Research Article

www.dergipark.org.tr/en/pub/mediterranean

MEDITERRANEAN
AGRICULTURAL SCIENCES
(2024) 37(3): 143-146
DOI: 10.29136/mediterranean.1551310

Comparative analysis of some multiple sequence alignment tools using *Gallus gallus* COX1 sequences

Kemal ESKIOGLU, Berkant Ismail YILDIZ, Demir OZDEMIR

Akdeniz University, Faculty of Agriculture, Department of Agricultural Biotechnology, 07058, Antalya, Türkiye

Corresponding author: D. Ozdemir, e-mail: dozdemir@akdeniz.edu.tr
Author(s) e-mail: kemaleskioglu94@gmail.com, berkantyildizz@gmail.com

ARTICLE INFO

Received: September 16, 2024
Received in revised form: November 4, 2024
Accepted: November 14, 2024

***Keywords*:**

Multiple sequence alignment
ClustalW
Clustal Omega
MUSCLE
MAFFT

ABSTRACT

Multiple Sequence Alignment (MSA) is an essential method in bioinformatics for detecting conserved sequence regions and deducing evolutionary relationships. However, performance variability exists among MSA tools, and different tools yield varying results depending on the dataset. This study conducts a comparative evaluation of four widely used MSA tools: ClustalW, Clustal Omega, MUSCLE, and MAFFT. The alignment quality and processing efficiency of these tools were assessed using 40 randomly selected *Gallus gallus* cytochrome c oxidase subunit 1 (COX1) DNA sequences. The findings offer valuable insights into the specific contexts in which these tools may be most effective. MAFFT demonstrated a notable advantage in processing speed, while Clustal Omega and MAFFT excelled in Column Score (CS). For Total Consensus (TC) score, ClustalW and MUSCLE showed superior performance, and Clustal Omega exhibited the highest performance based on Root Mean Square Deviation (RMSD) values. No significant difference was observed between the tools in terms of the Sum-of-Pairs (SP) score. This study serves as a valuable resource for researchers seeking to optimize the use of MSA tools for their specific applications.

1. Introduction

Multiple Sequence Alignment (MSA) is a cornerstone technique in bioinformatics, widely employed for identifying conserved sequence regions, elucidating evolutionary relationships, and predicting the structure and function of biological macromolecules (Levasseur et al. 2008; Pervez et al. 2014). With the advent of advanced sequencing technologies, the accumulation of vast amounts of DNA data has made MSA an indispensable tool across various research fields. In particular, MSA has become increasingly important in plant and animal biotechnology, where it aids in studying genetic diversity and enhancing disease resistance in plant and animal genomes (Ferrer-Costa et al. 2005; Prykhozhij 2015; Chowdhury and Garai 2017; Park et al. 2017).

However, the effectiveness of MSA heavily depends on the performance of the alignment tools used. Over time, several MSA algorithms have been developed, each employing different strategies to align sequences. One of the earliest and most widely used tools, ClustalW (www.clustal.org), follows a progressive alignment approach (Thompson et al. 1994), though it can produce suboptimal alignments due to the “once a gap, always a gap” issue, particularly with highly divergent sequences. Clustal Omega (www.clustal.org), an improved version, addresses some of these limitations by offering faster and more accurate alignments, especially for large datasets (Sievers and Higgins, 2018). MUSCLE (www.drive5.com) utilizes an iterative strategy that refines the guide tree and alignment to improve accuracy progressively, while MAFFT incorporates Fast Fourier Transform (FFT) to balance speed and accuracy (Katoh et al. 2002).

Despite the diversity of available tools, comparative studies reveal that no single method consistently outperforms others across all datasets and scenarios (Nuin et al. 2006; Aniba et al. 2010; Thompson et al. 2011; Pais et al. 2014). This variability necessitates a careful evaluation and selection of MSA tools based on the specific requirements of each study. To address this need, our study conducts a comprehensive evaluation of four prominent MSA tools: ClustalW, Clustal Omega, MUSCLE, and MAFFT, using COX1 DNA sequences from 40 randomly selected *Gallus gallus* specimens. By assessing alignment accuracy and computational efficiency across various metrics, including Time, Sum-of-Pairs (SP) score, Column Score (CS), Total Consensus (TC) score, and Root Mean Square Deviation (RMSD), we aim to provide a detailed analysis of each tool's strengths and weaknesses. This comparison will offer valuable guidance for researchers, helping them select the most appropriate MSA tool for their specific applications, thereby enhancing the reliability and reproducibility of bioinformatics analyses.

2. Material and Method

2.1. Dataset selection

In this study, DNA sequences from *Gallus gallus* were utilized. A dataset of 40 randomly selected COX1 (1548 bp)

(Gene ID: 807639) DNA sequences was obtained from the NCBI (National Center for Biotechnology Information) database. These sequences, which vary in length and genetic variation, were chosen to provide a comprehensive evaluation of multiple sequence alignment (MSA) tool performance.

2.2. Alignment tools and parameters

The DNA sequences were aligned using four MSA tools: ClustalW, Clustal Omega, MUSCLE, and MAFFT, all of which were integrated into the Python environment through relevant libraries and packages. ClustalW and MUSCLE were implemented using the Biopython (v1.78) library, while Clustal Omega and MAFFT were accessed via Bioconda.

ClustalW: Alignments were performed using the stepwise alignment method, with default parameters set for gap opening and extension costs at 10 and 0.1, respectively.

Clustal Omega: This tool utilized the incremental alignment method, designed for fast and accurate alignments. The gap opening and extension costs were also set to 10 and 0.1, respectively.

MUSCLE: Iterative alignment was employed, with 16 iterations by default. The gap opening cost was set to 1.0.

MAFFT: Fast Fourier Transform (FFT)-based alignment was used, with gap opening and extension costs set to 1.0 and 0.1, respectively.

2.3. Alignment process and evaluation

To assess the alignment quality, several metrics were calculated using custom Python scripts.

Sum-of-Pairs (SP) Score: SP scores were computed by summing the pairwise sequence similarities, and these were compared against reference alignments to evaluate accuracy.

Column Score (CS): This metric assessed the accuracy of each aligned column by determining whether the reference alignment columns were correctly aligned.

Total Column (TC) Score: The TC score was calculated by determining the proportion of sequences aligned in the same position across each column.

Root Mean Square Deviation (RMSD): RMSD was calculated using the scipy.spatial.distance module from SciPy (v1.10.0), measuring the average squared structural deviations between aligned sequences.

Computation Time: The completion time for each alignment and analysis process was recorded in minutes using the “time” and “timeit” Python modules. The efficiency of each tool was assessed by comparing these processing times.

All performance evaluations were automated through Python scripts, and the resulting data was analyzed directly to compare the efficiency and accuracy of each alignment tool.

3. Results and Discussion

In this study, Time, SP, CS, TC, and RMSD metrics were compared to assess the performance of multiple sequence alignment (MSA) tools, including ClustalW, Clustal Omega, MUSCLE, and MAFFT. The "Time" metric, measured in minutes, represents the duration of the alignment process for each tool. Among the tools, MAFFT demonstrated the fastest processing time at 0.6 minutes, while ClustalW, MUSCLE, and Clustal Omega required 6.62, 5.97, and 4.05 minutes, respectively. The SP score, which measures alignment accuracy, was identical across all tools at 966.250 indicating similar performance in this aspect. The CS metric, which evaluates alignment quality, indicated that Clustal Omega and MAFFT had the lowest CS values (0.03), whereas ClustalW and MUSCLE exhibited slightly higher values (0.04). Lower CS values generally correspond to better alignment quality. The TC metric, reflecting the total number of columns in the alignment, was 641 for both Clustal Omega and MAFFT, while ClustalW and MUSCLE yielded 630 columns, suggesting variations in the granularity of alignment results across tools. Overall, MAFFT stood out for its rapid processing time, and both Clustal Omega and MAFFT delivered superior alignment quality with lower CS values. While the tools showed comparable performance in terms of alignment accuracy, as indicated by identical SP scores, differences were observed in processing time and the detail of alignment results (Figure 1).

When comparing the tools based on RMSD, the performances of ClustalW, Clustal Omega, MUSCLE, and MAFFT varied significantly (Figure 2). RMSD is a metric that quantifies the differences between aligned sequences, where lower values indicate better alignment quality. The results revealed an RMSD value of 0 between ClustalW and Clustal Omega, indicating identical alignment results for these two tools. In contrast, ClustalW produced higher RMSD values compared to MUSCLE and MAFFT, with an RMSD of 9.59 between ClustalW and MUSCLE, and 15.98 between ClustalW and MAFFT. The RMSD between MUSCLE and MAFFT was 18.42, indicating that MAFFT introduces more alignment discrepancies compared to MUSCLE. Overall, these findings suggest that Clustal Omega delivers superior alignment quality, exhibiting the most consistent and lowest RMSD values across the tools. While ClustalW also demonstrated low RMSD values in certain cases, comparable to Clustal Omega, it generally showed higher values compared to other tools. The higher RMSD values observed for MUSCLE and MAFFT suggest that their alignments deviate more from the others, reflecting lower alignment quality relative to Clustal Omega.

In general, the findings from our study indicate that each multiple sequence alignment (MSA) tool presents distinct advantages. MAFFT excels in terms of processing time, while Clustal Omega and MAFFT perform better in terms of CS scores. On the other hand, ClustalW and MUSCLE perform well in terms of TC scores, with Clustal Omega also demonstrating superior performance based on RMSD. These results align with previous studies that report varying performances for different tools depending on the dataset used. Mohamed et al. (2018) compared six well-known MSA tools, including Clustal Omega, MAFFT, BROBCONS, KALIGN, RETALIGN, and MUSCLE. They found that BROBCONS outperformed the others in terms of both TC and SP scores, as well as processing time. MAFFT ranked third across these metrics, while Clustal Omega ranked lowest in terms of TC and SP scores, and fifth in processing time. This contrasts with our findings, where Clustal Omega and MAFFT performed better in terms of CS and RMSD, although not in SP scores. Sievers and Higgins (2018) classified MSA tools into two categories: those optimized for fast processing and large alignments, and those optimized for higher accuracy with fewer sequences. MUSCLE and MAFFT were cited as examples of the first group, while T-Coffee and MAFFT L-INS-i were placed in the second group. Our findings, that MAFFT is the fastest tool, are consistent with Katoh et al. (2005), who also identified MAFFT as the fastest in terms of processing time. However,

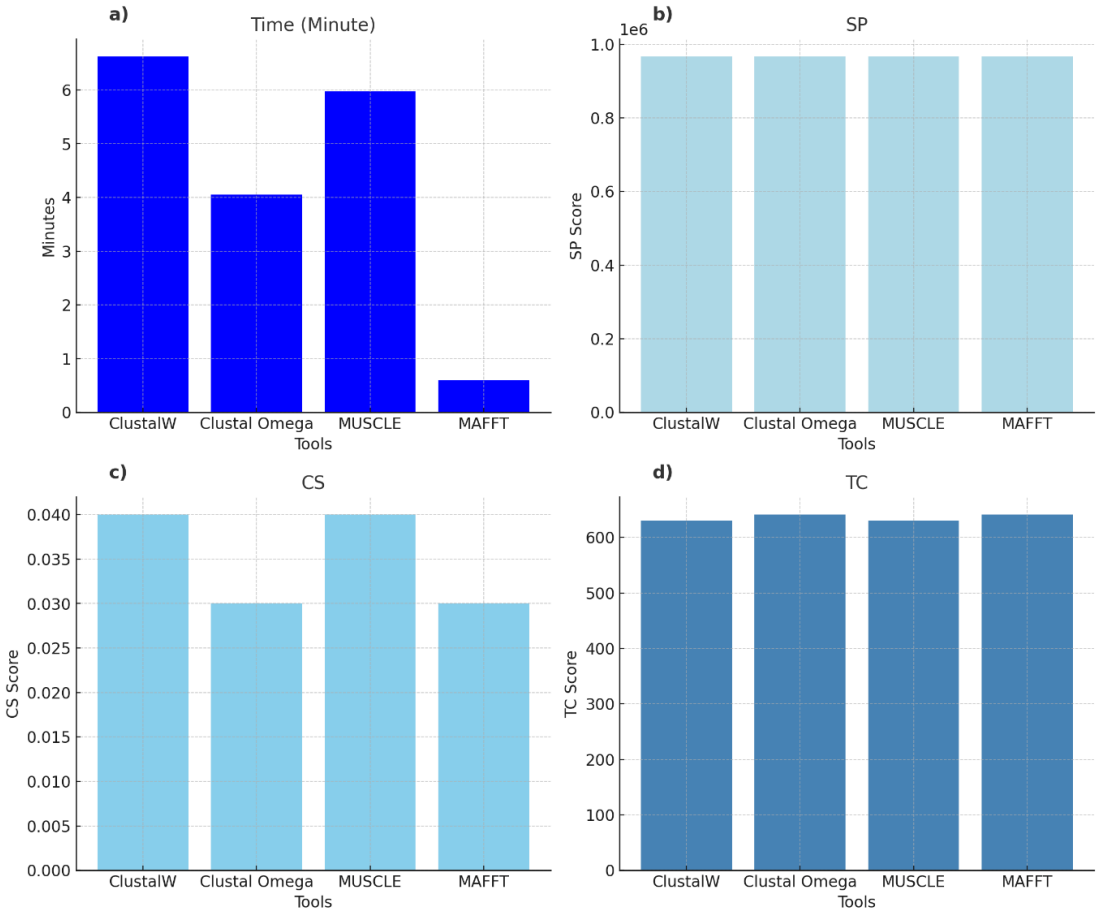

Figure 1. Performance comparison of alignment tools: a) Time b) Sum-of-Pairs (SP) c) Column Score (CS) d) TC scores (Total Consensus).

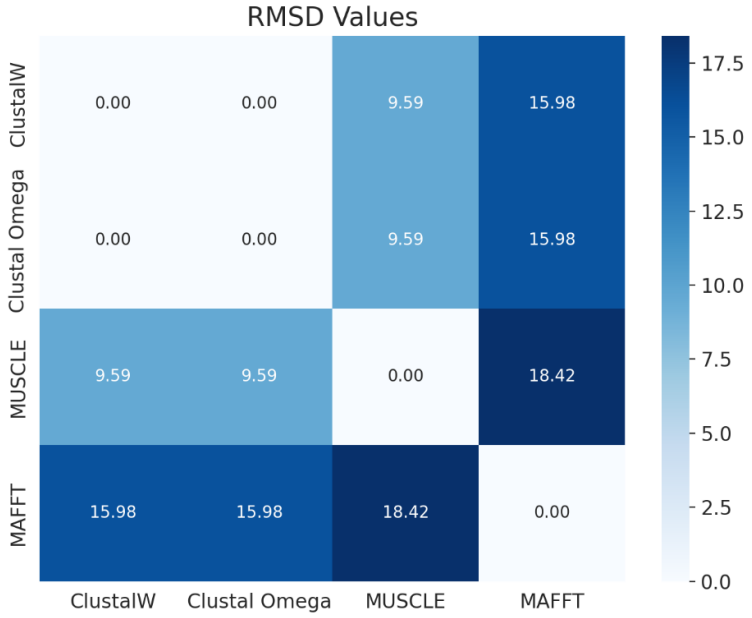

Figure 2. Pairwise values of the alignment tools.

Sievers and Higgins (2018) also noted that ClustalW2, MAFFT, and MUSCLE exhibited lower performance in terms of SP scores, while Clustal Omega slightly outperformed these tools with default settings, which is consistent with our results for CS and RMSD, though not for SP scores. In another study, Pais et al. (2014) compared the performance of ClustalW, Clustal Omega, DIALIGN-TX, MAFFT, MUSCLE, POA, Probalign, PROBCONS, and T-Coffee. They found that PROBCONS, T-

Coffee, Probalign, and MAFFT had superior accuracy, while ClustalW and MUSCLE were identified as the fastest tools. These findings reinforce the hypothesis proposed by Nuin et al. (2006), Aniba et al. (2010), Thompson et al. (2011), and Pais et al. (2014) that the performance of MSA tools is context dependent. In conclusion, the selection of an appropriate tool should be based on the specific requirements of the dataset and analysis to achieve optimal results, as each tool has its own strengths and limitations.

4. Conclusion

In this study, we evaluated the performance of four multiple sequence alignment (MSA) tools: ClustalW, Clustal Omega, MUSCLE, and MAFFT. Our findings demonstrate that each tool offers distinct advantages depending on the dataset and research context, underscoring the importance of selecting the appropriate tool for specific research needs. MAFFT emerged as the fastest tool in terms of processing time, while Clustal Omega and MAFFT outperformed the others in terms of CS score, and ClustalW and MUSCLE excelled in terms of TC score. Additionally, Clustal Omega showed superior alignment quality based on RMSD scores. In conclusion, the optimal choice of MSA tool should be made based on the characteristics of the dataset, the goals of the research, and the computational resources available. This study provides valuable insights into the comparative performance of these widely used MSA tools, helping researchers make informed decisions in tool selection. Continued development and optimization of these tools can further enhance their applicability, benefiting a wide range of fields from fundamental biological research to applied biotechnology, ultimately contributing to the more efficient utilization of biological data.

References

Aniba MR, Poch O, Thompson JD (2010) Issues in bioinformatics benchmarking: the case study of multiple sequence alignment. Nucleic Acids Research 38: 7353-7363.

Chowdhury B, Garai G (2017) A review on multiple sequence alignment from the perspective of genetic algorithm. Genomics 109: 419-431.

Ferrer-Costa C, Orozco M, de la Cruz X (2005) Use of bioinformatics tools for the annotation of disease-associated mutations in animal models. Proteins: Structure, Function, and Bioinformatics 61: 878-887.

Katoh K, Misawa K, Kuma KI, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research 30: 3059-3066.

Katoh K, Kuma KI, Toh H, Miyata T (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Research 33(2): 511-518.

Levasseur A, Pontarotti P, Poch O, Thompson JD (2008) Strategies for reliable exploitation of evolutionary concepts in high throughput biology. Evolutionary Bioinformatics 4: EBO-S597.

Mohamed EM, Mousa HM, Keshk AE (2018) Comparative analysis of multiple sequence alignment tools. International Journal of Information Technology and Computer Science 10: 24-30.

Nuin PA, Wang Z, Tillier ER (2006) The accuracy of several multiple sequence alignment programs for proteins. BMC Bioinformatics 7: 1-18.

Pais FSM, Ruy PDC, Oliveira G, Coimbra RS (2014) Assessing the efficiency of multiple sequence alignment programs. Algorithms for Molecular Biology 9: 1-8.

Park D, Park SH, Ban YW, Kim YS, Park KC, Kim NS, Kim Choi IY (2017) A bioinformatics approach for identifying transgene insertion sites using whole genome sequencing data. BMC Biotechnology 17: 1-8.

Pervez MT, Babar ME, Nadeem A, Aslam M, Awan AR, Aslam N, Hussain T, Naveed N, Qadri S, Waheed U, Shoaib M (2014) Evaluating the accuracy and efficiency of multiple sequence alignment methods. Evolutionary Bioinformatics, 10: EBO-S19199.

Prykhozhij SV, Rajan V, Gaston D, Berman JN (2015) CRISPR multitargeter: a web tool to find common and unique CRISPR single guide RNA targets in a set of similar sequences. PloS one 10: e0119372.

Sievers F, Higgins DG (2018) Clustal Omega for making accurate alignments of many protein sequences. Protein Science 27: 135-145.

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research 22: 4673-4680.

Thompson JD, Linard B, Lecompte O, Poch O (2011) A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives. PloS one, 6: e18093.