I.J. Information Technology and Computer Science, **2018, 8, 24-30**
Published Online August 2018 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2018.08.04

Comparative Analysis of Multiple Sequence Alignment Tools

Eman M. Mohamed
Faculty of Computers and Information, Menoufia University, Egypt
E-mail: eman.mohamed@ci.menofia.edu.eg.

Hamdy M. Mousa, Arabi E. keshk
Faculty of Computers and Information, Menoufia University, Egypt
E-mail: hamdimmm@hotmail.com, arabikesk@yahoo.com.

Received: 24 April 2018; Accepted: 07 July 2018; Published: 08 August 2018

***Abstract*—**The perfect alignment between three or more sequences of Protein, RNA or DNA is a very difficult task in bioinformatics. There are many techniques for alignment multiple sequences. Many techniques maximize speed and do not concern with the accuracy of the resulting alignment. Likewise, many techniques maximize accuracy and do not concern with the speed. Reducing memory and execution time requirements and increasing the accuracy of multiple sequence alignment on large-scale datasets are the vital goal of any technique. The paper introduces the comparative analysis of the most well-known programs (CLUSTAL-OMEGA, MAFFT, BROBCONS, KALIGN, RETALIGN, and MUSCLE). For programs' testing and evaluating, benchmark protein datasets are used. Both the execution time and alignment quality are two important metrics. The obtained results show that no single MSA tool can always achieve the best alignment for all datasets.

***Index Terms*—**Multiple Sequence Alignment, Accuracy, Progressive Alignment, Iterative alignment, and Bioinformatics.

I. Introduction

In bioinformatics, the process of sequence alignment is to put amino acids or nucleotides of RNA, DNA, and protein in the same column because of similarity using gaps in which alignment scores increased. MSA is used to predict the similarity between three or more biology sequence, which it is a generalization to pairwise sequence alignment (PSA). Table 1 describes the main differences between PSA and MSA. MSA developed to predict the functional or structural similarity of more than two sequences, predicted the structure of new sequences, grouping protein into families, and indict the relationship between different sequences.

Alignments can be classified into two types global or local. In global alignment, the sequences are completely compared for increasing the alignment score globally and taking full advantage of the number of matched up residues. The Needleman-Wunsch algorithm is a popular global alignment algorithm built-in dynamic programming technique [1]. This algorithm maximizes the number of amino acid matches and minimizes the number of required gaps to finds globally optimal alignment. Local alignments are more useful for aligning sub-regions of the sequences, whereas local alignment maximizes sub-regions similarity alignment. One of the most known of Local alignment is Smith-Waterman algorithm [2].

Table 1. Pairwise vs. multiple sequence alignment

PSA	MSA
Compare two biological sequences.	Compare more than two biological sequences.
Generally categorized as local or global alignment.	
Simple algorithms used. Global → Needleman-Wunsch Local → Smith-Waterman algorithm	Techniques: Dynamic alignment Progressive alignment, Iterative Alignment
Alignment Tools : Blast- EMBOSS Needle- EMBOSS Water, k-tuple, k-mer algorithms	Alignment Tools : MUSCLE, MAFFT, CLUSTAL family, T-coffee, KALIGN, RETALIGN, FSA

Dynamic Programming, Progressive Alignment, and Iterative Alignment are the main techniques for solving MSA. These techniques have different attributes. The main objectives of MSA techniques are to increase the alignment score and reduce execution time for all categories of biological sequences [3]. The author tries to improve the efficiency of the dynamic algorithm using only three main diagonals by ignoring useless data [4]. The paper enhances the performance of the Needlelman-Wunsch algorithm by using software pipelining technique and OpenMP programming [5]. The authors propose the parallel form for edit distance algorithm for PSA to reduce runtime and improve the accuracy of alignment [6].

This paper presents a Comparative study of the most well-known programs for multiple sequence alignment. The MSA programs comparison is necessary for biologist users to select the best MSA software corresponding to their needs. Whereas, there are many MSA programs tries

to improve alignment score. However, there is no single program generate optimal alignment for any biology case study.

This study compares and evaluates six well-known MSA software namely, CLUSTAL-OMEGA, MAFFT, MUSCLE, KALIGN, BROBCONS, and RETALIGN. MSA programs are available as web interfaces. In this study, the sum of pairs score (SP score) and Column Score (CS) are used for measuring the quality of the alignment. This comparison examines on BALIBASE 3.0 references.

The remainder of this paper is organized as follows; section II explains the three standard MSA methods such as dynamic, progressive and iterative alignment. In section III, the most well-known tools will be described. This tools namely: CLUSTAL-OMEGA, MAFFT, BROBCONS, KALIGN, RETALIGN, and MUSCLE. Section VI reviews the description and characteristics of BALIBASE v3 datasets. The practical results are shown in section V. The overall performance of the alignment obtained is analyzed based on the SPscore and TCscore (Total column score).

II. MSA Methods

There are different methods of MSA with different attributes and drawbacks. Some of these MSA methods are useful based on speed and accuracy. This section focuses on standard MSA methods

A. Dynamic programming alignment

Dynamic programming (DP) is used for finding optimal alignment of every sub-problem instead of re-computing them. DP searches for the alignment by giving some scores of matches and mismatches. DP obtains an accurate alignment and maximizes score function. To find similarity, it is essential to create the pairwise alignment of the two sequences by calculating a similarity score. The similarity score is attained by using the scoring system or substitution matrix [7]. The scoring system firstly gives a score values for a match, a mismatch, and a gap [8]; as in this example assign +2 for the match, -1 for mismatch and -2 for gap penalty.

Sequence 1: A T C G A G T A

Sequence 2: A - C G T - T A

Thus, for the alignment the similarity score is 5*2+1*-1+2*-2=+5. A substitution matrix is a grid that represents the collection of scores for the substitution of every nucleotide or amino acids with one another. The substitution matrix has the one row and one column for each possible letter in alphabet letters (ex. four rows and four columns for DNA, RNA) [7]. For example, the *i, j* element of the matrix has a value of +2 if match and -1 if a mismatch, The BLOcks SUbstitution Matrix (BLOSUM) is another amino acid substitution matrix. The matrix that constructed with no more than x% of sequences similarity is called BLOSUM-x. For example, using BLOSUM62 for alignments of sequences that have less divergent alignments and BLOSUM50 for alignments of sequences that have more divergent alignments [7].

However, DP methods are needed high computational power for large-scale datasets; the dynamic programming method gives the best possible alignment that maximizes the similarity score [9].

B. Progressive alignment

Progressive is a heuristic approach, which builds alignment progressively [10]. Progressive MSA performing alignment based on separating MSA into subsequences. In the first step, subsequence aligns in a pairwise manner using methods such as the Needleman-Wunsch, Smith-Waterman, k-tuple, or k-mer algorithm. The second step shows the relationship between the subsequences using clustering methods such as k-means. Next, a guide tree is constructed based on the similarity score. Finally, all subsequences alignment assembles one by one according to the guide tree. However, progressive MSA is very fast, it is not an optimal alignment technique. Progressive MSA provides near optimal alignment depended on the initial pairwise sequence alignment [10]. CLUSTALW [11], CLUSTAL-OMEGA [12], MAFFT [13], KALIGN [14], MUSCLE [15], BROBCONS [16] and RETALIGN [17] are popular progressive MSA programs.

C. Iterative Alignment

Iterative MSA is an extension method of progressive MSA, which modifies the construction of guide tree [18]. In iterative MSA, the dynamic programming applies to improve the alignment accuracy. In the first step, construct an initial MSA then, divide the initial MSA into subgroups. The second step realigns the subgroups using dynamic programming. Finally, rebuilding MSA until finding the best alignment score or for predefined iterative times [18]. MUSCLE [15], DIALIGN and T-Coffee [19] are popular iterative MSA programs.

III. MSA Programs

In this paper, the most well-known tools will be described. This tools namely: CLUSTAL-OMEGA, MAFFT, BROBCONS, KALIGN, RETALIGN, and MUSCLE. Table 2 describes the some of MSA tools for their method, type of sequences, and download server. These tools are publicly available on web servers, so users need not install some of MSA tools.

A. CLUSTAL-OMEGA

CLUSTAL family is very popular progressive alignment methods, especially the weighted variant CLUSTALW [11] and CLUSTAL-OMEGA [12]. Many web servers could access CLUSTAL-OMEGA and it is a current standard version. The next step, using the UPGMA method to construct a guide tree. The final step outputs multiple sequence alignment by a progressive alignment using the HHalign package [10]. The following steps illustrate the CLUSTAL-OMEGA algorithm.

Input: n DNA or n RNA or n Protein Sequences, S1, S2, ...Sn
Stage 1: Apply Pairwise alignments by the k-tuple method to generate

distance score matrix.
Stage 2: Sequence clustering by a mBed method from the distance matrix generated in Stage 1.
Stage 3: Sequence clustering by a k-means method from the distance matrix generated in Stage 2.
Stage 4: Guide tree construction by UPGMA method from the distance matrix generated in Stage 3.
Stage 5: Progressive alignment by HHalign package.
Output: naligned DNA or naligned RNA or naligned Protein Sequences S'1, S'2,..., S'n

Table 2. MSA techniques comparison

Technique	input format	output format	Web	Max # seq	File Size	seq type	Method	Server
CLUSTAL-OMEGA	FASTA, EMB, GenBank	ClustalW/ Pearson/FASTA/ MSF/	yes	max 4000 sequences	max file size of 4 MB.	Protein, DNA, RNA	global/ Progressive	http://www.clustal.org/omega/ https://www.ebi.ac.uk/Tools/msa/clustalo/
MUSCL	FASTA, EMB, GenBank	Fasta , Clustalw, MSF/html	yes	up to 500 sequences	max file size of 1 MB.	Protein	Progressive Step1 and Step2 iterative Step 3	http://www.drive5.com/muscle/ https://www.ebi.ac.uk/Tools/msa/muscle/
MAFFT	FASTA, EMB, GenBank	ClustalW/ Pearson/FASTA	yes	up to 500 sequences	max file size of 1 MB.	Protein, DNA, RNA	global/ Iterative	http://mafft.cbrc.jp/alignment/server/ https://www.ebi.ac.uk/Tools/msa/mafft/
KALIGN	FASTA, EMB, GenBank	MACSIM/ ClustalW/ Pearson/FASTA	yes	up to 2000 sequences	max file size of 2 MB.	Protein, DNA, RNA	Progressive	http://msa.sbc.su.se/cgi-bin/msa.cgi https://www.ebi.ac.uk/Tools/msa/kalign/
RETALIGN	FASTA	ClustalW	no version 0.22	max 1000 sequences	not limited	Protein	Progressive Corner-cutting Multiple Sequence Alignment	http://phylogenycafe.elte.hu/RetAlign/
PROBCONS	MFA	MFA/ ClustalW	yes version 1.12	max 1000 sequences	not limited	Protein	Probabilistic Consistency-based Multiple Alignment of Amino Acid Sequences	http://probcons.stanford.edu

B. MUSCLE

MUSCLE (MUltiple Sequence Comparison by Log-Expectation) is used for multiple protein sequences [15]. This algorithm builds initial alignment based on similarities of paired alignments then calculates distance matrix and generates the rooted tree. It uses Kimura distance for the aligned pair and k-mer distance for unaligned. Distance matrices are clustered using UPGMA that improve tree by recalculating similarities. The following steps illustrate the MUSCLE algorithm.

Input: n Protein Sequences, S1,S2, ...Sn
Stage 1: This stage builds a draft progressive alignment.
1.1 (accuracy) it uses log-expectation score instead of PPS score in profile - profile alignment;
1.2 (efficiency) uses k-mer distance instead of alignment score for sequence similarity or by constructing a global alignment of the pair and determining the fractional identity
1.3 A tree is constructed from the distance matrix using UPGMA or neighbor-joining, and a root is identified.
Stage 2: This stage attempts to improve the tree and builds a new progressive alignment according to this tree
2.1 Use alignment to compute more accurate pairwise distance between sequences.
2.2 A tree is constructed by computing a Kimura distance matrix and applying a clustering method to this matrix.
2.3 From new distance matrix, build the guide tree and a new alignment.
Stage 3: Refinement: performs iterative refinement using a variant of tree-dependent restricted Partitioning
3.1 tries to improve alignment
3.2 The tree is broken into subtrees, and the sub-alignments refined
Output: n aligned Protein Sequences S'1, S'2,..., S'n

C. PROBCONS

PROBCONS (PROPabilistic CONSistency-based multiple alignments of amino acid sequences) is a novel tool for generating multiple alignments of protein sequences based on probabilistic consistency. PROBCONS has accomplished the most elevated correctness's of all alignment methods until now [16]. The main stages of PROBCONS are presented in the following:

Input: n Protein Sequences, S1, S2, ...Sn
Stage 1: Compute posterior probability matrices for each pair of sequences.
Stage 2: Compute the expected accuracy of each alignment.
Stage 3: Apply the probabilistic consistency transformation to posterior matrices.
Stage 4: Compute a guide tree using the expected accuracies.
Stage 5: Progressively align the sequences using the guide tree.
Output: n aligned Protein Sequences S'1, S'2,...,S'n

D. MAFFT

MAFFT (Multiple Alignment using Fast Fourier Transform) quickly identifies some of the more obvious regions of homology [13]. After identifying these regions slower dynamic programming approaches are utilized to join these portions into a full arrangement. Thus, the main advantage of the initial version of MAFFT was speed. It is one of the more accurate programs too. It is available as a standalone or web interface. It returns many output formats, including interactive phylogenetic trees.

Input: n DNA or n RNA or n Protein Sequences, S1,S2, ...Sn
Stage 1: calculation of a crude pairwise distance matrix based shared 6-tuples
Stage 2: construction of a UPGMA (Unweighted Pair Group Method with Arithmetic Mean) guide tree
Stage 3: dynamic programming used in progressive alignment with this initial guide tree
Stage 4: improved pairwise distance matrix inferred from the alignment of step 3
Stage 5: improved guide tree constructed from the new distances that were computed in step 4
Stage 6: A repeat of the progressive alignment algorithm (like step 3, but with the new guide tree).
Stage 7: Then MAFFT repeats the following:
7.1 break the alignment into 2 groups based on the tree
7.2 realign these groups
7.3 Accept this alignment if it improves the score.
Output: naligned DNA or naligned RNA or naligned Protein Sequences S'1, S'2,..., S'n

E. KALIGN

KALIGN is very similar to standard progressive methods. This technique depends on the Wu-Manber string-matching algorithm so it improves MSA speed and accuracy [14]. This algorithm calculates the pairwise distances, then construct a guide tree and based on tree order the sequences/profiles are aligned.

Input: n DNA or n RNA or n Protein Sequences, S1,S2, ...Sn
Stage 1: Apply Pairwise alignments by the k-tuple method to generate distance score matrix adopted from ClustalW.
Stage 2: The guide tree is constructed using either UPGMA or Neighbour-Joining method.
Stage 3: Progressive alignment by Wu-Manber approximate string-matching algorithm.
Stage 4: The distances between two strings are measured using Levenshtein edit distance
Output: naligned DNA or naligned RNA or naligned Protein Sequences S'1, S'2,..., S'n

F. RETALIGN

RETALIGN (RETicular ALIGNnment) is a progressive corner-cutting method for multiple sequence alignment. During the progressive alignment, it focuses on defining the set of optimal and sub-optimal alignment [17]. Therefore, it does not define the dynamic table compact part. This technique uses a network to store the alignments, so the alignments can be used in an efficient way during the progressive stage. This technique depends on the size of the network (threshold parameter). The better alignment scores mean larger the threshold parameter.

Input: n Protein Sequences, S1, S2, ...Sn
Stage 1: Build or load a guide tree for the sequences
Stage 2: Transform the sequences at the leaves of the guide tree into simple 'linear' networks
Stage 3: Visit the internal nodes of the guide tree in reverse traversal order. For each internal node ***v*** with children ***u1*** and ***u2***, labeled with the networks of alignments ***A1*** and ***A2***, respectively, calculate the x_v -network of ***A1*** and ***A2*** using the generalized Waterman-Byers algorithm
Stage 4: Return the best-scored alignment from the ***x***-network calculated at the root of the guide tree.
Output: n aligned Protein Sequences S'1, S'2,..., S'n

IV. BALIBASE DATASET

BALIBASE is a benchmark dataset that is used to measure the accuracy of MSA tools, which has more refined test cases. Protein datasets are available in TFA format in BALIBASE 3.0. The structure of the sequences are known and their reference alignment is available in the form of MSF format. Evaluation is carried out by comparing the Structure-based Reference Alignment with its Sequence Based tool alignment [20].

Table 3 introduces the main characteristics of BALIBASE 3.0. The BALIBASE dataset contains a C application program called bali_score that compute the SPscore (sum-of-pairs score) and TCscore (Total Column score) (ftp://ftp -igbmc.u -strasbg.fr/pub/BAliBASE3) [21].

Table 3. BALIBASE 3.0 characteristics.

Ref name	Seq identity	files number	Description
RV11	<20% identity	38	Sequences with variability length.
RV12	20-40% identity	44	
RV20	Up to 3 orphans	41	Family with Orphans
RV30	<25% residue identity	30	Divergent families up to 4 sub-groups
RV40	up to 400 residues	49	N/C-terminal large extension
RV50	up to 100 residues	16	Internal large insertions

The source code for the scoring schemes used here is available from ftp://ftp-igbmc.u-strasbg.fr/pub/msa_reference /bali_score_src_v4.tar.gz.

A. SPscore

Multiple sequence alignment is dealing with the alignment of greater than two sequences. To measure the quality/accuracy of multiple sequence alignment by giving it a score numeric value. For MSA, typically the most popular scoring method in bioinformatics called SP (sum-of-pairs) function. SP function of a multiple alignment S= (S_1, ..., S_N) is the scores summation of aligned pairwise sequences. MSA goal is to get the highest SPscore [22]. For example, the following is four DNA aligned sequence.

S1'="TACAT-AA"
S2'="-AC-TCA-"
S3'="AA-ATCAA"
S4'="TCATCAA"
SP(T,-,A,T)=score(T,-)+ score(T,A)+ score(T,T)+ score(-,A) + score(-,T) + score(A,T)=-2-1+2-2-2-1=-6;
SP(A,A,A,C)= score(A,A)+ score(A,A)+ score(A,C) +score(A,A)+ score(A,C)+ score(A,C)=2+2-1+2-1-1=+3

$$SPscore(a_1,\cdots\cdots,a_n)=\frac{\sum_{i,j}^{n} S(a_i,a_j)}{\sum_{i,j} S_r} \qquad (1)$$

where S_r is dataset reference score and $S(a_i, a_j)$ score between pairwise sequences a_i and a_j [23].

B. TCscore

TCsore is the total number of matched columns in alignment to reference alignment. The score C_i of the i[th] column is equal to 1 when column i in test alignment matches the same column in reference alignment r_i, otherwise, it is equal to 0, which m_r is a number of columns in reference alignment [22, 23].

$$S_i = \sum_{i}^{n} \begin{cases} 1 & if\ c_i = r_i \\ 0 & otherwise \end{cases} \quad (2)$$

$$TCscore = \sum_{i}^{n} \frac{S_i}{m_r} \quad (3)$$

V. Practical Results

In the experiments, the BALIBASE 3.0 references are used to evaluate MSA programs namely; CLUSTAL-OMEGA, MAFFT, KALIGN, BROBCONS, RETALIGN, and MUSCLE. To evaluate the performance of previous programs using SPscore and TCscore for each method.

A. Overall alignment accuracy evaluation

The overall accuracy of 218 test file alignment was measured using average SPscore and average TCscore.

The simulated datasets in BALIBASE 3.0 were applied to MSA tools. In this comparison, BROBCONS achieves the highest SPscore and TCscore followed by KALIGN and MAFFT algorithms as shown in Fig. 1. Table 4 recorded the average TCscore and SPscore for six alignment tools are recorded.

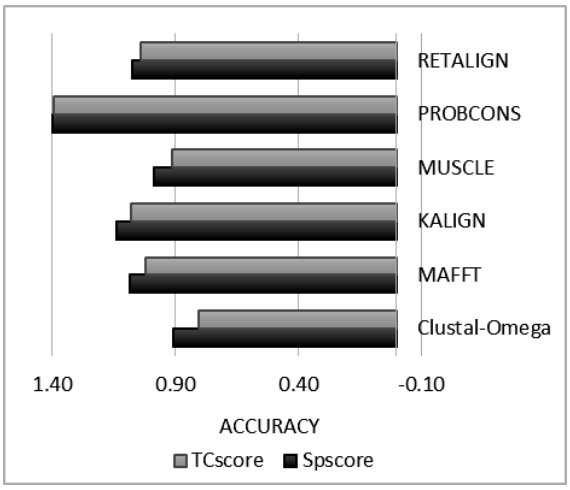

Fig.1. Accuracy Based on SPscore and TCscore.

B. Effect of number of sequences

To study the effect of sequence numbers on alignment accuracy, this evaluation measured using average SPscore for 4 to 40 sequences, 41 to 80 sequences, 81 to 120 sequences and more than 120 sequences. The accuracy has a weak effected when the number of sequences is increased as shown in Fig. 2. BROBCONS and MUSCLE achieve the highest average SPscores in this study of sequence number effect.

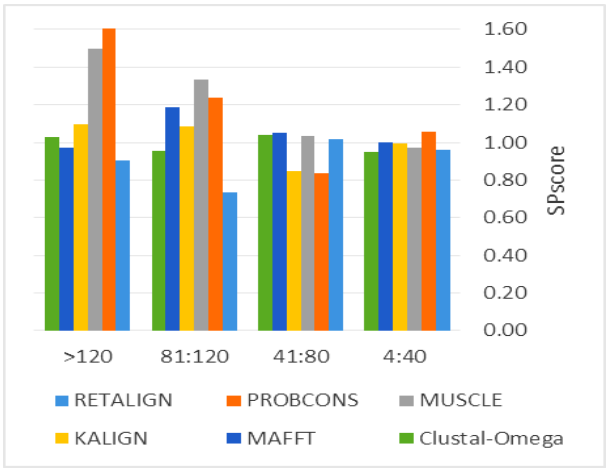

Fig.2. SPscore based on the average for a number of sequences

C. Effect of sequence length

Effect of sequence length was generated for all MSA tools. The sequence length in average is varying from 66 to 1630 KB. A weak effect of accuracy results based on sequence length is shown in Fig. 3. This study indicates that BROBCONS take the highest average SPscore.

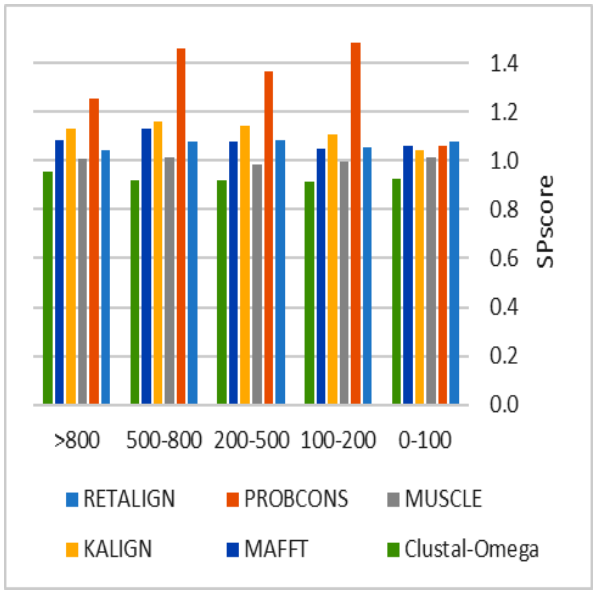

Fig.3. SPscore based on average for sequence length.

D. Effect of Execution time

The effect of overall alignment accuracy, the effect of sequences number and sequence length indicated that BROBCONS was achieved the highest accuracy. However, BROBCONS is the slowest one. BROBCONS execute at a maximum time and KALIGN faster than BROBCONS and MAFFT as shown in Fig. 4.

Table 4. The SPscore and TCscore of MSA programs on the benchmark BALIBASE 3.0 references

	RV11		RV12		RV20		RV30		RV40		RV50	
	SPS	TCS	SPS	TCS	SPS	TCS	SPS	TCS	SPS	TCS	SPS	TCS
CLUSTAL-OMEGA	0.80	0.65	0.96	0.85	0.92	0.82	0.93	0.83	0.99	0.90	0.86	0.76
MAFFT	1.03	0.97	1.03	0.96	1.11	1.05	1.08	1.02	1.15	1.09	1.11	1.05
KALIGN	1.00	0.95	1.08	1.01	1.26	1.19	1.29	1.25	1.11	1.05	1.10	1.04
MUSCLE	0.90	0.82	1.00	0.92	1.01	0.95	1.00	0.94	1.04	0.96	0.96	0.89
PROBCONS	1.17	1.17	1.07	1.02	1.57	1.55	1.50	1.52	1.56	1.53	1.52	1.56
RETALIGN	1.01	1.03	1.05	1.01	1.16	1.10	1.15	1.12	1.05	0.99	1.04	0.99

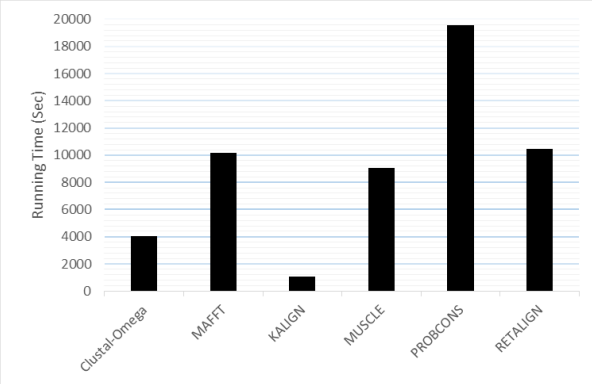

Fig.4. The average execution time of all alignment for BALIBASE datasets.

Table 5. Characteristic evaluation of MSA Tools

MSA Tool	Accuracy	Time
BROBCONS	The highest alignment accuracy	Highest
KALIGN	Less accuracy as compared with PROBCONS and MAFFT	Lowest
MAFFT	High alignment quality	Higher than KALIGN
MUSCLE	More accurate than CLUSTAL-OMEGA	Less time with a minimum number of iteration
RETALIGN	More accurate than CLUSTAL-OMEGA	Higher than KALIGN
CLUSTAL-OMEGA	Less accuracy	Less time

VI. Discussion

The evaluation results yield the more expected results. BROBCONS takes the highest time to complete alignments but produces the highest level of accuracy. The user must identify the main objective of alignment to select the suitable alignment tool. If user major objective is execution time then KALIGN is the best possible solution. On the other hand, the most accurate results are BROBCONS, KALIGN, and MAFFT. The major objective of MSA is finding the accurate similarity in less execution time. So the concept of parallel MSA is a suitable architecture to decrease execution time.

The evaluation results between MSA tools is dependent on many factors. In this paper, we focus on average execution time, similarity scores (SPscore and TCscore), number of sequences and average sequences length. Most of MSA tools treat of the dependence of the initial pairwise sequence alignment such as CLUSTAL-OMEGA.

This experiment showed that increasing sequence length and increasing number of sequences had a weaker effect on alignment results. Therefore, the fact of increasing sequence length and number of sequence did not achieve more accurate alignment results. Nevertheless, the study of the effect of sequence length and a number of sequences measured using SPScore represented that BROBCONS is the more accurate tool. Table 5 describes the characteristics comparison for the six MSA tools.

VII. Conclusion

This paper studies and evaluates groups of MSA tools. Some experiments are conducted using BALIBASE datasets for analyzing accuracy, execution time, effects of sequence length and number of sequences in these MSA tools. The Results showed that the accuracy of BROBCONS outperformed all the studied MSA tools, but it was a very slow tool. Among other tools, KALIGN, MAFFT, and RETALIGN gave the highest SPscore, respectively. Our prior analysis and evaluation results allow the user to select the suitable alignment tool and know the strength and weaknesses of six MSA tools. It is also necessary implemented these MSA in parallel because of existing a large amount of data and run time. The paper proposes that BROBCONS implementation using GPU will solve the time-consuming problem and will be efficient MSA system for large-scale datasets.

References

[1] S. B. N. a. C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence", Journal of Molecular Biology, Vol. 48, No. 3, pp. 443–453, 1970.

[2] T. F. S. a. M. S. Waterman, "Identification of common molecular subsequences", Journal of Molecular Biology, vol. 147, no. 1, pp. 195–197, 1981.

[3] G. B. a. D. G. H. I. M. Wallace, "Multiple sequence alignments," Current Opinion in Structural Biology, Vol. 15, No. 3, pp. 261-266, 2005.

[4] Arabi E. keshk,"Enhanced Dynamic Algorithm of Genome Sequence Alignments", IJITCS, vol.6, no.6, pp.40-46, 2014. DOI: 10.5815/ijitcs.2014.06.06.
[5] Jayapriya J, Michael A," A Novel Distance Metric for Aligning Multiple Sequences Using DNA Hybridization Process", International Journal of Intelligent Systems and Applications (IJISA), Vol.8, No.6, pp.40-47, 2016.
[6] Xu Li, ZhenzhouJi,"Efficient Parallel Design for Edit distance algorithm in DNA Sequence Alignment", IJEM, Vol. 1, No.4, pp.32-38, 2011.
[7] Gupta, S. K., Kececioglu, J. D. and Schaffer, A. A. "Improving the practical space and time efficiency of the shortest-paths approach to sum-of-pairs multiple sequence alignment", J. Comput. Biol., Vol. 2, pp. 459–472, 1995.
[8] S. Altschul, "Gap costs for multiple sequence alignment", J. Theor. Biol, Vol. 138, pp. 297–309, 1989.
[9] W. J. W. a. D. J. Lipman, " Rapid similarity searches of nucleic acid and protein data banks", Proceedings of the National Academy of Sciences of the United States of America, Vol. 80, No. 3, pp. 726-730, 1983.
[10] D.-F. F. a. R. F. Doolittle, "Progressive sequence alignment as a prerequisite to correct phylogenetic trees", Journal of Molecular Evolution, Vol. 25, No. 4, pp. 351-360, 1987.
[11] D. G. H. a. T. J. G. J. D. Thompson, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties, and weight matrix choice", Nucleic Acids Research, Vol. 22, No. 22, pp. 4673–4680, 1994.
[12] F. Sievers, A. Wilm, D. Dineen,T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Soding, J.D. Thompson and D.G. Higgins, "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega", Molecular Systems Biology 7: 539, 2011.
[13] K. K. a. D. M. Standley, "MAFFT multiple sequence alignment software version 7: improvements in performance and usability", Molecular Biology and Evolution, Vol. 30, No. 4, pp. 772–780, 2013.
[14] T. L. a. E. L. L. Sonnhammer, "Kalign—an accurate and fast multiple sequence alignment algorithm", BMC Bioinformatics, Vol. 6:298, 2005.
[15] R. C. Edgar, "MUSCLE: a multiple sequence alignment method with reduced time and space complexity", BMC Bioinformatics, Vol. 5:113, 2004.
[16] M. C.B. Do, "ProbCons: probabilistic consistency-based multiple sequence alignment," Genome Research, Vol. 15, No. 2, pp. 330-340, 2005.
[17] Adrienn Szabó, Ádám Novák, István Miklós, Jotun Hein," Reticular alignment: A progressive corner-cutting method for multiple sequence alignment", BMC Bioinformatics, Vol. 11:570, 2010.
[18] Hirosawa,M., Totoki,Y., Hoshida,M. and Ishikawa,M., "Comprehensive study on iterative algorithms of multiple sequence alignment", CABIOS, Vol. 11, pp. 13–18, 1995.
[19] S. M. I. X. e. a. P. Di Tommaso, " T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension", Nucleic Acids Research, Vol. 39, No. 2, pp. w13-w17, 2011.
[20] K. P. R. R. P. O. Thompson JD, "BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark", Proteins, Vol. 61, No. 1, pp. 36-127, 2005.
[21] J. P. F. P. O. Thompson,"BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs", Bioinformatics, Vol. 15, No. 1, pp. 87-88, 1999.
[22] J. Thompson, F. Plewniak, et al., "BAliBASE: A Comprehensive comparison of multiple alignment programs", Nucleic Acids Research, Vol. 27(13), pp. 2682-2690, 1999.
[23] J.D. Thompson, B. Linard, D. Lecompte and O. Poch, "A comprehensive benchmark study of multiple sequence alignment methods: Current challenges and future perspectives", PLoS ONE, Vol. 6, pp. 1-14, 2011.

Authors' Profiles

Eman M. Mohamed is a Ph.D. student at Menoufia University Faculty of Computers and Information, Egypt. She received his BSc. and MSc. in Computer Science from Menoufia University, Faculty of Computers and Information in 2008 and 2012. Her research interest includes Cloud Computing, Big Data, Bioinformatics, Data Privacy, and Security.

Hamdy M. Mousa received the B.S. and M.S. in Electronic Engineering and Automatic control and measurements from Menoufia University, Faculty of Electronic Engineering in 1991 and 2002, respectively and received his Ph.D. in Automatic control and measurements Engineering (Artificial intelligent) from Menoufia University, Faculty of Electronic in 2007. His research interest includes intelligent systems, Natural Language Processing, privacy, Security, embedded systems, GSP applications, intelligent agent, Robotics.

Arabi E. keshk received the B.Sc. in Electronic Engineering and M.Sc. in Computer Science and Engineering from Menoufia University, Faculty of Electronic Engineering in 1987 and 1995, respectively and received his Ph.D. in Electronic Engineering from Osaka University, Japan in 2001. His research interest includes software testing, software engineering, distributed system, database, data mining, and bioinformatics

How to cite this paper: Eman M. Mohamed, Hamdy M. Mousa, Arabi E. keshk, "Comparative Analysis of Multiple Sequence Alignment Tools", International Journal of Information Technology and Computer Science(IJITCS), Vol.10, No.8, pp.24-30, 2018. DOI: 10.5815/ijitcs.2018.08.04