method for almost four decades. Sanger Sequencing involves a similar process of PCR, polymerase chain reaction, that is widely used nowadays. While PCR requires two primers, a forward primer, and a reverse primer, Sanger Sequencing only requires one primer and dNTP. In the 1990s, a new sequence alignment computer program called “Clustal” was established. The original software “Clustal” was created by Des Higgins in 1988. Then the second version, “Clustal V”, was released with higher speed and simplicity. It was able to store and reuse old alignments. The software kept developing with new features. “Clustal X”, released in 1997, includes the graphics. Recently, a new software called “MAFFT (Multiple Alignment using Fast Fourier Transform)” was published in 2002. It includes a wide range of options for better sequence alignment such as PartTree, DPPartTree, FFT-NS-1, and more.

Genomic databases have been expanding with large amounts of data. Sequence alignment and Multiple sequence alignment(MSA) are some of the most actively researched areas in bioinformatics and they are the central tasks for studies in modern biology. With the rapid development of computer science and the increased raw data, molecular biologists are seeking more and more effective computer science string algorithms that can perform both sequence alignment and multiple sequence alignment(MSA) successfully. This paper introduces and analyzes the two renowned and frequently used sequence alignment software: Clustal and MAFFT.

What is Multiple Sequence Alignment(MSA)?

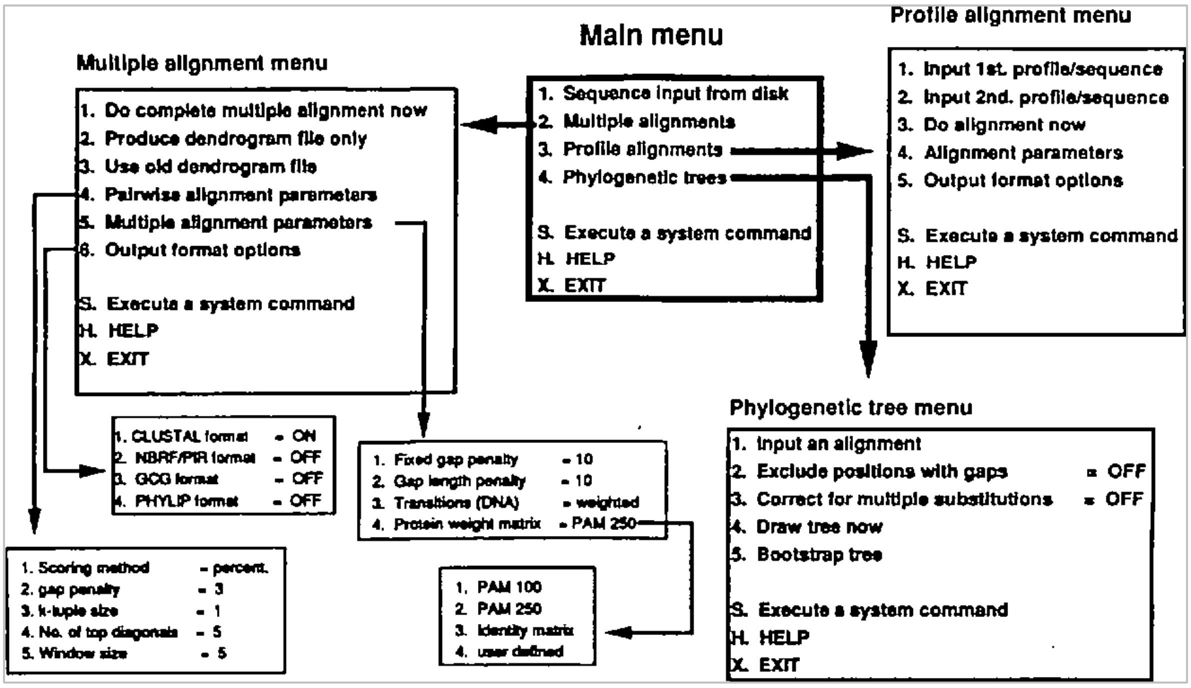
Multiple Sequence Alignment(MSA) refers to the algorithmic alignment of three or more evolutionary related sequences(protein or nucleic acid) of similar lengths. These sequences take into account evolutionary events such as mutations and rearrangements. MSA can be applied to both DNA, RNA, and protein sequences. It requires a complex combination of computation biological problems to compute an accurate MSA. The output of MSA displays homology that can be inferred and the evolutionary relationships between the sequences used.

About CLUSTAL

The most commonly used multiple sequence alignment software is Clustal. There are numerous versions of Clustal over the development of the algorithm, but Clustal X, in particular, is a new interface version of Clustal W with a graphical user interface. Similar to the other versions of Clustal, Clustal X performs multiple sequence alignment and displays the result of the alignment. Clustal X is improved in a way that it consists of a broader range of options, including the option to select a region of the alignment and realigning the region with different gap penalties, while keeping the other regions fixed.

Clustal uses pairwise progressive sequence alignment. It first does a pairwise sequence alignment that can be done from the sequence set. Based on the pairwise similarity of the sequences, a guide tree is created. The guide tree then constructs a multiple sequence alignment.

Another version of Clustal is “Clustal V”. It is a newer version of the original Clustal package of programs that features the ability to calculate phylogenetic trees and to store and reuse old alignments. Its main advantage comes from its simplicity, speed, sensitivity, and capacity. The sequences are aligned in a progressive manner, where sequences are aligned based on their sizes. In Clustal V, guide trees are created using the UPGMA method of Sneath and Sokal(1973) to be utilized for multiple alignment and phylogenetic purposes. Phylogenetic trees can be calculated using the Neighbor-Joining method of Saitou and Nei.



2. Exclude positions with gaps = OFF
3. Correct for multiple substitutions = OFF

1. CLUSTAL format = ON
2. NBRF/PIR format = OFF
3. GCG format = OFF
4. PHYLIP format = OFF

1. Fixed gap penalty = 10
2. Gap length penalty = 10
3. Transitions (DNA) = weighted
4. Protein weight matrix = PAM 250

1. Scoring method = percent.
2. gap penalty = 3
3. k-tuple size = 1
4. No. of top diagonals = 5
5. Window size = 5

Figure 1. The full menu of CLUSTAL V

Clustal can be used for various purposes. Recently, Clustal has been deeply associated with polymerase chain reaction (PCR). In particular, in 2013, a few researchers in China investigated the new method of differentiating Astragail and Hedysari Radix by using PCR amplification. They used 30 samples for Astragali Radix and 28 samples for Hedysari Radix. These sequences were amplified using PCR with a specific primer. They used PCR to amplify the sequence but they also utilized specifically Clustal W to align the sequences. In addition, in 2017, these researchers again explored for rapid identification of Cervus Nippon, C. elaphus. The PCR for these sequences was established based on the single-nucleotide polymorphisms(SNP) in COI and SRY sequence. SNPs in the COI and SRY sequences were found by Clutal X 2.1 program.

One of the main menus of Clustal is multiple alignments. In Clustal, the multiple alignments are aligned using the progressive manner, where sequences are aligned in larger and larger groups. These groups are based on the order in a 'guide tree', which is constructed using the UPGMA method of Sneath and Sokal (1973) that is from a pairwise alignment called the Wilbur and Lipman (1983). The final multiple alignment is done by creating multiple alignment aligned by larger and larger alignments. After the conventional dynamic programming algorithm, the residue on the sequences is contributed to the calculation of the alignment score. The score is determined by the average of all the scores between the residues at each position. When calculating, a full amino acid weight matrix, such as PAM 250 matrix, and two gap penalties are used.

Another main menu of Clustal is the phylogenetic tree. However, the ‘guide trees’ from the multiple alignments should not be used to create phylogenetic trees because not only it creates a great error but also it gives incorrect topologies if the rate of evolution varies. Instead, one can either calculate right after the multiple alignments or save the alignment in NBRF/PIR format and use it later. The Neighbor-Joining method of Saitou and Nei (1987) is a distance method that can correct multiple substitutions and gives a correct topology considering the variation of evolution.

The confidence levels on the tree are calculated using the bootstrap procedure which is similar to the Felsenstein procedure (1985).

Lastly, the main menu has a profile alignment option, which is also an alignment of old alignments. This can be used to add new sequences, predict its structure, and create multiple alignments. One advantage of profile alignment is that it allows you to control the multiple alignment processes better. Since the alignments are aligned in the same way as the progressive alignment, the same input and output format can be used.

About MAFFT

In addition to Clustal, MAFFT has been another well-known multiple sequence alignment software recently. In order to successfully extract biological information from large numbers of sequences and MSAs, programs such as MAFFT are necessary for its sophistication of algorithms. It features various multiple sequence alignment methods including progressive alignment.

MAFFT nowadays includes an improved large option for refining sequences and MSAs. PartTree and DPPartTree option both cluster sequences and compute the distance between the clusters. PartTree uses k-mer-based distance while DPPartTree uses dynamic programming (DP). Although the DPPartTree option is slower, it is more likely accurate. The FFT-NS-1 option is far more accurate than the former two options. It computes pairwise distance. The G-INS-1 option is accurate as well, but it takes a longer computational time. It is important to select appropriate strategies based on the sequences. If the sequences are large and homologous, options such as PartTree and DPPartTree are preferred. If the sequences are similar to each other, option FFT-NS-2 would be a wise choice.

There are two viewers that visualize the phylogenetic trees and are used for sequence selection: Phylo.io and Archaeopteryx. Archaeopteryx was preferred but recently Phylo.io became more popular.

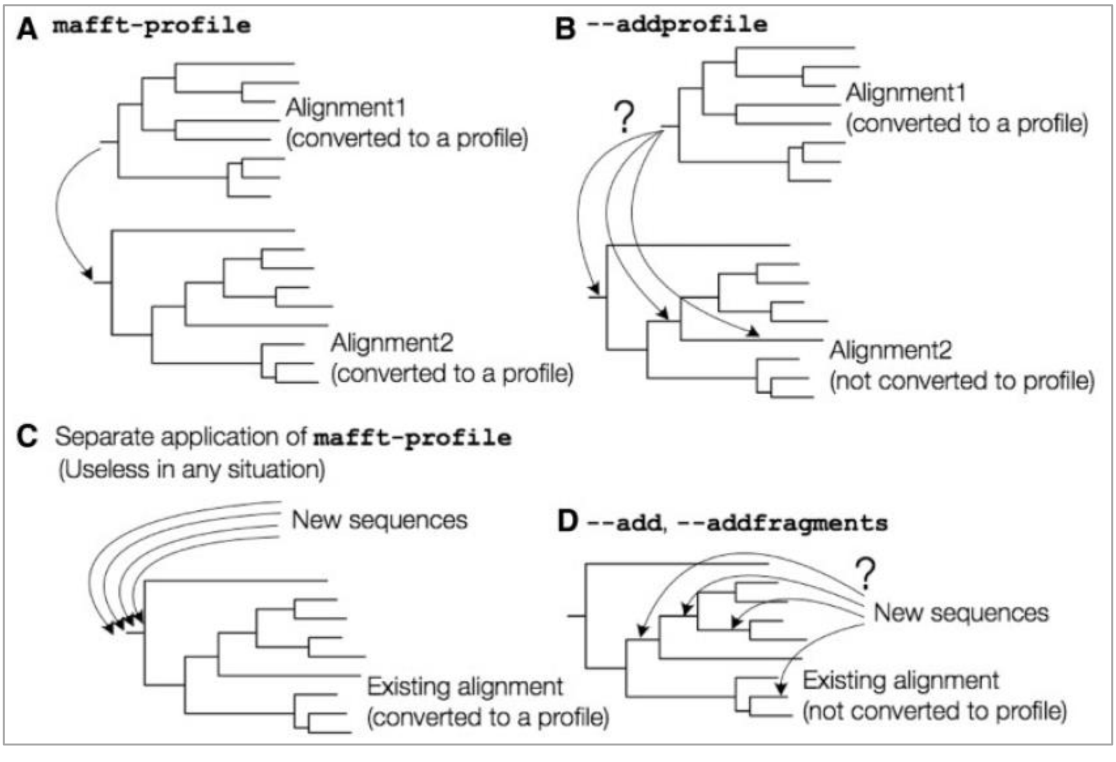


Figure 2. Assumptions on the phylogenetic relationship in different options of MAFFT. (A) mafft-profile, (B) ‒‒addprofile, (C), misuse of mafft-profile, and (D) ‒‒add or –addprofile.

In figure 2A, it displays the subprogram of MAFFT called “mafft-profile”, which converts two separate alignments to profiles and then aligns those two profiles assuming that they are phylogenetically isolated from each other. However, reckless handling of this subprogram can cause problematic misalignments.

In MAFFT version 7, the latest version of MAFFT, has various options for various alignment strategies, such as progressive methods and --addprofile. --addprofile option can accept two existing alignments and predicts a phylogenetic relationship. In Figure 2B, the two alignments, alignment 1 and alignment 2, are determined whether they can form a monophyletic cluster or not. If alignment 1 can form a monophyletic cluster, it can be placed in any phylogenetic position in alignment 2. However, if alignment 1 can’t form a monophyletic cluster, --add option is available.

Other than rapid algorithms and parallelization approach, “existing alignment” is by far the most efficient and useful approach. Existing alignment algorithm uses already existing aligned and annotated sequences, which is fairly short(consisting up to ~1000), to build a larger MSA containing newly sequenced data. This method is much more efficient in that it does not rebuild the entire MSA using the ungapped sequences and it is also stable even with the low-quality sequences from errors or misassemblies (often biologically important information is contained in low-quality sequences). To make the final MSA less affected by these low-quality sequences, one can select reliable sequences to first build a backbone MSA.

Figure 2C illustrates the misuse of the existing alignment. The mafft-profile option described in figure 1A is inappropriate for adding new sequences because it assumes a phylogenetic relationship.

In order to overcome this limitation of profile alignment, a new option that adds the unaligned sequences to an existing MSA was needed. Figure 2D exhibits the --add option, which assumes that each new sequence derived from a tree of an existing alignment. This option calculates the alignment at the nodes where the descendants are present.

Conclusion

With an increase in a large sequence database, multiple sequence alignment (MSA) plays an important role in evolutionary analyses of biological sequences. Therefore, various multiple sequence alignment software is necessary to effectively analyze a large database. The two most useful sequence alignment software—Clustal and MAFFT—are addressed in this paper. Clustal is software used actively in the late 20th century. It includes multiple versions such as Clustal V, Clustal X, and etc. Another software MAFFT has been developed recently with numerous new features such as progressive alignment.

Although Clustal and MAFFT provide numerous benefits to conduct multiple sequence alignment, there are certain drawbacks to the method. T-coffee is another multiple sequence alignment software that uses progressive alignment. Though MAFFT and Clustal may be the fastest compared to T-coffee, they require more memory in order to run. Both Clustal and MAFFT have been optimized for the GPU(graphics processing unit) system. GPU is an electronic circuit used in computing technology. GPU is known to be more powerful than the CPU(Central Processing Unit) because it can do thousands of operations at once. However, one main drawback of GPU is the simplified structure of GPU chips. This leads to many programming constraints. Clustal uses the progressive alignment method, which makes multiple alignments with all of the sequences in the set. However, it does not use all of the information the sequences contain. This could be considered inefficient if the sequences do have important information. In addition, Clustal W does not eliminate the gaps in between the sequences already aligned.

Therefore, the alignment produced by Clustal W is heavily influenced by the input alignment.

There will be myriads of challenges in the future when it comes to utilizing multiple sequence alignment, including dealing with large sequences, integrating large amounts of experimental data, accurately aligning non-coding and non-transcribed sequences, and integrating alternative methods. Thereby, knowledgeable and practical solutions are necessary to pave the way to enhance alignment construction in future evolutionary biology studies.

In order to mitigate some of the drawbacks, several ways include using a better guide tree and combining the existing multiple sequence alignment methods in the future. Progressive sequence alignment, one of the most commonly used multiple sequence alignment methods, uses a guide tree to progressively align the sequences according to the topology of the tree. Using a better guide tree, as well as combining the advantages of the existing MSA methods will result in higher accuracy.

Reference

1. Wang, Y., Wu, H., & Cai, Y. (2018). A benchmark study of sequence alignment methods for protein clustering. *BMC bioinformatics*, *19*(19), 529.
2. Bray, N., Dubchak, I., & Pachter, L. (2003). AVID: A global alignment program. *Genome research*, *13*(1), 97-102.
3. Agrawal, A., & Huang, X. (2009, March). Pairwise statistical significance of local sequence alignment using multiple parameter sets and empirical justification of parameter set change penalty. In *BMC bioinformatics* (Vol. 10, No. S3, p. S1). BioMed Central.
4. Paul, F., Otte, J., Schmitt, I., & Dal Grande, F. (2018). Comparing Sanger sequencing and high-throughput metabarcoding for inferring photobiont diversity in lichens. *Scientific reports*, *8*(1), 1-7.
5. Katoh, K., Rozewicki, J., & Yamada, K. D. (2019). MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings in bioinformatics*, *20*(4), 1160-1166.
6. Higgins, D. G., Bleasby, A. J., & Fuchs, R. (1992). CLUSTAL V: improved software for multiple sequence alignment. *Bioinformatics*, *8*(2), 189-191.
7. Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, *30*(4), 772-780.
8. Pearson, W. R. (2013). An introduction to sequence similarity ("homology") searching. *Current protocols in bioinformatics*, *42*(1), 3-1.
9. Kemena, C., & Notredame, C. (2009). Upcoming challenges for multiple sequence alignment methods in the high-throughput era. *Bioinformatics*, *25*(19), 2455-2465.
10. Thompson, J. D., Linard, B., Lecompte, O., & Poch, O. (2011). A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives. *PloS one*, *6*(3), e18093.
11. Zhan, Q., Ye, Y., Lam, T. W., Yiu, S. M., Wang, Y., & Ting, H. F. (2015, December). Improving multiple sequence alignment by using better guide trees. In *BMC bioinformatics* (Vol. 16, No. S5, p. S4). BioMed Central.
12. Long, P., Cui, Z. H., Li, Q. Q., Xu, J. P., Zhang, C. H., Zhou, L. S., & Li, M. H. (2013). Study on identification of Astragali Radix and Hedysari Radix by PCR amplification of specific alleles. *Zhongguo Zhong yao za zhi= Zhongguo zhongyao zazhi= China journal of Chinese materia medica*, *38*(16), 2581-2585.
13. Wei, Y. C., Jiang, C., Yuan, Y., Zhao, Y. Y., Jin, Y., & Huang, L. Q. (2017). Identification of Cervus nippon, C. elaphus and their hybridize samples based on COI and SRY gene. Zhongguo Zhong yao za zhi= Zhongguo zhongyao zazhi= China journal of Chinese materia medica, 42(23), 4588-4592.