Middle East Journal of Science

https://dergipark.org.tr/mejs

MEJS

e-ISSN:2618-6136

Research Article

THREE MSA TOOLS ANALYSIS IN DNA AND PROTEIN DATASETS

***Fırat Aşır**[1] **Tuğcan Korak**[2] **Özgür Öztürk**[3]

[1]Department of Histology and Embryology, Medical School, Dicle University, Diyarbakır, Turkey
[2]Department of Medical Biology and Genetics, Medical School, Kocaeli University, Kocaeli, Turkey
[3]Pharmaceutical Biotechnology, Center for System-based Drug Research, Department of Pharmacy, Ludwig Maximilian University of Munich, Germany

* Corresponding author; firatasir@gmail.com

Abstract: *Multiple sequence alignment (MSA) is the alignment of three or more sequences of DNA, RNA, and protein. MSA is used to construct phylogenetic trees and to compare evolutionary relationships between sequences analyzing similarities and dissimilarities. A variety of multiple sequence alignment tools are available, each using different methods and parameters to align sequences. In this article three MSA tools; CLUSTALW, SAGA, and MAFFT* ***were*** *used to analyze five datasets BALiBASE_R9, DIRMBASE, SABmark, DNABali, and ProteinBali. The* ***results*** *showed that MAFFT may be more useful to align DNA and protein sequences than the other two tools.*

Keywords*: multiple sequence alignment, MAFFT, SAGA, CLUSTALW*

Received: August 17, 2021 Accepted: November 25, 2021

1. Introduction

Multiple sequence alignment (MSA) refers to the alignment of more than two DNA, RNA, or amino acid sequences to determine structural, functional, and evolutionary relationships among organisms and also to predict secondary and tertiary protein structure [1-5]. MSA may offer functional and conserved regions in a whole sequence family and illustrate evolutional history or comparison of the sequences by considering re-arrangements, insertions, deletions, or mutations [6, 7]. MSA is also adapted to next-generation sequencing methods for structural and functional comparisons [8].

Over the past decade, various MSA algorithms or strategies have been developed to perform convenient alignment of different datasets. One of them is CLUSTALW that uses progressive alignment. It starts to align the most similar sequences and then continues with less similar ones in order to obtain global alignment. It is the third generation of CLUSTAL software and has improvements on down or up weighting to similar and divergent sequences, respectively. In addition, amino acid substitution matrices are altered in accordance with the divergence of aligned sequences. Moreover, residue-specific and locally reduced gap penalties in hydrophilic sites promote new gaps in possible loop sites [9]. Sequence alignment by genetic algorithm (SAGA) is an MSA software that uses an automatic scheduling scheme to manage the usage of various operators in order to combine or mutate alignments among generations. It produces alignments in a similar manner to evolution and provides a gradual

development of fitness of the alignment populations through objective function which measures multiple alignment quality. Objective function score and comparison with reference alignments based on sequences of the known tertiary structure are the strengths of the SAGA [10]. Moreover, Multiple Alignment using Fast Fourier Transform (MAFFT) is another software that applies progressive and iterative approaches with several modifications. It provides fast recognition of homologous regions by the Fast Fourier transform algorithm and increases alignment accuracy of distantly related sequences or sequences that have large insertions, and decreases CPU time by the simplified scoring system. Although both use the progressive method MAFFT uses less CPU time than CLUSTALW with similar accuracy [11]. Besides these software, there are also various MSA approaches such as Clustal Omega [12], MUMMALS [13], ProbCons [14], T-Coffee [15], DIALIGN [16], MUSCLE [5], PROMALS3D [17], Kalign [18], M-Coffee [19], Align-M [20], PRANK [21, 22], 3DCoffee [23], Expresso [24] and HAlign [25] etc. Each of them has its own advantages and disadvantages to optimize sequence alignments, align distantly related sequences and minimize the computational time [11].

Choosing the best tool for MSA requires consideration of several aspects based on the study's scope [2]. Thus, analyzing the performance of the MSA tools with different datasets is essential to illustrate or facilitate software selection in further studies. In this study, CLUSTALW, SAGA, and MAFFT were compared by consistency, the column with gap and sum of pair scores with BAliBASE, SABmark, DIRMBASE, ProteinBali, and DNABali datasets.

2. Materials and Methods

Reference data (aligned data, data with dashes) were acquired from benchmarks BAliBASE (Version 3.0 R9), SABmark (Version 1.63), DIRMBASE (Version 1.0), and the manually constructed ProteinBali and DNABali datasets. All datasets are compatible with the following three MSA tools; CLUSTALW (Version 2.1), SAGA (Version 0.95), and MAFFT (Version 6). The reference data in this study were randomly picked up from datasets to evaluate the performance of MSA tools.

BAliBASE benchmark includes sequences that are specifically designed for MSA. It contains high-quality manually refined reference alignments by considering 3D superpositions and distinctive reference datasets with different properties. [26]. Box10, 22, 32 were acquired from BAliBASE benchmark. ProteinBali was randomly constructed from a different subset of BAliBASE benchmark and includes the following protein sequences: box46, box50, box56. DNABali (Reference Protein-Coding DNA Alignments Databases: http://dna.cs.byu.edu/mdsas/download.shtml) was randomly designed from BAliBASE benchmark and consists of the following DNA sequences: RV61_sushi_ref6, RV64_kringle_1_ref6, and RV70_photo_ref7. DIRMBASE offers locally related DNA sequences including highly conserved motifs generated by a random model of sequence evolution [27]. dna-400-30-4-0, r1-dna-400-30-4-1, r1-dna-400-30-4-2, r1-dna-400-30-4-3 and r1-dna-400-30-4-4 were selected from DIRMBASE. SABmark dataset offers MSA of protein sequences with low homology [28, 29,]. D1a6m__-d1ash, d1ash__-d1dlwa, d1dlwa_-d1ew6a, d1ew6a_-d1gtea1, d1gtea1-d1gvha1 were selected from SABmark.

In order to perform MSA tools, reference data were converted to FASTA format using Jalview software. Unaligned (undashed) data were converted from Multiple Sequence Format (MSF) to FASTA format by using Jalview (only those that were not FASTA format). Each data was individually uploaded to SAGA, CLUSTALW, and MAFFT. For SAGA and DNABali (DNA datasets), data were converted

into protein sequences, then uploaded to tools. In each software, some parameters were changed in order to obtain the highest scores. Input parameters were adjusted according to the dataset either DNA or protein. Results were recorded in FASTA format (if available), otherwise in clustal.aln format. Clustal.aln formats were converted to FASTA format by Jalview. Firstly, reference data, then FASTA formatted data was uploaded to SuiteMSA. For each individual data, consistency, the sum of pair scores, and column score with gaps were recorded. Scores attained from Suit MSA were arranged in a table according to datasets and tools (Figure 1). Regarding all scores, tables and graphs were constructed for each individual dataset by SAGA, CLUSTALW, and MAFFT. Mean values of data were considered in the comparison of tools.

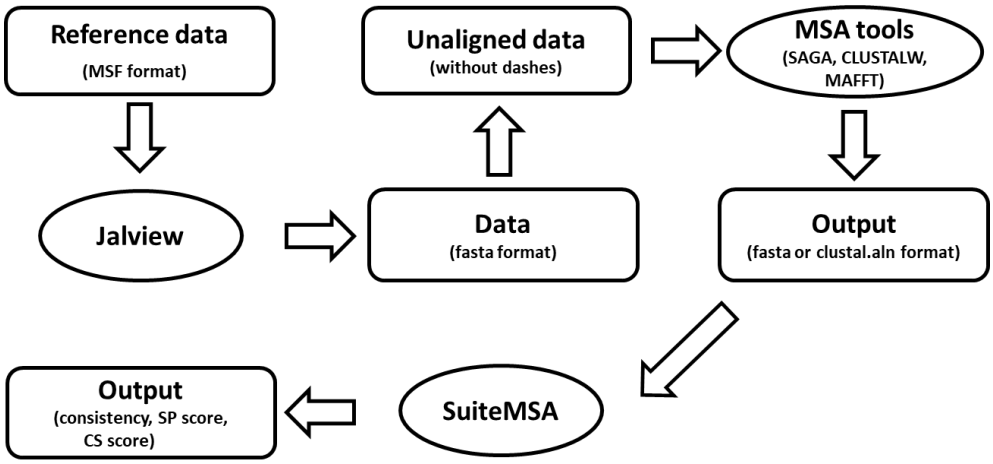

Figure 1: A roadmap of data conversion and processing for sequences to be aligned with MSA tools. Rectangles stand for data, the circle represents software.

3. Results

For each dataset, scores from SUITMSA were recorded and a graph corresponding to each data was constructed regarding CLSUTALW, MAFFT, and SAGA. Table 1 shows all data exploited in this study. From BALiBASE benchmark, box10, box22, and box32 protein data were used and their all scores from each MSA tools were recorded. From DIRMBASE benchmark, dna0, dna1, dna2, and dna3 nucleotide data were used and SuitMSA comparator results were shown. In SABmark, randomly five pairwise sequences (d1h97a_-d1irdb_, d1irdb_-d1itha_, d1itha_-d1jboa_, d1jboa_-d1ngka_, d1ngka_-d1qlab1_) were chosen and labeled as 7-8, 8-9, 9-10, 10-11 and 11-12. DNABali and ProteinBali datasets were chosen as same way in SABmark.

Table 1. Show all five datasets (BALiBASE, DIRMBASE, SABmark, DNABali, and ProteinBali) and three tools (CLUSTALW, MAFFT, and SAGA) and scores of tools according to each data set.

		CLUSTALW			MAFFT			SAGA		
Data set	**Individual data**	**Const (%)**	**SOP**	**CS w gap**	**Const (%)**	**SOP**	**CS w gap**	**Const (%)**	**SOP**	**CS w gap**
BAliBASE	Box10	15.416	0.411	0.044	6.282	0.538	0.064	9.081	0.515	0.071
	Box22	11.377	0.405	0.003	14.379	0.509	0.095	10.256	0.494	0.062
	Box32	19.981	0.392	0.187	21.978	0.462	0.213	17.325	0.433	0.184
	mean	**15.591**	**0.403**	**0.078**	**14.213**	**0.503**	**0.124**	**12.221**	**0.481**	**0.106**
DIRMBASE	Dna0	1.502	0.016	0.000	3.953	0.233	0.000	1.590	0.043	0.011
	Dna1	1.059	0.021	0.000	6.877	0.123	0.023	2.602	0.034	0.007
	Dna2	1.302	0.030	0.000	19.908	0.438	0.202	1.294	0.125	0.010
	Dna3	0.866	0.003	0.000	6.212	0.076	0.030	2.143	0.037	0.012
	mean	**1.182**	**0.018**	**0.000**	**9.238**	**0.218**	**0.064**	**1.907**	**0.060**	**0.010**
SABmark	7_8	38.667	0.419	0.378	52.469	0.581	0.573	63.432	0.681	0.676
	8_9	67.114	0.689	0.674	63.758	0.674	0.659	65.126	0.734	0.748
	9_10	13.58	0.206	0.149	44.809	0.451	0.383	46.123	0.487	0.409
	10_11	9.877	0.000	0.000	8.287	0.000	0.000	12.019	0.000	0.000
	11_12	10.145	0.000	0.000	18.71	0.000	0.000	16.540	0.000	0.000
	mean	**27.877**	**0.263**	**0.240**	**37.607**	**0.341**	**0.323**	**40.648**	**0.380**	**0.367**
DNABali	Kringle	2.864	0.247	0.000	4.057	0.233	0.000	2.150	0.205	0.000
	Photo	21.460	0.774	0.0172	27.556	0.832	0.254	5.551	0.583	0.059
	Sushi	38.669	0.093	0.000	38.834	0.097	0.000	30.942	0.037	0.000
	mean	**20.998**	**0.371**	**0.006**	**23.482**	**0.387**	**0.085**	**12.881**	**0.275**	**0.020**
ProteinBali	Box46	5.087	0.356	0.000	5.089	0.514	0.000	4.023	0.624	0.000
	Box50	20.186	0.678	0.193	54.381	0.828	0.503	35.984	0.184	0.346
	Box56	3.423	0.512	0.024	26.277	0.619	0.236	16.671	0.503	0.190
	mean	**9.565**	**0.515**	**0.072**	**28.582**	**0.654**	**0.246**	**18.893**	**0.437**	**0.179**

Const: Consistency; SOP: sum of pair scores. CS w gap stands for column score with gaps

3.1. BALIBASE Results

Mean consistency scores of CLUSTALW (15.591%) were better than MAFFT's (14.213%) and SAGA's scores (12.221**%**) (Figure 2a). For any tool, the sum of pair cores was close to each other, meaning that individual data differences were not reflected in the sum of pair scores thus putting tools at the forefront rather than data in scoring (Figure 2b). MAFFT performed the best column scores (Figure 2c).

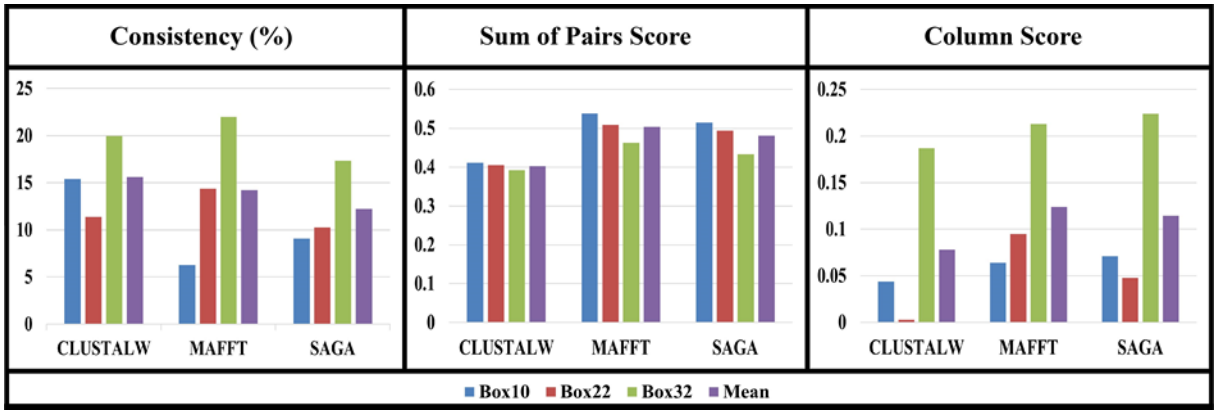

Figure 2. Graphical representation of a) consistency, b) sum of pairs score, c) column score with gaps of BALIBASE data with CLUSTALW, MAFFT, and SAGA.

3.2. DIRMBASE Results

Figure 3a shows the consistency (%) of individual DIRMBASE datasets and means by CLUSTALW, MAFFT, and SAGA. MAFFT resulted in the highest consistency score (7.238%). Figure 3b shows the sum of pair scores of three MSA tools. CLUSTALW rendered the lowest score, while MAFFT performed the highest score as in consistency. Column scores with all tools were very close to zero but MAFFT got a relatively better score than the other two tools (Figure 3c).

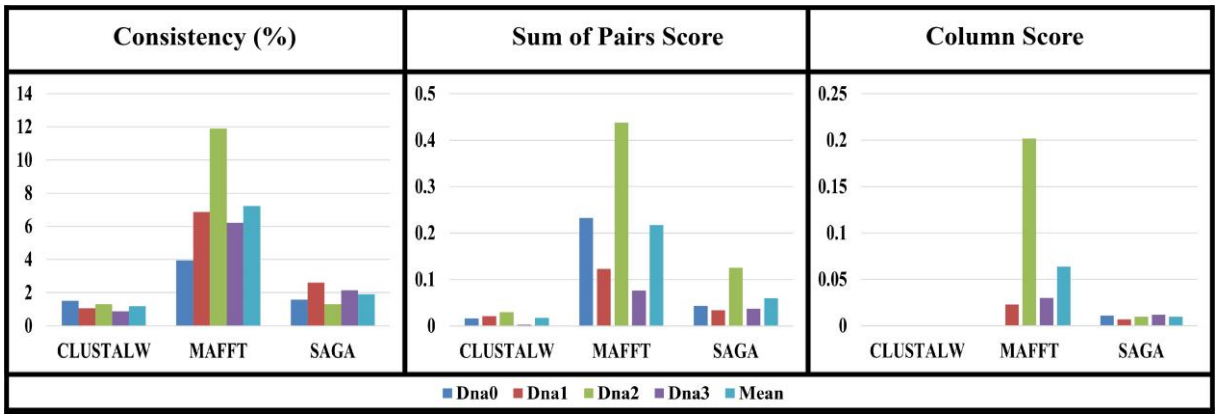

Figure 3. Graphical representation of a) consistency, b) sum of pairs score, c) column score with gaps results of DIRMBASE data with CLUSTALW, MAFFT and SAGA.

3.3. SABmark Results

Higher scores were recorded in SABmark dataset than scores in BALiBASE and DIRMBASE. Although higher than scores of CLUSTALW, consistency scores of SAGA and MAFFT were close to each other (Figure 4a). Figure 4b shows the sum of pair scores with three MSA tools. SAGA and MAFFT resulted in higher scores than CLUSTALW. SAGA's score was slightly higher than MAFFT's score, making SAGA the best tool according to the sum of pair scores. Among tools, SAGA and MAFFT recorded similar results which were higher than that of CLUSTALW (Figure 4c).

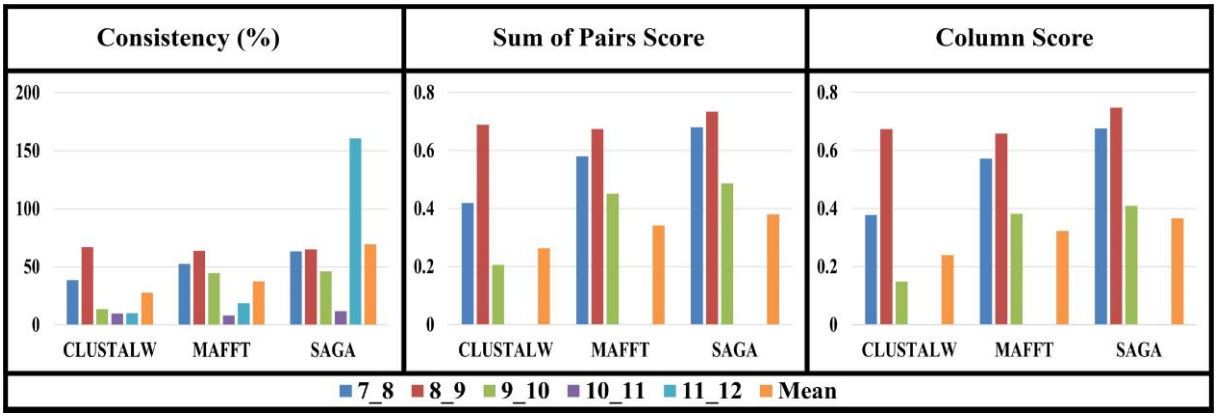

Figure 4. Graphical representation of a) consistency, b) sum of pairs score, c) column score with gaps results of SABmark data with CLUSTALW, MAFFT and SAGA.

3.4. DNABali Results

Figure 5a shows consistent results for three MSA tools based on DNABali dataset. The highest mean consistency score was recorded by MAFFT while SAGA performed the lowest score. Figure 5b shows the sum of the pair score of individual data by tools. SAGA got the lowest sum of pair scores while MAFFT rendered the highest score. MAFFT generated the highest column sores with a gap, which was not statistically significant (Figure 5c).

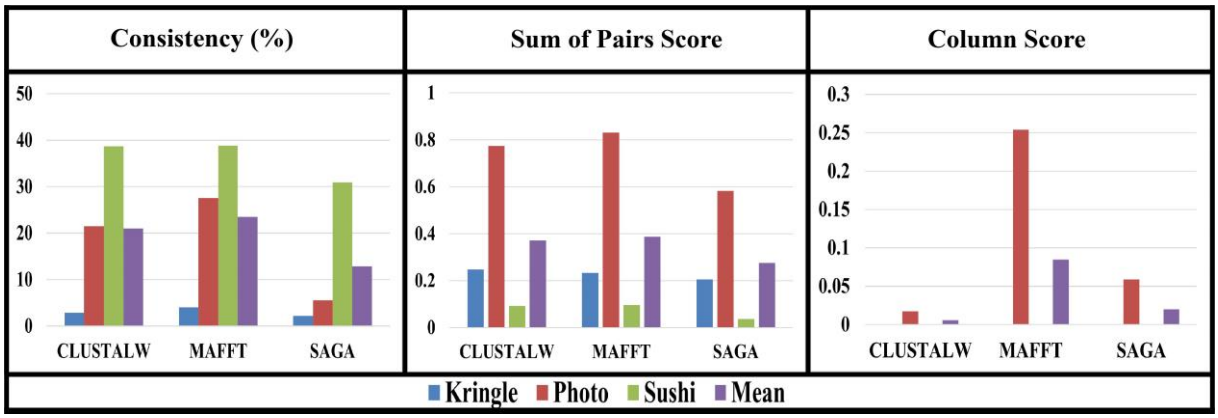

Figure 5. Graphical representation of a) consistency, b) sum of pairs score, c) column score with gaps results of DNABali data with CLUSTALW, MAFFT and SAGA.

3.5. PROTEINBali Results

Figure 6a shows consistent results for three MSA tools based on PROTEINBali dataset. The highest mean consistency score was recorded with MAFFT while the score of ClustalW was the lowest value. Regarding the sum of pair scores, MAFFT generated the highest mean value among others while SAGA resulted in the lowest value (Figure 6b). The highest score was recorded with MAFFT while the lowest score was recorded with ClustalW (Figure 6c).

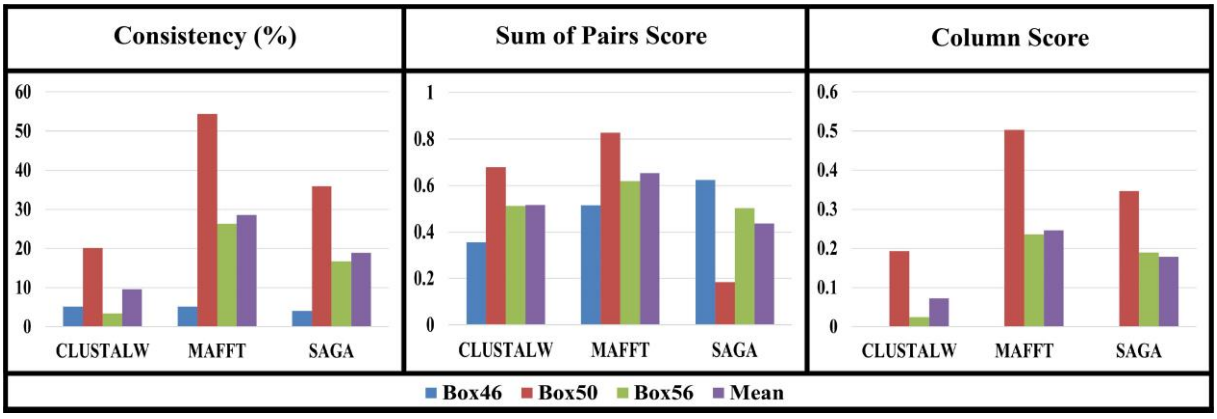

Figure 6. Graphical representation of a) consistency, b) sum of pairs score, c) column score with gaps results of PROTEINBali data with CLUSTALW, MAFFT and SAGA.

We constructed Table 2 to show each tool's performance by consistency (%), SP score, and CS score with each dataset. MAFFT performed best scores in most cases.

Table 2. Arrangement of MSA tools by consistency, SP and CS scores from best to worst performance.

	Consistency (%)			SP score			CS score		
BALIBASE	CLUSTALW (15.591)	MAFFT (14.213)	SAGA (12.221)	MAFFT (0.503)	SAGA (0.481)	CLUSTALW (0.403)	MAFFT (0.124)	SAGA (0.106)	CLUSTALW (0.078)
DIRMBASE	MAFFT (9.238)	SAGA (1.907)	CLUSTALW (1.182)	MAFFT (0.218)	SAGA (0.060)	CLUSTALW (0.018)	MAFFT (0.064)	SAGA (0.010)	CLUSTALW (0.000)
SABmark	SAGA (40.648)	MAFFT (37.607)	CLUSTALW (27.877)	SAGA (0.380)	MAFFT (0.341)	CLUSTALW (0.263)	SAGA (0.367)	MAFFT (0.323)	CLUSTALW (0.240)
DNABali	MAFFT (23.482)	CLUSTALW (20.998)	SAGA (12.881)	MAFFT (0.387)	CLUSTALW (0.371)	SAGA (0.275)	MAFFT (0.085)	SAGA (0.020)	CLUSTALW (0.006)
ProteinBali	MAFFT (28.582)	SAGA (18.893)	CLUSTALW (9.565)	MAFFT (0.654)	CLUSTALW (0.515)	SAGA (0.437)	MAFFT (0.246)	SAGA (0.179)	CLUSTALW (0.072)

4. Discussion

Multiple sequence alignment (MSA) is an algorithm used for identifying shared regions of homology, determination of the consensus sequence, predicting the secondary and tertiary structures, and constructing a phylogenetic tree in mathematics and bioinformatics [28]. MSA algorithms align three or more sequences of DNA, RNA, or protein with different ways such as dynamic programming, progressive alignment construction, iterative methods, hidden Markov models, and genetic algorithms and simulated annealing techniques [29]. MSA tools consume a lot of CPU time to give the best results, and the processing time can increase quadratically depending on the length and number of sequences. Due to increasing biological data, the performance of MSA algorithms is gaining importance. In multiple alignments, consistency, the sum of pair score (PS), and column score (CS) are popular parameters used to analyze the performance of MSA tools [30, 31]. These scores are useful when reference alignment of the same sequences is available. Consistency shows the percentage of identical columns in the test alignments and the columns of reference alignments [32]. The SP score is generated by comparing identically aligned residue pairs in the test and the reference alignments. SP score is calculated based on the test alignments and a determinant parameter to analyze the performance of the tool in the sequence alignment. The SP score equals score 1 when test alignment is identical to reference alignment while zero score means no identity between alignments. CS calculates scores the fraction of identically aligned positions. 2.3 CS score is calculated according to the individual comparison of columns and the calculated score then is divided into the number of columns analyzed. Of these, SP scores is a determinant of tool quality, while other scores are supporting SP score [33, 34].

In our study, consistency, SP score, and CS score were recorded to evaluate the quality and reliability of three MSA tools. We analyzed CLUSTALW, MAFFT and SAGA performance on five datasets from benchmarks BALiBASE (protein-based), DIRMBASE (nucleotide-based), SABmark (protein-based) and manually constructed DNABali (nucleotide-based) and ProteinBali (protein-based) (Table 1). BALIBASE is manually constructed and known to include a high-quality sequence of proteins with linear motifs (protein interaction sites, cell compartment targeting signals, post-translational modification sites, or cleavage sites). BALIBASE includes protein families sequences that are disordered and hard to align by conventional multiple sequence alignment algorithms [35]. MAFFT resulted in higher performance in SP and CS scores than CLUSTALW and SAGA, while CLUSTALW was best in consistency for BALIBASE dataset, however, consistency of MAFFT was close to CLUSTALW's consistency (Table 1 and 2, Figure 2). DIRMBASE is a database that contains highly

conserved motives and locally related sequences and is used for local alignments [36]. In DRIMBASE dataset, MAFFT outperformed the other tools by far for each parameter (Table 1 and 2, Figure 3). All tools got their highest score in consistency, SP score, and CS scores in SABmark dataset (Table 1 and 2, Figure 4). The reason for that is SABmark includes sequences with up to 50% similarity because over this region makes programs perform better since it is easy to align [37]. Also, pairwise alignment was opted for SABmark dataset, this improved scores for all tools. SAGA got higher performance than CLUSTALW and MAFFT in consistency, SP and CS scores in SABmark dataset, but there was no big gap between SAGA's and MAFFT's performances. For DNABali and ProteinBali dataset, MAFFT was far better than other tools for each parameter (Table 1 and 2, Figure 5 and 6, respectively).

By analyzing the performance of tools with consistency, we observed that MAFFT got the highest performance in 3 out of 5 datasets, while SAGA and CLUSTALW resulted in once. For the SP scores, which is a very important indicator for the effectiveness of an MSA **tool**, MAFFT performed the higher score than CLUSTALW and SAGA 4 out of 5 datasets. Likewise, CS score also revealed that MAFFT's performance was better than other tools in the 4 datasets. We have analyzed three popular MSA tools in five different datasets ~~with different~~. We conclude that MAFFT was better than CLUSTALW and SAGA to align multiple sequences of DNA and protein families.

5. Conclusion

Our results showed that MAFFT seems to be a better tool for both DNA and protein sequence**s** alignment than CLUSTALW and SAGA. To get the best alignment, the type of sequence and the tool specific to data should be picked up, otherwise, it may give false or non-optimal results. Tool requirements and parameters should not be ignored during multiple alignments.

5.1. Supplementary information about tools

CLUSTALW and its parameters:

Online webserver: http://www.genome.jp/tools/clustalw/

For protein data, protein parameter with BLOSUM matrix, for nucleotide data DNA parameter with IUB matrix was selected. The slow/accurate option was used to get more accurate results. Gap open penalty and gap extension value were adjusted to 30.00 and 6.00, relatively. For additional options, -OUTORDER=INPUT was used. Output format was selected as FASTA. The remaining setting was set to default.

MAFFT and its parameters:

Online webserver: https://www.ebi.ac.uk/Tools/msa/mafft/

For protein data, protein parameter with BLOSUM62 matrix, for nucleotide data nucleic acid parameter with none matrix was selected. Gap open penalty and gap extension values were adjusted to 3.00 and 0.5, respectively. Maximum iteration number was set to 100, FFT was local pair, ORDER was input, output was FASTA format. Rest was adjusted as default.

SAGA and its parameters:

Online webserver: http://rsdb.csie.ncu.edu.tw/tools/msa.htm

For protein data, protein and nucleic acid (as converted to protein sequence) parameter with PAM250 matrix (SAGA runs only for protein) has opted. Gap open penalty and gap extension value was adjusted to 8.00 and 12, relatively. Output was saga_aln. Settings remaining are left to default.

In order to convert DNA sequences to protein sequences at default settings. We use http://web.expasy.org/translate/

For SuiteMSA, we download SuiteMSA package (v1.3.22B) [zip file]
http://bioinfolab.unl.edu/~canderson/SuiteMSA/

Acknowledgment

We would like to thank Jens ALLMER for the conceptualization and permission for this study to be published. This study has not received any funding.

Conflict of interest:

The authors declare no conflict of interest.

The Declaration of Ethics Committee Approval

The author declares that this document does not require an ethics committee approval or any special permission. Our study does not cause any harm to the environment.

Authors' Contributions:

All authors read and approved the final manuscript. All authors mentioned in the paper have significantly contributed to the research:

F.A.: Conceptualization, Investigation, Software, Validation, Visualization, Project administration, Supervision, Writing – review and editing (%35)

T.K.: Conceptualization, Investigation, Methodology, Visualization, Writing – original draft, Writing – review, and editing (%35)

Ö.Ö.: Investigation, Methodology, Validation, Visualization, Data curation Visualization, Writing – original draft, Writing – review and editing (%30)

References

[1] Notredame, C. "Recent Evolutions of Multiple Sequence Alignment Algorithms", *PLOS Computational Biology*, 3(8), e123, 2007.

[2] Edgar, R.C., Batzoglou, S. "Multiple sequence alignment", *Current opinion in structural biology*, 16(3), 368-373, 2006.

[3] Moretti, S., et al. "The M-Coffee web server: a meta-method for computing multiple sequence alignments by combining alternative alignment methods", *Nucleic Acids Research*, 35(Web Server issue), W645-8, 2007.

[4] Chowdhury, B., Garai, G. "A review on multiple sequence alignment from the perspective of genetic algorithm", *Genomics*, 109(5), 419-431, 2017.

[5] Edgar, R.C. "MUSCLE: a multiple sequence alignment method with reduced time and space complexity", *BMC Bioinformatics*, 5, 113-31, 2004.

[6] Kumar, S., Filipski, A. "Multiple sequence alignment: in pursuit of homologous DNA positions", *Genome Research*, 17(2), 127-35, 2007.

[7] Chatzou, M., et al. "Multiple sequence alignment modeling: methods and applications", *Briefings in Bioinformatics,* 17(6), 1009-1023, 2016.

[8] Bawono, P., et al. "Multiple Sequence Alignment", *Methods Mol Biol*, 1525, 167-189, 2017.

[9] Thompson, J.D. et al. "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(22), 4673-80, 1994.

[10] Notredame, C, Higgins, D.G. "SAGA: Sequence Alignment by Genetic Algorithm", *Nucleic Acids Research*, 24(8), 1515-1524, 1996.

[11] Katoh, K., et al. "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform", *Nucleic Acids Research*, 30(14), 3059-66, 2002.

[12] Sievers, F., et al. "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega", *Molecular Systems Biology*, 7, 539-44, 2011.

[13] Pei, J., Grishin, N.V. "MUMMALS: multiple sequence alignment improved by using hidden Markov models with local structural information", *Nucleic Acids Research*, 34(16), 4364-4374, 2006.

[14] Do, C.B., et al. "ProbCons: Probabilistic consistency-based multiple sequence alignment", *Genome Research*, 15(2), 330-40, 2005.

[15] Notredame, C., et al. "T-Coffee: A novel method for fast and accurate multiple sequence alignment", *Journal of Molecular Biology*, 302(1), 205-17, 2000.

[16] Morgenstern, B., et al. "DIALIGN: finding local similarities by multiple sequence alignment", *Bioinformatics*, 14(3), 290-4, 1998.

[17] Pei, J., et al. "PROMALS3D: a tool for multiple protein sequence and structure alignments", *Nucleic Acids Research*, 36(7), 2295-300, 2008.

[18] Lassmann, T., Sonnhammer, E.L.L. "Kalign – an accurate and fast multiple sequence alignment algorithm", *BMC Bioinformatics*, 6(1), 298-306, 2005.

[19] Wallace, I.M., et al. "M-Coffee: combining multiple sequence alignment methods with T-Coffee", *Nucleic acids research*, 34(6), 1692-1699, 2006.

[20] Van Walle, I., et al. "Align-m--a new algorithm for multiple alignment of highly divergent sequences", *Bioinformatics*, 20(9), 1428-35, 2004.

[21] Löytynoja, A., Goldman, N. "Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis", *Science*, 320(5883), 1632-5, 2008.

[22] Löytynoja, A., Goldman, N. "An algorithm for progressive multiple alignment of sequences with insertions", *Proceedings of the National Academy of Sciences of the United States of America*, 102(30), 10557-62, 2005.

[23] O'Sullivan, O., et al. "3DCoffee: Combining Protein Sequences and Structures within Multiple Sequence Alignments", *Journal of Molecular Biology*, 340(2), 385-395, 2004.

[24] Armougom, F., et al. "Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee", *Nucleic acids research*, 34(Web Server issue), W604-W608, 2006.

[25] Zou, Q., et al. "HAlign: Fast multiple similar DNA/RNA sequence alignment based on the centre star strategy", *Bioinformatics*, 31(15), 2475-81, 2015

[26] Pais, F.S., Ruy, P.C., Oliveira, G. and Coimbra, R.S. "Assessing the efficiency of multiple sequence alignment programs", *Algorithms for Molecular Biology*, 9(1), 4-11, 2014.

[27] Subramanian, A.R., et al. "DIALIGN-TX: greedy and progressive approaches for segment-based multiple sequence alignment", *Algorithms for Molecular Biology*, 3, 6-16, 2008.

[28] Carillo, H., Lipman, D.J. "The multiple sequence alignment problem in biology", *SIAM Journal on Applied Mathematics*, 48, 1073–1082, 1988.

[29] Daugelaite, J., et al. "An Overview of Multiple Sequence Alignments and Cloud Computing in Bioinformatics", *ISRN Biomathematics*, 2013, 615-630, 2013.

[30] Hogeweg, P., Hesper, B. "The alignment of sets of sequences and the construction of phyletic trees: an integrated method", *Journal of Molecular Evolution*, 20, 175–18, 1984.

[31] Karplus, K., Hu, B.R. "Evaluation of protein multiple alignments by SAM-T99 using the BAliBASE multiple alignment test set", *Bioinformatics*, 17, 713–720, 2001.

[32] Lassmann, T., Sonnhammer, E.L.L. "Quality assessment of multiple alignment programs". *FEBS Letters*, 529, 126–130, 2002.

[33] Pais, F.S., et al. "Assessing the efficiency of multiple sequence alignment programs", *Algorithms for Molecular Biology*, 9(1), 4, 2014.

[34] Anderson, C.L., et al. "SuiteMSA: visual tools for multiple sequence alignment comparison and molecular sequence simulation", *BMC Bioinformatics*, 12(1), 184, 2011.

[35] Bahr, A., et al. "BAliBASE (Benchmark Alignment dataBASE): enhancements for repeats, transmembrane sequences and circular permutations", *Nucleic Acids Research*, 29 (1), 323-6, 2001.

[36] Menke, M., et al. "Matt: local flexibility aids protein multiple structure alignment", *PLOS Computational Biology*, 4(1), e10, 2008.

[37] Van Walle, I., et al. "SABmark--a benchmark for sequence alignment that covers the entire known fold space", *Bioinformatics*, 21(7), 1267-1268, 2005.