A comprehensive performance analysis of various Multiple Sequence Alignment tools

Arnav Aima, Sakshi Dubey, Suhani Mehta
Bioinformatics, University of Florida

***Abstract*—Multiple Sequence Alignment (MSA) is a biological concept of aligning three or more sequences. It is of great importance for studying the genetic functions, structures and evolution process of the biological sequences. The aligning of multiple sequences is one of the most fundamental problems in Bioinformatics and is widely used in numerous biological applications like predicting the shapes of proteins, understanding functional sites in protein sequences and constructing phylogeny trees. MSA demands extremely sensitive computational methodologies to come up with precise results. The objective of this project is to conduct a comparative study on MUSCLE, T-Coffee and Kalign MSA tools on the basis of computation time and accuracy to identify the best use-case for each.**

***Index Terms*— MUSCLE, T-Coffee, Kalign, QScore, TC Score**

I. INTRODUCTION

THIS report throws light on the comparative study of Multiple Sequence Alignments. Multiple Sequence Alignment is the alignment methodology of two or more biological sequences. The objective of aligning MSA is to achieve a high Sum of Pairs (SP score) between the aligned sequences. There are various MSA tools to analyze the multiple sequence alignments which are widely used in bioinformatics. In this project, we have precisely focused on analyzing the performance of MUSCLE, T-Coffee & Kalign software. We have used BAlibase database to experiment with different tools. This project has taken SP scores, T Column Score, Cline's Score, and CPU time as the evaluative parameters.

II. MULTIPLE SEQUENCE ALIGNMENT TOOLS

A. MUSCLE: Multiple Sequence Comparison by Log-Expectation (MUSCLE) is a multiple sequence alignment tool which is used for aligning protein and nucleotide sequences.

Algorithm: MUSCLE algorithm works in three different stages. There is always an option to terminate the algorithm after entirely completing each step.

[1] Draft Progressive: The first stage is dedicated to building a progressive alignment. First, it calculates the similarity measure matrix. MUSCLE uses either counting of a k-mer method or constructing global alignment to compute the similarity of each pair of sequences & to determine the fractional identity. Furthermore, a triangular distance matrix is calculated with the help of pair-wise similarities.

[2] Improved Progressive: This is basically an iterative state. It is meant to improve the tree and to build a new progressive alignment. The second stage attempts to improve the tree and builds a new progressive alignment according to the previous tree. This is done to calculate the similarity of each pair of sequences in the current alignment. Now a tree is constructed with the help of Kimura distance matrix & clustering method. The improvement in MUSCLE is achieved if the tree converges and the iteration terminates.

[3] Refinement: The process of refinement in MUSCLE involves partitioning. Initially, sequences are divided into disjoint sets by deleting an edge. Now edges are traversed in decreasing distance from the root. Now there is a profile extraction phase. Current Multiple Alignment is used to extract the multiple alignments of each subset. The obtained multiple sequence alignments are now realigned. This is the final phase of the MUSCLE algorithm to finally accept or reject the solution. It calculates the SP score of the resulting alignment. The new alignment is retained if the score increases otherwise it is discarded. The algorithm will terminate if no change is retained while visiting all the edges. Another terminating condition could be that a

user-defined maximum number of iterations is achieved.

Complexity: In the first two stages of the algorithm, the time complexity is $O(N^2L + NL^2)$ and the space complexity is $O(N^2 + NL + L^2)$.In the final stage, the refinement stage also adds to the time complexity of another term, $O(N^3L)$.

MUSCLE usually gives better sequence alignments than some of the other available tools. Also, the idea of visiting edges in order of decreasing distance from the root has the effect of first re-aligning individual sequences than any other closely related groups.

B. KALIGN

Kalign[1] is a progressive alignment algorithm implemented in standard C, that uses multi-pattern matching and global dynamic programming with affine gap penalties. The unique feature of this algorithm is the usage of the approximate string-matching algorithm of Wu-Manber[2], which makes it very fast. An extension to Baeza-Yates-Gonnet algorithm[3], Wu-Manber string matching algorithm measures the Levenstein edit distance between two sequences. Complexity of this algorithm is $O(tk)$, where t is the text string length and k represents total errors allowed. Wu-Manber algorithm overcomes the drawbacks of standard k-tuple method, like failing to detect any similarity when there are no patches of high identity. Hence Kalign is extremely powerful for very large number of sequences that have high pairwise distance.

Kalign also pays attention to the locality of matches so it can disregard most of the spurious errors. Scoring matrix used is GONNET250. The algorithm first calculates the pairwise distances, which is then used to construct a guide tree that dictates the order in which sequences are aligned. It measures distance similarity as the sum of 3 highest scoring diagonals in order to exclude many of the spurious matches reported. These scores are utilized by UPGMA clustering method to construct the guide tree.

C. T-COFFEE

Tree-based Consistency Objective Function for Alignment Evaluation(T-Coffee) is a software to align Multiple Sequence Alignments via the library of pair-wise alignments. This method works on library alignment information which is used to guide the progressive alignment. This methodology uses intermediate alignments which are based on both immediate neighbors & how the whole alignment is aligned with each other. T-Coffee has an advantage of considering the information from all of the sequences at every step instead of considering information at that particular stage.

Algorithm:

1. It finds a library of sequences between pairs of sequences
2. It creates a new scoring matrix for each pair of sequences using the library
3. It uses the scoring matrix to indirectly align the pair of sequences using an extended library.

T-Coffee has many variations depending on the type of information like a comparison of sequences, pair wise sequence alignments, sequence profiles & protein structures. All the variation shave been evolved for balancing the tradeoff between computational cost & alignment accuracy.

Some of the popular T-Coffee variations have been discussed below.

[1] M-Coffee: M Coffee is a special type of T-Coffee that's used for evaluating & combining the most common multiple sequence alignment packages. M - Coffee is used to get a better alignment than the individual ones. It is well effective to reflect regions of high integrity where the various packages agree upon.

[2]Expresso and 3D-Coffee: This is carried out on the availability of structural information.

[3] R-Coffee: This kind of T-Coffee is mainly used to align RNA sequences while using secondary structure information

[4]PSI-Coffee: PSI-Coffee is used when accuracy is of prime importance. It is used to align the distantly related proteins using homology. It gives the most accurate results among all other kinds of T-Coffee but it is pretty slow.

[5]TM-Coffee: It aligns trans-membrane proteins using homology extension

T-Coffee is a progressive method for sequence alignment with many advantages. Firstly, it can combine sequences from heterogeneous sequence

alignments. It is observed that because of this nature local & global alignments can lead to a significant increase in accuracy. Secondly, T-Coffee uses a position-specific scoring scheme instead of substitution matrix for aligning sequences. The methodology of extending library lets the tool to analyze the information gathered from other sequences too. Thirdly, T-coffee uses a primary weighting scheme. One of the major shortcomings of T-Coffee method is that this tends to overweight small segments where high similarity could be a possible option. Apart from the fact that T-Coffee is not an ideal multiple sequence alignment tool, it is still widely used for carrying out bioinformatics research for aligning a wide range of multiple alignment sequences.

III. DATASET

This project aims at the comparative analysis of three different MSA tools namely MUSCLE, T-Coffee & Kalign. We have used the BAliBASE(Benchmark Alignment Database) to study our experimental observations. BAlibase is a database which is manually categorized in sub-blocks on different parameters like conservation sequence length, similarity, the presence of insertions and N/C terminal extensions, etc. Since BAlibase database contains high-quality documented alignment, it is also an excellent choice to identify the strong and weak points of the numerous alignment programs. The BAlibase Database has been divided into several hierarchical reference sets. All the subsets have been further divided into smaller sub groups according to their sequence length & percent similarity. We have used BAlibase 3.0 for our experimental work.
The different reference sets are organized in separate directories:

#RV	Sets of Sequences	#Reference	Details
RV11	38	Reference1	Equi distant sequences with very divergent sequences (<20% identity)
RV12	44	Reference 2	Equi-distant sequences with medium to divergent sequences (20-40% identity)
RV20	41	Reference 3	families aligned with a highly divergent "orphan" sequence
RV30	30	Reference 4	subgroups with <25% residue identity between groups
RV40	49	Reference 5	sequences with N/C-terminal extensions
RV50	15	Reference 6	internal insertions.

IV. EXPERIMENTATION

For experimentation purpose, we have used the BAliBase dataset sequence in fasta format. All the RV11. RV12, RV20, RV30, RV40, RV 50 set of sequences of BAlibase database are aligned using MUSCLE, T-Coffee & Kalign Multiple Sequence Alignment Tools independently. Furthermore, we compared the already aligned sequence of BAlibase database with the newly aligned results of mentioned three MSA tools. We have carried out the comparative study on the following parameters:

[1] QScore: QSCORE multiple alignment scoring software is used to compare two multiple sequence alignments. In Qscore, one alignment is considered to be test alignment and second alignment is considered to be the "reference" alignment. Q score is referred to as SPS (Sum of Pair Score) in BAlibase. It basically shows how many pairs of residues are aligned correctly.
For instance, if a match is scored as 2 and aligning residue with gap is scored as 1 in a comparative study, then the total score gets normalized by maximum possible score. Thus, the resultant values lies between 0 and 1. On the other side, this parameter has a drawback for practical use. It doesn't penalize over-alignment and it can give good score to an alignment that erroneously aligns non homologous regions.

[2] TC (Total Column) score: Total Column score is referred to as sum of scores of each column. It is basically the ratio of correctly identified columns in reconstructed alignment and total number of columns in reference alignment. Although, it is widely use for

practical experimentation yet its tendency to be very sensitive to misalignment limit its use.

[3]Cline's Score(CS): Cline's Score is a distance-based scoring scheme for pairwise alignments called shift scores. It also overcomes the drawbacks of SPS and TC scores. It can also take negative values in case of large shifts.

V. RESULTS & DISCUSSION

We have performed the experimentation of BAlibase database with MUSCLE, Kalign and T-Coffee. Here is the evaluative performance comparison on several parameters.

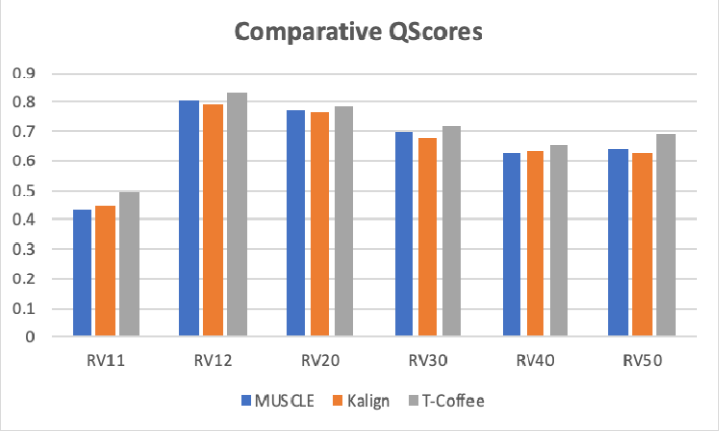

Figure 1: Comparative Q Score Graph

As we can observe from Figure 1, across all the data sets, T-Coffee gives the highest Q Scores (Sum of Pair Score). MUSCLE performs marginally better for RV12, RV20, RV30 and RV50, where as Kalign performs better for RV11 and RV40. Looking at the datasets themselves, we can say that Kalign performs better for highly divergent sequence as compared to MUSCLE. It is important to note that even though T-Coffee is consistently better across the datasets, the difference isn't too big. As per our study, we believe T-Coffee gives better result because it has the ability to align both the local and global sequences.

We have analyzed the MSA tools on behalf of Figure 2 for Total Column score. For Total Column (TC) score, Kalign under performs for all datasets apart from RV30. Comparing MUSCLE with T-coffee, there's only a marginal difference, apart from RV50, which has internal insertions. Kalign especially underperforms for RV12 dataset, which has medium to divergent sequences.

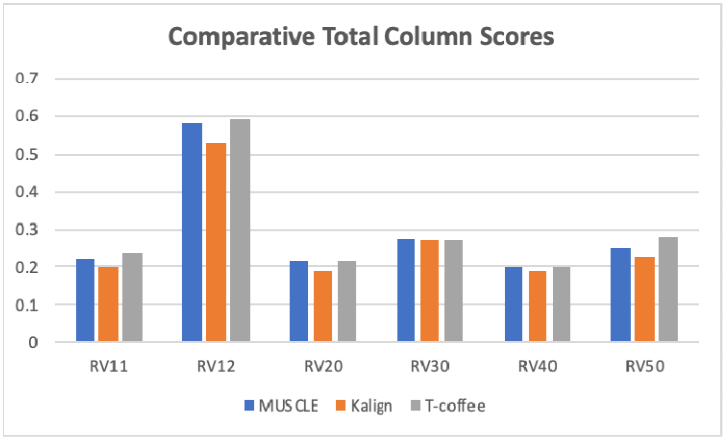

Figure 2: Comparative Total Score Graph

This underperformance could be because Kalign algorithm takes only top 3 highest scoring diagonals, in order to disregard many of the spurious matches reported.

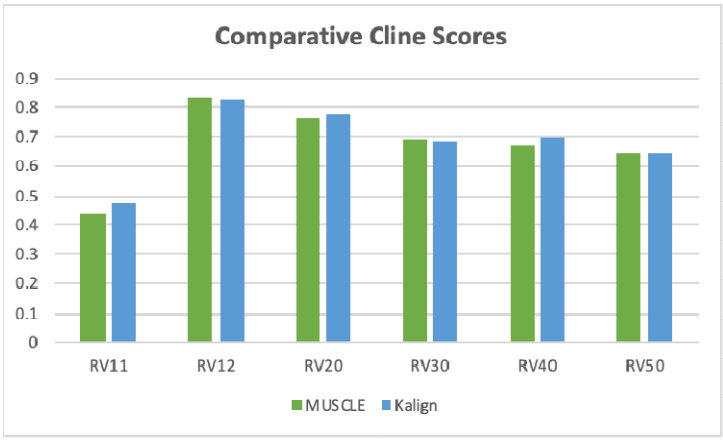

Figure 3: Comparative Cline Score Graph

As we can observe from the Figure 3, we have plotted Comparative Score Graph between MUSCLE and Kalign. For Cline score, there isn't much difference in the performance of MUSCLE and Kalign. Kalign performs better for highly divergent sequences of RV11 and RV20 and also for RV40. For RV12 and RV30 MUSCLE performs better but the difference isn't that noticeable. This suggests that performance of both algorithms is pretty similar, but Kalign is slightly more accurate for divergent sequences.

VI. CONCLUSION

When it comes to Q score, T-Coffee is the winner for BaliBase whereas Kalign and MUSCLE have pretty similar performance. For Total Column score, MUSCLE and T-Coffee were pretty similar and good enough. However, TC score is pretty sensitive to even a slight

shift in sequence. Hence it is easy to deduce that T-Coffee gives more accurate results.

However, when we consider the speed of execution, T-Coffee is extremely slow, in fact, it could be 50 times as slow as Kalign. So we recommend using T-Coffee only when extremely accurate results are needed, and time is not a constraint.

To make a better decision between MUSCLE and Kalign, we analyze the difference between their respective Cline Scores. Cline score overcomes the drawbacks of Q score and TC score. We can hence assert that even though there isn't a big difference, Kalign definitely performs better for highly divergent sequences. Combined with the fact that Kalign is almost 3 times faster than MUSCLE, we can safely say that amongst the 2, Kalign is more robust and fast.

The performance of Kalign is dependent on the underlying approximate string matching algorithm from Wu and Manber, which is extremely fast and a lot more accurate as compared to most other tools. Moreover, it outperforms for highly divergent sequences because it also takes into account the locality of the matches along with the total number. It is even more impressive when we consider that Kalign hasn't actually been trained on BaliBase dataset, unlike others. Hence we also recommend integrating the Wu-Manber string matching algorithm into MUSCLE.

Overall, we recommend using Kalign in most situations.

VII. DIVISION OF WORK

Task 1: Every team member was in charge of learning about one algorithm and executing the dataset on the corresponding MSA tool.
MUSCLE - Arnav
T-Coffee - Sakshi
KAlign - Suhani

Task 2: Converting the BaliBase reference files to FASTA format and calculating Q score is done by Arnav
Task 3: Calculation of TC Score and Cline score is analyzed by Suhani
Task 4: Collation of all materials, comparative conclusions & report writing is done by Sakshi

.**VI. REFERENCES:**

[1] Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32:1792–1797. doi: 10.1093/nar/gkh340.

[2] Lassmann T, Sonnhammer EL. Kalign--an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics*. 2005;6:298. Published 2005 Dec 12. doi:10.1186/1471-2105-6-298

[3] Notredame C, Higgins DG, Heringa J (2000-09-08). "T-Coffee: A novel method for fast and accurate multiple sequence alignment". *J Mol Biol*. 302 (1): 205–217. doi:10.1006/jmbi.2000.4042. PMID 10964570.

[4] Thompson J.D., Plewniak,F. and Poch,O. (1999) BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. Bioinformatics, 15, 87–88.

[5] S. Wu, U. Manber," A fast algorithm for multi-pattern searching," Tech. R. TR-94-17, Dept. of Comp. Science, Univ of Arizona, 1994

[6] Ricardo Baeza-Yates and Gaston H. Gonnet, \A New Approach to Text Searching", Comm. of ACM, 35 10, Oct. 1992, 74-82.

[7] Edgar RC (2004). "MUSCLE: a multiple sequence alignment method with reduced time and space complexity". *BMC Bioinformatics*. 5 (1): 113. doi:10.1186/1471-2105-5-113. PMC 517706. PMID 15318951

[8] Bawono P, van der Velde A, Abeln S, Heringa J. Quantifying the displacement of mismatches in multiple sequence alignment benchmarks. *PLoS One*. 2015;10(5):e0127431. Published 2015 May 19. doi:10.1371/journal.pone.0127431

[9] Shu N, Elofsson A. KalignP: improved multiple sequence alignments using position specific gap penalties in Kalign2. *Bioinformatics*. 2011;27(12):1702-3.
[10] *Loytynoja, A. (2005).* "An algorithm for progressive multiple alignment of sequences with insertions". Proceedings of the National Academy of Sciences. ***102*** *(30).* PMC 1180752. PMID 16000407