Cumhuriyet Science Journal

e-ISSN: 2587-246X
ISSN: 2587-2680

Cumhuriyet Sci. J., 42(3) (2021) 526-535
http://dx.doi.org/10.17776/csj.842265

Multiple sequence alignment quality comparison in T-Coffee, MUSCLE and M-Coffee based on different benchmarks

Tuğcan KORAK [1, *], ***Fırat AŞIR*** [2], ***Esin IŞIK*** [3], ***Nur CENGİZ*** [4]

[1]*Kocaeli University, Department of Medical Biology, Faculty of Medicine, Kocaeli/ TURKEY*

[2]*Dicle University, Department of Histology and Embryology, Faculty of Medicine, Diyarbakır/ TURKEY*

[3]*University of Zurich, Institute of Molecular Cancer Research, Zurich/ SWITZERLAND*

[4]*University Hospital Cologne, Institute of Human Genetics, Köln/ GERMANY*

Abstract

Multiple sequence alignment (MSA) is a fundamental process in the studies for determination of evolutionary, structural and functional relationships of biological sequences or organisms. There are various heuristic approaches comparing more than two sequences to generate MSA. However, each tool used for MSA is not suitable for every dataset. Considering the importance of MSA in wide range of relationship studies, we were interested in comparing the performance of different MSA tools for various datasets. In this study, we applied three different MSA tools, T-Coffee, MUSCLE and M-Coffee, on several datasets, BAliBase, SABmark, DIRMBASE, ProteinBali and DNABali. It was aimed to evaluate the differences in the performance of these tools based on the stated benchmarks regarding the % consistency, sum of pairs (SP) and column scores (CS) by using Suite MSA. We also calculated the average values of these scores for each tool to examine the results in comparative perspective. Eventually, we conclude that all three tools performed their best with the datasets from ProteinBali (average % consistency: 29.6, 32.3, 29.7; SP: 0.74, 0.73, 0.74; CS with gaps: 0.27, 0.27, 0.26 for T-Coffee, MUSCLE, M-Coffee, respectively), whereas the lowest performance was obtained in datasets from DIRMBASE (average % consistency: 1.8, 1.1, 4.3; SP: 0.05, 0.04, 0.04 CS with gaps: 0.01, 0, 0.008 for T-Coffee, MUSCLE, M-Coffee, respectively)

Article info

History:
Received: 17.12.2020
Accepted: 29.08.2021

Keywords:
Multiple sequence alignment,
T-Coffee,
MUSCLE,
M-Coffee.

1. Introduction

Multiple sequence alignment (MSA) is a fundamental process in the studies for determination of evolutionary, structural and functional relationships [1-3]. It is generally used to predict the function and structure of proteins from biological sequences [4, 5]. While next generation sequencing methods have been developing, MSA plays a key role in function and structure comparison in this technology [6]. In addition, different MSA strategies can be developed and designed for specific targets. For instance, ODOTool, developed by Ugurel et al. (2020) [7] for bacterial single nucleotide polymorphism determination, is recently used for the analysis of mutations in genomes of SARS-CoV2 that causes COVID-19 pandemic. Also, various MSA algorithms have been developed and served as tools such as T-Coffee [8], MUSCLE [9] , M-Coffee [10], CLUSTALW [11], Clustal Omega [12], Align-M [13], DIALIGN [14], Kalign [15], MAFFT [16, 17], ProbCons [18], PROMALS3D [19], 3DCoffee [20], HAlign [21], Expresso [22], PRANK [23, 24] and MUMMALS [25] etc.

T-Coffee (Tree-based Consistency Objective Function For alignment Evaluation) is one of the MSA methods that benefits from the progressive-alignment strategy [8, 10]. In this strategy, firstly a phylogenetic tree is constructed between sequences and then an alignment is established according to their order in the tree [26]. For the majority of cases, this approach is successful but its weak point is greediness. If errors occur in initial alignments, they cannot be corrected later, while the remaining sequences are added in [8, 10]. This is the first motivation for T-Coffee, which aims to minimize the greedy character of this algorithm and, hence, provides better use of informationin the initial stages. Furthermore, global alignments which align entire sequences with each other do not assure to obtain optimal solutions. Besides, local alignments which

*Corresponding author. *e-mail address: tugcankorak@gmail.com*

align a part of sequence have great performance when net block of ungapped alignments found in each sequence. The combination of best features of these two alignments could form a powerful method to align multiple sequences and this is the second motivation to design a new method, T-Coffee [8].

Additionally, some tools follow iterative approach, in which progressive alignment in a group of sequences is repeated for certain times until reaching the best optimal alignment as seen in MUSCLE (MUltiple Sequence Comparison by Log-Expectation) [26]. MUSCLE is an MSA tool which was developed by Robert C. Edgar in the beginning of 2000s. The algorithm of MUSCLE aims to decrease the time and computational costs with high throughput and accuracy, as most of the previous tools were unable to provide all of them at once. Four different benchmark datasets are used as a reference to test the algorithm by Robert C. Edgar. Those datasets are BALiBASE, SABmark, SMART and PREFAB. It was shown that 500 sequences with an average length of 350 amino acids are aligned only in seven minutes which is a significant improvement compared to the best MSA tools present in those times [9, 10].

The raise in the genomic, structural and functional knowledge and in the computing power resulted in a new approach named as meta-method, in which the requirement to arbitrarily pick a single method to perform MSA is eliminated. This method is called as M-Coffee, standing for Meta-Consistency Objective Function for Alignment Evaluation. It is a consistency-based meta-method, where results from diverse individual MSA tools are combined by T-Coffee to have a final single MSA [1, 3, 27, 28]. This is a major improvement in MSA approach, considering that there are no certain criteria to select the most proper method for every single study among the various approaches with their own advantages and disadvantages [1, 3, 27]. By using M-Coffee, it is possible to include the results from wide range of MSA tools and receive a final alignment incorporating all these tools [1, 3, 10].

The evaluation of a certain MSA tool regarding certain criteria, such as computational cost or accuracy, requires the comparison of the reference datasets obtained from multiple sequence alignment benchmarks. DIRMBASE is one of the various available systems providing benchmark datasets. It is set up by randomly putting highly conserved motives created by random model of sequence evolution (ROSE) into the long DNA sequences that are unlikely to align [29]. Benchmark Alignment dataBASE (BAliBase) was the initially developed large scale benchmark tool applied in the assessment of MSA quality. The reference alignments obtained from the BAliBase are constructed by considering three dimensional superposition of the alignments [30]. The Sequence Alignment Benchmark (SABmark) is another benchmark which supplies the alignments of proteins that are not close to each other regarding their homology [31]. The datasets in SABmark are divided into two sets which are Twilight Zone, with the alignments of low to low similarity, and Superfamilies, with the alignments of low to intermediate similarity [32].

Although MSA is used in wide range of bioinformatics, verification of an MSA reconstruction quality is impeded due to the deficiency of good reference MSAs [33, 34], and also MSA programs do not offer application to compare MSAs. Drawing inspiration from these, Suite MSA, which is a java-based execution, provides verification and rapid comparison of many MSAs by using alignment statistics. This comparison helps researchers to visually localize the regions where inconsistency occurs between an alternative MSA and a reference MSA. Beside these, Suite MSA contains graphical user interface (GUI) and phylogeny editor to make simulation of biological sequence evolution with determination of variable simulation parameters to generate reference MSAs. Also, the reference MSA can be acquired from a benchmark MSA database or can be manually created [33].

In this study, we applied and compared three different MSA tools, namely T-Coffee, MUSCLE and M-Coffee using different datasets. Suite MSA was run to evaluate the performances of the tools based on the reference datasets obtained from BAliBase, SABmark, DIRMBASE and the constructed ProteinBali and DNABali benchmarks.

2. Materials and Methods

The reference datasets were obtained from BAliBase (Version 3.0 R9), SABmark (Version 1.63), DIRMBASE (Version 1.0), and the constructed ProteinBali and DNABali benchmarks, that are all compatible with the three MSA tools (T-Coffee (Version 10.00.r1613), MUSCLE (Version 3.8.31) and M-Coffee (Version 10.00.r1613 mode mcoffee)), were used in this study. DIRMBASE dataset contains locally related DNA sequences including ROSE motifs and motifs of length 60 [29]. SABmark dataset provides MSA of protein sequences that have low similarity [31, 32]. BAlibase dataset designed specifically for MSA and offers high quality manually refined reference alignments by considering three dimensional superpositions. It provides simulation of

real problems that could be encountered during MSA and divided into reference datasets with different characteristics [30]. The reference data used in the current research were randomly selected from these datasets to be used in MSA tools. Box10, 22, 32 were selected from BAlibase; d1a6m__-d1ash, d1ash__-d1dlwa, d1dlwa_-d1ew6a, d1ew6a_-d1gtea1, d1gtea1-d1gvha1 were selected from SABmark; dna-400-30-4-0, r1-dna-400-30-4-1, r1-dna-400-30-4-2 and r1-dna-400-30-4-3 were selected from DIRMBASE. ProteinBali was randomly constructed for protein sequences from a different subset of BAliBase benchmark and includes box001, 022, 034, 036, 046, 050, 054, 076, 0133, 0153. Finally, DNABali (Reference Protein-Coding DNA Alignments Databases: balibase_mdsa_all.tar.gz, http://dna.cs.byu.edu/mdsas/download.shtml) was randomly constructed for DNA sequences from BAliBase benchmark and includes RV61_sushi_ref6, RV64_kringle_1_ref6, RV65_zf_1_ref6, RV66_sushi_2_ref6 and RV70_photo_ref7.

The reference sequence from BAliBase was converted from ".msf" format into ".fasta" format using Jalview (http://www.jalview.org/Download), an MSA editor. Jalview enables researchers to carry out desired editing in MSA, to analyze the MSA and even to construct proper annotations [35]. Subsequently, the dashes in all reference datasets were deleted in order to obtain them in unaligned form that is required to upload them in MSA tools.

2.1. Processing through T-Coffee

Reference data in unaligned form from all datasets were uploaded to T-Coffee Multiple Sequence Alignment Server (http://tcoffee.crg.cat/). The output format was selected as fasta-aln and other parameters remained as default.

2.2. Processing through MUSCLE

The datasets in the unaligned form were uploaded to MUSCLE web server (http://toolkit.tuebingen.mpg.de/musc le). As an option "output sequences in input order" was marked and other parameters remained as default values which are "3" as maximum number of iteration and .fasta format as an output format.

2.3. Processing through M-Coffee

Since M-Coffee which is available from the web server (http://tcoffee.crg.cat/apps/tcoffee/do:mcoffee) was developed based on T-Coffee, their running procedure resembles to each other. In addition to all parameters selected for T-Coffee, as described above, a set of MSA tools was constructed by selecting Mpcma_msa, Mmafft_msa, Mclustalw_msa, Mdialigntx_msa, Mpoa_msa, Mmuscle_msa, Mprobcons_msa and Mt_coffee_msa. Because M-Coffee is a consistency-based meta-method in which results from diverse individual MSA tools are combined by T-Coffee to have a final single MSA.

At the end, results of each tool were obtained from result folders by choosing fasta-aln. Next, this text file was converted to .fasta format by Jalview to become compatible with Suite MSA program. Then, Suite MSA was run to compare alignments with reference alignment. All these steps were repeated for each dataset.

2.4. Running suite MSA

Suite MSA has a wide range of applications. In this study, MSA Comparator was used for the comparison of the obtained alignment with the reference alignment. However, there was need to check whether the names, order and content of the aligned sequences except for the placement of the gaps in both reference and result files were exactly same. If not, they should be adjusted to have the identical names and sequences.

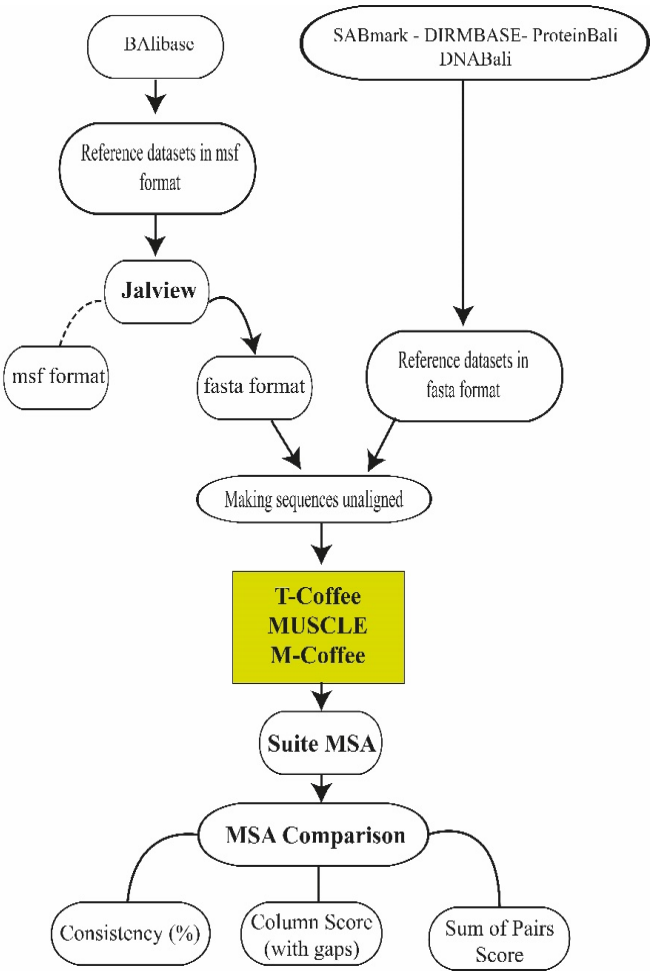

Figure 1. The workflow to evaluate the performance of MSA tools based on different benchmarks.

After ensuring these aspects, initially reference file and then result file were uploaded into the program. In the open window, there were several buttons for functions which would provide information about the compared files when selected. One of them was "Show sum of pairs" which informs about the % consistency, sum of pairs score (SP), column score (CS) without gaps and with gaps. These values were used to plot the graphs and bar charts providing the statistical analysis of the tools. All the steps above are summarized in Figure 1.

3. Results and Discussion

In our study, we used % consistency, SP score and CS score to evaluate the quality and reliability of three MSA tools. % consistency represents percentage of columns in the obtained MSA which are 100% identical to the columns of reference MSA. SP score is calculated as a whole score of the alignment and a determinant parameter to understand how successful the tool in aligning. The SP score receives score 1 when the identical alignment is obtained from the comparison and score 0 refers to incorrect alignment. The greater SP score shows the greater number of correctly aligned sequences. However, in CS calculation for similarity, each column is scored independently from each other and then the total score is divided into number of columns analyzed. Therefore, SP score is used as the major indicator of quality, while other values are also analysed to support the result [30, 36].

Table 1. The acquired values for % consistency, sum of pairs score and column score

		T-Coffee			M-Coffee			MUSCLE		
		Cons.	SOP	CS	Cons.	SOP	CS	Cons.	SOP	CS
BAliBase	box10	8.581	0.53	0.069	9.008	0.546	0.072	8.031	0.496	0.064
	box22	10015	0.486	0.051	12957	0.503	0.086	11623	0.464	0.044
	box32	19225	0.472	0.221	20675	0.473	0.217	20966	0.468	0.216
DIRMBASE	4-0	1603	0.032	0.009	2306	0.049	0.023	0.183	0.036	0
	44200	2403	0.04	0.009	7143	0.064	0.011	30317	0.069	0
	44231	1426	0.096	0.011	2936	0.024	0	2308	0.024	0
	44259	1908	0.037	0.012	5137	0.032	0	0.182	0.035	0
SABmark	1_2	31481	0.371	0.338	35669	0.419	0.369	30189	0.355	0.317
	2_3	16456	0.075	0.067	33113	0.43	0.357	31013	0.28	0.248
	3_4	8125	0.034	0.032	45261	0.046	0.035	20863	0.19	0.114
	4_5	18132	0	0	17582	0	0	20968	0	0
	5_6	18947	0	0	18717	0	0	25907	0	0
DNABali	RV61	42805	0.031	0	37821	0.034	0	43492	0.038	0
	RV64	0.566	0.112	0	2718	0.242	0	2773	0.2	0
	RV65	9681	0.183	0	9538	0.356	0	41456	0.126	0
	RV66	2463	0.134	0	6971	0.423	0	20852	0.187	0
	RV70	5981	0.568	0.072	44420	0.647	0.133	9776	0.641	0.107
ProteinBali	Box001	59366	0.837	0.635	59.24	0.835	0.622	59.97	0.822	0.623
	Box022	16834	0.556	0.015	3887	0.567	0.03	5321	0.546	0.023
	Box034	14575	0.798	0.145	16383	0.789	0.159	17863	0.779	0.145
	Box036	62.08	0.925	0.679	62261	0.926	0.667	60938	0.929	0.628
	Box046	2713	0.523	0	3948	0.534	0	4628	0.503	0
	Box050	44348	0.815	0.428	32378	0.783	0.308	47.72	0.801	0.472
	Box054	17477	0.606	0.209	20613	0.614	0.22	23922	0.594	0.244
	Box076	1964	0.743	0	0.508	0.736	0	0.116	0.745	0
	Box0133	69459	0.866	0.308	71662	0.876	0.368	70248	0.864	0.32
	Box0153	44369	0.75	0.26	26039	0.753	0.278	32796	0.749	0.297

Cons.: consistency (%), SOP: sum of pairs score, CS: column score.

As a consequence of performing Suite MSA on the alignments obtained from each tool for each datasets, a large quantity of data was generated. This data was organized in the Table 1 showing the values of % consistency, SP score and CS for each individual dataset. Then, these values were plotted as two different sets of graphs. The first graph set in Figure 2 and Figure 3 enables to compare the performance of each tool on individual datasets in terms of % consistency, SP and CS values. The second graph set shown in Figure 4 shows general comparison of these three MSA tools by taking average % consistency, CS and SP scores into consideration. These average values obtained for each tool on different datasets from the given benchmark were plotted as bar charts.

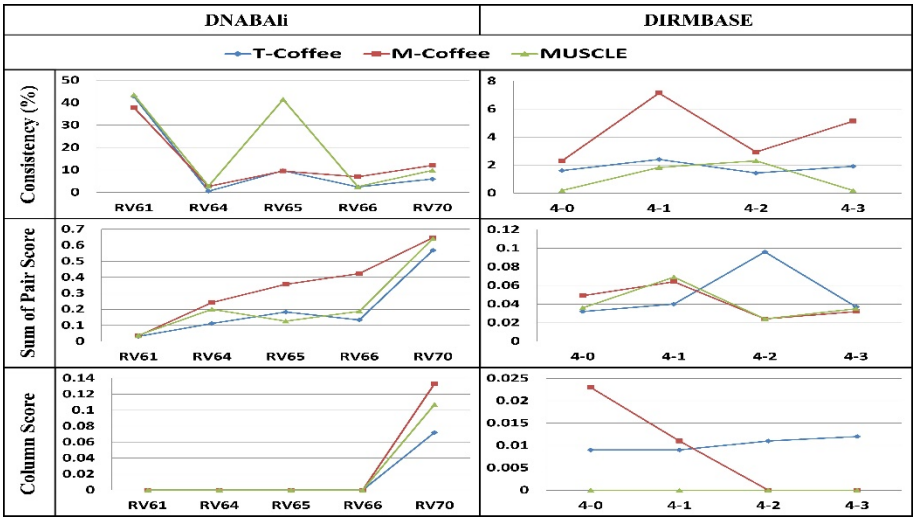

Figure 2. This figure indicates the results for the evaluation of the given tools based on the benchmark systems that provide nucleotide datasets. It consists of the graphs representing the results of the tools indicated by each column. The rows of the figure illustrates the acquired values for % consistency, sum of pairs score and column score. Each graph located in the cells of the figure shows the trend of the scores and accuracy over datasets. (For DNABali, RV61, RV64, RV65, RV66 and RV70 refer to RV61_sushi_ref6, RV64_kringle_1_ref6, RV65_zf_1_ref6, RV66_sushi_2_ref6 and RV70_photo_ref7, respectively. For DIRMBASE, 4-0, 4-1, 4-2, 4-3 refer to r1-dna-400-30-4-0, r1-dna-400-30-4-1, r1-dna-400-30-4-2 and r1-dna-400-30-4-3, respectively)

The scores acquired from DNABali (Figure 2) revealed that all of three tools perform similarly regarding their % consistency in each dataset, except for RV65 in which MUSCLE gave the highest percentage value. In contrast, RV65 had the lowest score for the SP with MUSCLE. M-Coffee gave the highest scores for SP in almost each dataset. Eventually, CS was obtained as 0 for datasets out of the RV70 dataset at which M-Coffee had the highest value.

Performance scores of the MSA tools compared in this study based on DIRMBASE datasets revealed the highest % consistency for M-Coffee for the whole dataset; while T-Coffee was more consistent than MUSCLE for datasets 4-0, 4-1 and 4-3. On the other hand, M-Coffee has a CS score equal to 0 in datasets of 4-2 and 4-3 while T-Coffee has relatively greater values. However, MUSCLE gave 0 for column scores in all datasets. Also, except for 4-2 and with superior value for T-Coffee, there is no distinctive difference in the SP scores of those three tools in all datasets of this benchmark. Based on these, if higher consistency is desired, M-Coffee seems to be more convenient tool for the given datasets of DIRMBASE. However, CS and SP scores are either variable in each dataset for the tools or very close to each other's.

For ProteinBali represented in Figure 3.panel, overall % consistency, CS and SP scores are very similar for each datasets for each tool. However, MUSCLE has slightly higher consistency and CS for box050 and box054 datasets. SP scores do not show significant differences among the tools based on the tested dataset.

In Figure 3. panel representing graphs of BAliBase dataset, M-Coffee has given the overall highest scores for all statistical tests. % consistency and SP scores are very close to each other for all tools although M-Coffee is slightly higher than others. Similarly, column scores are also very close to each other but M-Coffee is significantly greater than others only for box22 in the given dataset.

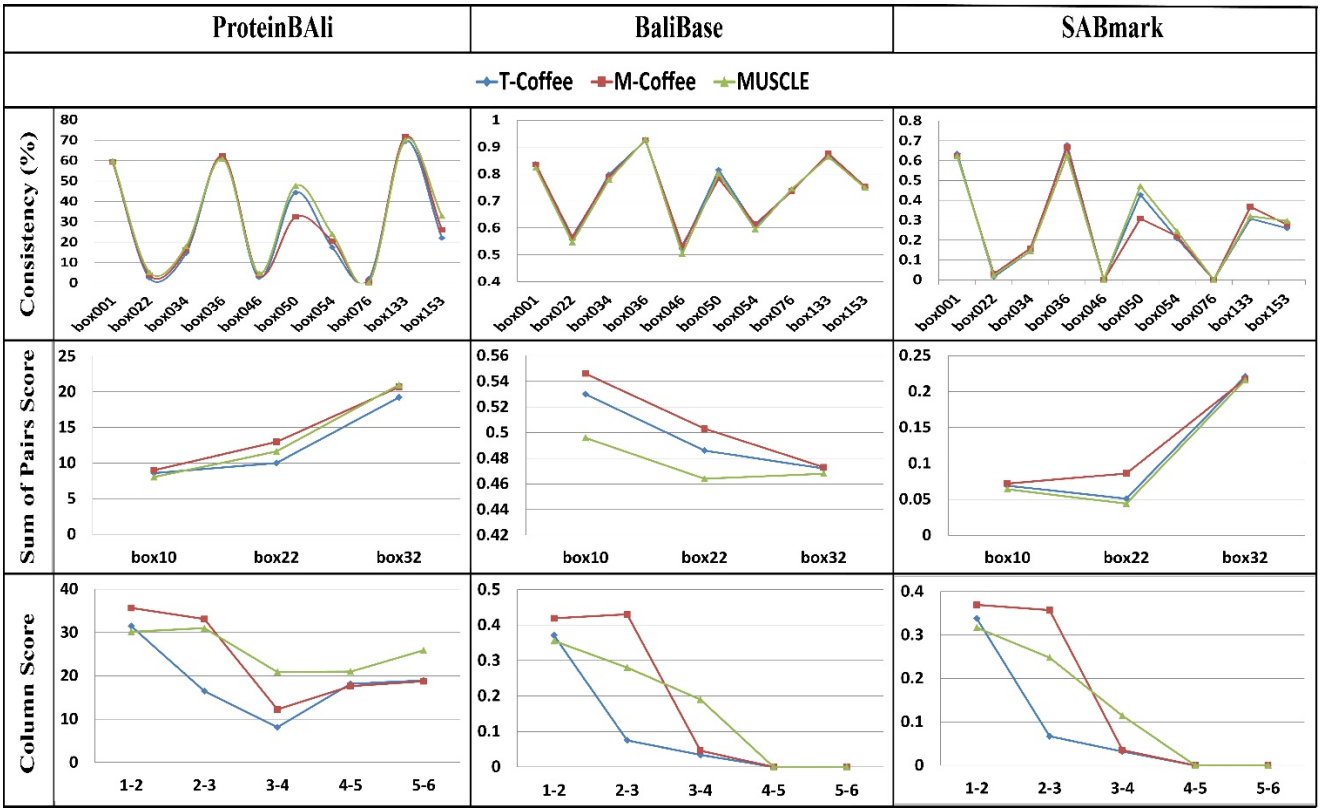

Figure 3. This figure indicates the results for the evaluation of the given tools based on the benchmark systems that provide amino acid datasets. It consists of the graphs representing the results of the tools indicated by each column. The rows of the figüre illustrates the acquired values for % consistency, sum of pairs score and column score. Each graph located in the cells of the figure shows the trend of the scores and accuracy over datasets. (For SABmark, 1, 2, 3, 4, 5, 6 refer to d1a6m__, d1ash__, d1dlwa_, d1ew6a_, d1gtea1 and d1gvha1, respectively)

For 1-2 dataset of SABmark in Figure 3.panel, M-Coffee has the highest % consistency and others resulted in relatively similar values to each other. For 2-3 dataset, M-Coffee and MUSCLE had similar percentages and significantly higher than T-Coffee. For the remaining datasets, MUSCLE gave greater consistency than other tools having similar consistencies. When the CS is compared, all tools possess approximately same values for 1-2 datasets but T-Coffee has lower CS for 2-3 datasets than others. Next, 3-4 dataset shows very close CS in T-Coffee and M-Coffee but CS is slightly higher in MUSCLE. Additionally, CS is 0 for the other datasets in each tool. SP score is very close in T-Coffee and MUSCLE for 1-2 datasets but it is slightly higher in M-Coffee. For 2-3 dataset, there are noticeable differences among scores; M-Coffee has the greatest and T-Coffee has the lowest sum of pairs score. Next, for 3-4 datasets the highest SP score belongs to MUSCLE than other tools in which scores are close to each other. The remaining datasets give 0 for SP score in each tool.

The information obtained from the bar charts illustrate that all tools have given the highest scores and % consistency with ProteinBali (Figure 4). The second highest scores are generated in SABmark and BAliBase datasets and their values are similar for % consistency and CS but BAliBase has greater results for the SP scores. Eventually, the lowest scores belong to DIRMBASE datasets. The % consistency values are in between 1% and 35% which is not high enough to mention about an acceptable consistency. SP scores varied between 0.04 and 0.75 and CS was between 0 and 0.27. SP scores (Figure 4) and CS indicate higher quality of tools for the given alignments if they are close to 1 and lower quality if they are close to 0. Our results clearly show that both SP and CS are very close to 0 for all tools and datasets except for scores for ProteinBali (Figure 4).

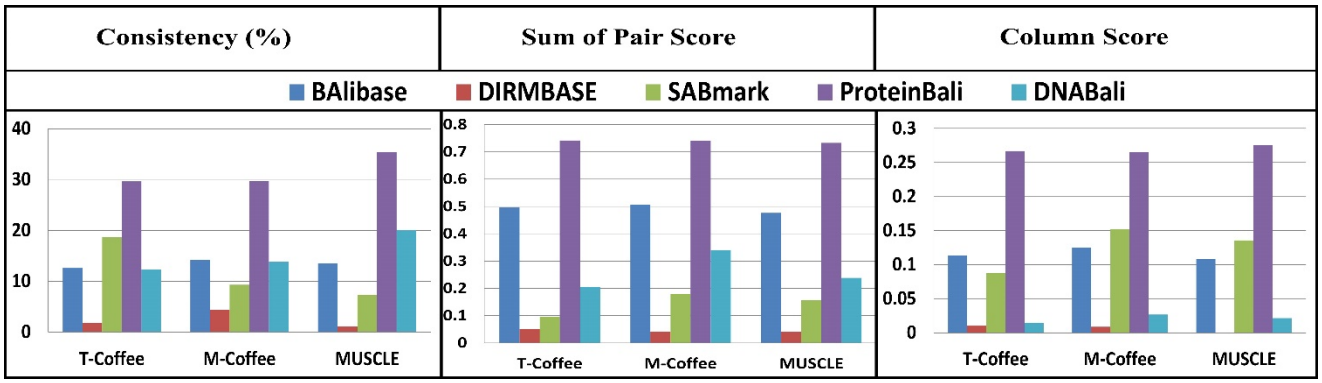

Figure 4. Bar chart graphs belonging to average % consistency, sum of pairs and column score values of the given tool for the given benchmark.

MSA is a process required in a wide range of bioinformatic studies (Bawono et al., 2017). There are numerous distinctive approaches for MSA, each of which possesses their own limitations and advantages. Despite the lack of certain criteria, aligners attempt to use the ones which provide high accuracy, take shorter computational time and do not cause problems with the computational memory. Therefore, the selection of the proper tool is not straightforward and requires consideration of several aspects based on the study and circumstance [10].

As expected, some of the tools perform better than the others on some datasets. Although there were various sequences in DNABali, some of them could be processed by all three tools due to some limitations. In addition to number of sequence limitation, which is at most 150 sequences for T-Coffee and M-Coffee, there is also number of character limitation in M-Coffee which cannot accept more than 2500 characters. Although MUSCLE was predicted to be more advantegous in terms of its ability to accept input with any length, it resulted in some outputs which were not applicable to Suite MSA. The reason of the error is because the residues within the sequence of RV62, RV63 and RV67 obtained from MUSCLE are not identical to their reference alignments. This seems to be caused by that several of the residues in the DNA sequences found in these datasets have been changed by MUSCLE into ambiguity codes. To use MSAcomparator, residues in each sequence have to be the same so the comparison scores cannot be obtained for these datasets of DNABali. Consequently, only 5 of the given datasets were processed by all three tools. The acquired scores from DNABali (Figure 2) revealed that MUSCLE seems to be advantageous for only one of the datasets to obtain higher % consistency and in the other datasets, the percentages are similar in three tools. However, M-Coffee seems to be preferable regarding with the other scores. For the scores of DIRMBASE datasets (Figure 2), if a higher consistency is desired, M-Coffee seems to be more convenient tool for given datasets of DIRMBASE. However, CS and SP scores vary greatly among the tested MSA tools for some sub-datasets and similar for the others Therefore, a general conclusion in terms of more preferable tool cannot be done based on these parameters.

From the ProteinBali scores (Figure 3), it can be deduced that since none of the tools has a score dominancy to others, they are almost equally convenient for this datasets. For the BAliBase dataset (Figure 3), even though M-Coffee had higher scores for the tested datasets, the values do not significantly outperform MUSCLE and T-Coffee. Scores of SABmark reveals that for the greater consistency, MUSCLE can be suitable for each given datasets of this benchmark. However, the general convenient tool cannot be determined for the other scores because obtained scores are variable in each given dataset. Additionally, CS and SP score is 0 for the all tools in 4-5 and 5-6 datasets even though % consistency different from 0 is obtained. These might be improved by alterations in the settings of the tools. Otherwise, these tools may not be preferred for the analysis of these datasets.

As seen from the above, the given datasets of benchmarks show differences in terms of % consistency, CS and SP score. Thus, it is difficult to deduce the proper tool for benchmarks. To approach more general and to distinct tool performances clearly, bar charts that represent average scores of each benchmark were plotted (Figure 4). It provides less detail about which tool performs higher scores on which benchmark. At a first glance, it is possible to see the tool and corresponding benchmark with greater scores. The information obtained from these bar charts illustrate that all tools have given the highest scores and % consistency with ProteinBali. Also, both SP and CS are very close to 0 for all tools and datasets except for scores for ProteinBali (Figure 4). Additionally, BAliBase resulted in greater scores compared to

remaining tools. Relatively higher scores for BAliBase dataset can be related to relatively higher % consistency values and high % consistency values might be resulted from higher sequence similarity levels. Low scores for others mean that MUSCLE, M-Coffee and T-Coffee tools are not very reliable for the given datasets. This was unexpected; however, it can be due to small size of our datasets. Because, the evaluation of such tools are carried out with much larger datasets containing hundreds or thousands of sequences. It is also difficult to associate the results to the average length of the sequences as the measurements of the sequence similarity are not available. Additionally, more information should be taken into consideration for comprehensive analysis such as scoring matrices or gap penalties, minimum level of sequence similarities. Number of iterations, for MUSCLE, could be another factor that may be affecting the quality of the alignment as it can be manually controlled, also can be determined automatically.

Finally, % consistency, CS and SP are low for each tool in different benchmarks. The underlying reasons could be related to quality of the given data. Next, the aligned sequence lengths can affect the obtained scores because tools work better in proper length intervals and the given sequence lengths may be out of this interval. Furthermore, level of homology with the reference sequence is important in the performance of the tools. For example, some tools give greater scores as the percentage of homolog sequences increase. Beside these, the suitability of given dataset to the tools should be considered when the low scores are encountered. However, this is not the case for T-Coffee, M-Coffee and MUSCLE which can be used for protein, DNA and RNA datasets. Nevertheless, when the datasets used in this study are considered, it can be obviously seen that all three tools are much more appropriate for the amino acid datasets (Figure 4). This situation may change with different benchmarks or datasets. Additionally, the differences in the algorithms may also affect the scores. However, this is less probable according to our results (Figure 2 and 3), since there is no clear difference when T-Coffee and M-Coffee that has progressive algorithm is compared with MUSCLE having an iterative strategy in its algorithm.

M-Coffee is a meta-method combining results from various tools and then apply T-Coffee on the results for these tools to reach the ultimate MSA. As expected, M-Coffee took longer time compared to MUSCLE and T-Coffee because of this distinct processing scheme. M-Coffee was expected to be more reliable and to give better results than particularly T-Coffee due to its strategy. However, there are some datasets of which values are greater with T-Coffee or MUSCLE than with M-Coffee. This can be resulted from again the similarity level of the sequences, as one of the limitations of this computationally intensive method is that it is not preferable when the distant sequences are attempted to align [1]. Consequently, these unexpected lower values for M-Coffee may be related to the percentages of sequence similarity.

This study has shown that the choice of the proper tool for MSA is not straightforward and several aspects such as the homology level or length of sequences should be taken into consideration. Although we carried out our study for only small size of data, our results are sufficient to support the idea that while the MSA tool is suitable for a dataset from a certain benchmark, it may be inappropriate for another dataset from the same benchmark system. This illustrates that there are still plenty of limitations to be eliminated in MSA tools. With the increasing information about the structure and function and also improvements in the computing power, it may lead to the development of the new strategies not only revealing more accurate alignments but also confirming and fixing the results acquired from previous studies.

All in all, available powerful MSA methods with the distinctive strategies like M-Coffee, T-Coffee and MUSCLE are fundamental steps through the further studies like three-dimensional structure determination. However, as seen in our results, the evaluation of them is dependent on the references from benchmark systems and is complicated process, since the tools are not suitable and reliable for all the datasets.

Acknowledgment

The authors acknowledge Prof.Dr. Jens Allmer for his guidance in the conceptualization of this study.

Conflicts of interest

All authors declare that they have no conflict of interest.

References

[1] Notredame C., Recent evolutions of multiple sequence alignment algorithms, *PLoS Comput. Biol.*, 3(8) (2007) e123.

[2] Edgar R.C., MUSCLE: multiple sequence alignment with high accuracy and high throughput, *Nucleic Acids Res*., 32(5) (2004) 1792-1797.

[3] Moretti S., Armougom F., Wallace I.M., Higgins D.G., Jongeneel C.V., Notredame C., The M-Coffee web server: a meta-method for computing multiple sequence alignments by combining alternative alignment methods, *Nucleic Acids Res.*, 35(Web Server issue) (2007) W645-648.

[4] Wang Y., Wu H., Cai Y., A benchmark study of sequence alignment methods for protein clustering, *BMC Bioinformatics*, 19(Suppl 19) (2018) 529.

[5] Maiolo M., Zhang X., Gil M., Anisimova M., Progressive multiple sequence alignment with indel evolution, *BMC Bioinformatics*, 19(1) (2018) 331.

[6] Bawono P., Dijkstra M., Pirovano W., Feenstra A., Abeln S., Heringa J., Multiple Sequence Alignment, *Methods Mol. Biol*., 1525 (2017) 167-189.

[7] Ugurel O.M., Ata O., Turgut-Balik D., An updated analysis of variations in SARS-CoV-2 genome, *Turk. J. Biol.*, 44(3) (2020) 157-167.

[8] Notredame C., Higgins D.G., Heringa J., T-Coffee: A novel method for fast and accurate multiple sequence alignment, *J. Mol. Biol*., 302(1) (2000) 205-217.

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

[10] Edgar R.C., Batzoglou S., Multiple sequence alignment, *Current Opinion in Structural Biology*, 16(3) (2006) 368-373.

[11] Thompson J.D., Higgins D.G., Gibson T.J., CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic Acids Res.*, 22(22) (1994) 4673-4680.

[12] Sievers F., Wilm A., Dineen D., Gibson T.J., Karplus K., Li W., Lopez R., McWilliam H., Remmert M., Söding J., Thompson J.D., Higgins D.G., Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega, *Mol. Syst. Biol.*, 7 (2011) 539.

[13] Van Walle I., Lasters I., Wyns L., Align-m--a new algorithm for multiple alignment of highly divergent sequences, *Bioinformatics*, 20(9) (2004) 1428-1435.

[14] Morgenstern B., Frech K., Dress A., Werner T., DIALIGN: finding local similarities by multiple sequence alignment, *Bioinformatics*, 14(3) (1998) 290-294.

[15] Lassmann T., Sonnhammer E.L.L., Kalign – an accurate and fast multiple sequence alignment algorithm, *BMC Bioinformatic*s, 6(1) (2005) 298.

[16] Katoh K., Misawa K., Kuma K., Miyata T., MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform, *Nucleic Acids Res.*, 30(14) (2002) 3059-3066.

[17] Katoh K., Standley D.M., MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability, *Molecular Biology and Evolution*, 30(4) (2013) 772-780.

[18] Do C.B., Mahabhashyam M.S., Brudno M., Batzoglou S., ProbCons: Probabilistic consistency-based multiple sequence alignment, *Genome Res.*, 15(2) (2005) 330-40.

[19] Pei J., Kim B.H., Grishin N.V., PROMALS3D: a tool for multiple protein sequence and structure alignments, *Nucleic Acids Res*., 36(7) (2008) 2295-2300.

[20] O'Sullivan O., Suhre K., Abergel C., Higgins D.G., Notredame C., 3DCoffee: Combining Protein Sequences and Structures within Multiple Sequence Alignments, *Journal of Molecular Biology*, 340(2) (2004) 385-395.

[21] Zou Q., Hu Q., Guo M., Wang G., HAlign: Fast multiple similar DNA/RNA sequence alignment based on the centre star strategy, *Bioinformatics*, 31(15) (2015) 2475-2481.

[22] Armougom F., Moretti S., Poirot O., Audic S., Dumas P., Schaeli B., Keduas V., Notredame C., Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee, *Nucleic acids research*, 34(Web Server issue) (2006) W604-W608.

[23] Löytynoja A., Goldman N., An algorithm for progressive multiple alignment of sequences with insertions, *Proc. Natl. Acad. Sci. U S A*, 102(30) (2005) 10557-10562.

[24] Löytynoja A., Goldman N., Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis, *Science*, 320(5883) (2008) 1632-1635.

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

[26] Kemena C., Notredame C., Upcoming challenges for multiple sequence alignment methods in the high-throughput era, *Bioinformatics*, 25(19) (2009) 2455-2465.

[27] Wallace I.,M., O'Sullivan O., Higgins D.G., Notredame C., M-Coffee: combining multiple sequence alignment methods with T-Coffee, *Nucleic Acids Research*, 34(6) (2006) 1692-1699.

[28] Rosenberg M.S., Sequence alignment: Methods, models, concepts, and strategies, In: Rosenberg M.S., (Ed) California: University of California Press, (2009).

[29] Subramanian A.R., Kaufmann M., Morgenstern B., DIALIGN-TX: greedy and progressive approaches for segment-based multiple sequence alignment, *Algorithms for molecular biology:AMB*, 3 (2008) 6.

[30] Pais F.S., Ruy P.C.,Oliveira G., Coimbra R.S., Assessing the efficiency of multiple sequence alignment programs, *Algorithms Mol. Biol.*, 9(1) (2014) 4.

[31] Menke M., Berger B., Cowen L., Matt: local flexibility aids protein multiple structure alignment, *PLoS Comput. Biol.*, 4(1) (2008) e10.

[32] Van Walle I., Lasters I., Wyns L., SABmark--a benchmark for sequence alignment that covers the entire known fold space, *Bioinformatics*, 21(7) (2005) 1267-1268.

[33] Anderson C.L., SuiteMSA User's Manual. Nebreska, USA: University of Nebraska–Lincoln, (2011).

[34] Saeed F., Perez-Rathke A., Gwarnicki J., Berger-Wolf T., Khokhar A., A high performance multiple sequence alignment system for pyrosequencing reads from multiple reference genomes, *Journal of Parallel and Distributed Computing*, 72(1) (2012) 83-93.

[35] Waterhouse A.M., Procter J.B., Martin D.M., Clamp M., Barton G.J., Jalview Version 2--a multiple sequence alignment editor and analysis workbench, *Bioinformatics*, 25(9) (2009) 1189-1191.

[36] Anderson C.L., Strope C.L., Moriyama E.N., SuiteMSA: visual tools for multiple sequence alignment comparison and molecular sequence simulation, *BMC Bioinformatics*, 12(1) (2011) 184.