International Journal of Computer Sciences and Engineering Open Access
Review Paper Volume-5, Issue-3 E-ISSN: 2347-2693

A Comparative Study of Multiple Sequence Alignments

R. Karmakar[1], T.K. Sadhu[2*], A. Hazra[3], S. Sahana[4], S. Karmakar[5]

[1]Department of Computer Science, University of Burdwan, Burdwan, W.B, India
[2*]Department of Computer Science, University of Burdwan, Burdwan, W.B, India
[3]Department of Computer Science, University of Burdwan, Burdwan, W.B, India
[4]Department of Computer Science, University of Burdwan, Burdwan, W.B, India
[5]Department of Computer Science, University of Burdwan, Burdwan, W.B, India

[*]*Corresponding Author: timirsadhu97@gmail.com*

Available online at: www.ijcseonline.org

Received:28/Feb/2017 Revised: 06/Mar/2017 Accepted: 22/Mar/2017 Published: 31/Mar/2017

***Abstract*-Multiple sequence alignment is a very useful tool[18]. It is used to solve different Biological sequence alignment problem like DNA and Protein sequences[3]. There are many ways to solve multiple sequence alignment problems. Dynamic programming method is used to produce MSA directly[23]. Nowadays, progressive alignment approach and iterative approach are the important methods to solve MSA problems. This paper discussion is about some progressive alignment and iterative Multiple Sequence Alignment algorithm methods and compare their performances.**

Keywords- Progressive MSA, DNA, progressive alignment.

I. INTRODUCTION

In computational biology, sequence alignment is very useful for biologists to get many different useful information like, it can give information about various organisms and their evolution.It can also tell how these organisms are genetically related to other organism[2]. DNA, Protein and RNA sequences contain useful information to understand different organism and their evolution. For improvements of our understanding about the evolutionary relationship of various organism and distance among different organisms by comparing these sequences.

If two sequence are align together then this alignment is called pairwise sequence alignment.When the number of sequences is more than two then the alignment is called Multiple sequence alignment.These sequences have an evolutionary relationship with other sequences.The purpose of MSA is to detect similarities between all the sequences and evolve evolutionary relationships between these sequences. Multiple Sequence Alignment is used in many areas of bioinformatics like, it can find sequence family, structural relationship between sequences and connection between different elements of these sequences[1]. In Bioinformatics, Multiple Sequence Alignment help biologists to find the answer about a lots of problems. It is most useful in disease discovery, disease can be predict by comparing DNA with patients DNA.MSA problems leads to an NP-hard problem. Thus most existing methods use heuristics approaches due to the exponential time complexity of global optimization[22].

II. EXISTING ALGORITHMS

Using appropriate method or a programming technique most interesting problems evolve an extremely long sequences which are various in both length and residues that can be solved. In this paper, five different most popular algorithms are discussed by reviewing some existing paper for presenting these algorithms. These MSA algorithms are:

ClustalW

ClustalW one of the most common algorithm to produce MSA. Progressive alignment process is used by the ClustalW to perform a Multiple Sequence Alignment. Assuming, there are n sequences as input sequences to be aligned, ClustalW running through three major steps, name of this steps are Build distance matrix, construct guide tree and building MSA.

In first step, all the sequences are aligned separately to produce pairwise alignments and building a distance matrix from pairwise alignments to measure the divergence of each pair of sequence. $n(n-1) / 2$ pairs of sequences are aligned separately and the results will be used as pairwise distances. These pairwise distances will be collected to build the distance matrix.

Next step is to build a guide tree. A guide tree is constructed from distance matrix by using neighbour-joining algorithm.This tree is useful to guide the final multiple

alignment process. Weights are assigned on each sequence depends on their distance from the root of the tree.

Final step of this algorithm is, using the guide tree sequences are aligned progressively according to their branching order. At each of alignment, penalties ae assigned for opening and extending gaps by dynamic programming. Each steps align two existing alignments or sequences.this process is going continous to produce the final MSA result.

Complexity of clustalW algorithm, for performing k^2 global alignments, which takes $O(k^2n^2)$ time for performs neighbour-joining, which takes $O(k^3)$ time and for performs at most k-profile alignments, each takes $O(kn+n^2)$ time $O(k^2n+kn^2)$ Thus, total times taken by clustalW algorithm is $O(k^2n^2+k^3)$ time.

T-coffee

T-Coffee is a progressive alignment approach like ClustalW to produce the MSA.T-Coffee is a consistency based MSA algorithm to produce more accurate result than other MSA algorithms .it use two types data sources to take input from libraries.One is global data sources and another is local data sources The global alignments are constructed using ClustalW on the sequences, two at a time [1].Global data souces is useful when full length of input sequence are require and L-align program to produce one or more local alignments and collect these alignments into a library called primary library of alignments. Computing the primary library weights a weight to each pair of aligned residues in the library in order to reflect the correctness of an aligned residue pair. T-coffee considers each of these pairs as a constraint. Because having run two different programs on each pair of sequences, there will have two libraries lists of constraints, one for local alignments and one for global alignments. Combination of the libraries the efficient combination of local and global alignment information. This is achieved by pooling the ClustalW and L-align primary libraries simple process with 3 simples rule duplicated pairs will be merged into a single entry that has been weighted equal to the sum of the two weights otherwise a new entry is created pairs of residues that did not occur will be weighted as 0. T-coffee combines two global and local primary libraries into a merged library with new weights. T-coffee then examines the consistency of each pair of residues with residue pairs from all other sequences. T-coffee takes each aligned residue pair from the library and checks the alignment of the two residues with residues from the remaining sequences. If L is the average sequence length of the sequences then the complexity of building the extended library pair wise is $O(N^3L^2)$.The extended library will be used as input to the optimization procedure. From the extended library, pair wise alignments score are pooled out to build a distance matrix between all the sequences. This distance matrix used to build phylogenetic tree using neighbor joining method. Using the help of this phylogenetic tree the order of building MSA in such a way that the closest sequences in the tree will be aligned first. The closest two sequences on the tree are aligned first using dynamic programming. Guide tree will suggest whether the next closest two sequences are aligned or sequence is added to the existing alignment of the first two sequences. This process is running continues until all the sequences have been aligned.

The complexity of T-Coffee is, if the number of sequence is N and average sequence length is Lthen whole procedure takes $O(N^2L^2) + O(N^3L) + O(N^3) + O(NL^2)$ .Where, $O(N^2L^2)$ takes for computation pair-wise library,$O(N^3L)$ is for the extension, $O(N^3)$ is for the computation of the NJ tree and $O(NL^2)$ for the computation of the progressive alignment.

MUSCLE

In MUSCLE, progressive alignment is used to build the progressive MSA and the result is refined by iterative method .This algorithm is goes through three different stages.First stage is making a draft progressive alignment. Second stage is improved progressive alignment and the third stage refinement.

In draft progressive al
ignment going through four steps these steps are Similarity Measure, Distance estimate, Tree construction and Progressive alignment.similarity of each sequence is calculated using k-mer technique . After measures similarity between sequences Distance estimate can be calculated by estimate distance from pair sequence similarity After computing distance matrix a guide tree is constructed from the calculated distance matrix using neighbour-joining approach. From the guide tree a progressive alignment is constructed.

Next stage is improved progressive alignment, this stage is going through four steps. these steps are similarity measure, Tree construction, Tree comparison and Progressive alignment. In the first step of this stage is similarity measure, this time similarity measure between pairs of sequences is calculated using fractional identity method. After measures similarity, Build a new distance matrix by calculating kimura distance matrix between each pair of sequences. A new guide tree is build by clustering new distance matrix using UPGMA. Now the, previous tree and the present tree are compared to check changing branching order of the set of the corresponding nodes. If second stage is repeatedly executed to check the number nodes has decreased or not, then the iteration is terminate and a new progressive alignment is build. New alignments will recreated only for the changed nodes while the rest of the existing alignment is retained of each subtree for which the branching order is

unchanged. When the alignment at the root is completed, the algorithm may terminate.

Final stage of this algorithm is refinement**.**Refinement stage is build using four steps .These steps are bipartition, profile extraction, re-alignment and accept or reject.First step is bipartition An edge is deleted from the tree then tree divided into two subtrees and the sequences is divided into two subsets.The second step is profile extraction .MUSCLE extracts the profile, the multiple alignment of each subset, from the current multiple alignment in which columns with only gaps will be removed. Then the two profiles will be re-aligned yelling a new multiple alignment. third step is re-alignment, the Sum-of-pairs score of the new multiple alignment is calculated. If the score is higher then MUSCLE accept the new alignment otherwise the alignment will be rejected.
Complexity of MUSCLE, if sequence number is N and L is length of the sequences then draft progressive alignment stage and improved progressive alignment stage takes $O(N^2L+NL^2)$ times and the Refinement process stage takes $O(N^3L)$ time.

K-align
K-align is another alignment method to compute the distance between sequences .It is suited for aligning when sequence number is very large.This algorithm is more similar to the standard progressive methods for sequence alignment, such as first calculate the pair wise distances using k-tuple method like clustalw and other progressive methods, then a tree is build using either UPGMA or neighbour-joining method, and progressive alignment is constructed from the tree. The main difference between K-align and other algorithms is that this algorithm uses Wu-Manber approximate string-matching technique for distance calculation and dynamic programming is used to align the profiles,levenshtein edit distance is used in this method to allows string matching with mismatches and distance between two strings are[13].

MAFFT
MAFFT is an iterative progressive alignment method like MUSCLE. The sequences are aligned progressively using iterative approach. MAFFT used Fast Fourier Transform(FFT) method to finds efficient homologous regions. In this algorithm the distance matrix is calculated using the 6-mer method. MAFFT going through three different steps. First Step is FFT-NS-1,in this step 6-mer method is used to build the distance matrix, The second step is FFT-NS-2, in this step build a tree using the value of distance matrix.The quality of the tree is improved by constructing another guide tree from FFT-NS-1.The third step is FFT-NS-i, in this step the quality of the sequence are improve by the iterative method.In case of producing MSA ,MAFFT is a high speed algorithm for producing MSA.

III. PERFORMANCE OF MSA ALGORITHMS

Performance of the algorithms says many of the limitations and strengths of the different programs.Performance of MSA programs is depend on different aspects like CPU time, performance percentage and scalability. Also most of the programs performance is dependent on various factors, like sequence length and the similarity percentage between the sequences .

Table 1: Comparison of MSA algorithms

MSA Algorithms	**Algorithm method and features**	**Time complexity**	**Computation time**
ClustalW	Progressive technique	$O(k^2n^2+k^3)$	Less as compared to T-Coffee
T-Coffee	Progressive method with extended library	$O(N^2L^2)$ + $O(N^3L)$ + $O(N^3)$ + $O(NL^2)$	Highest
MUSCLE	Iterative method	$O(N^2L + NL^2+ N^3L)$	Depends on number of iterations.
K-align	Wu-manber string matching for distance estimation	$O(tk)$	Lowest
MAFFT	Fast Fourier Transform	$O(N^2)$	Higher than K-align but produce more accurate results

CPU time is the total amount of time required to align all the sequences in the benchmark.
To check the overall performance of the programs it depends on correctness of the results of all alignments. Scalability of the sequences are depended on handle and analyze the produced data properly.

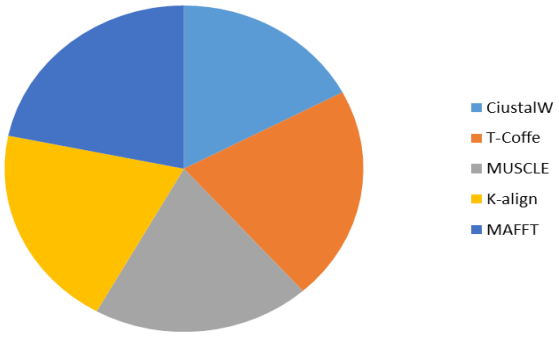

Fig.1: Overall performance chart

Performance evaluation of the programs describes different limitations and strengths of a program.

IV. DISCUSSION

In dynamic programming method to produce MSA for N sequences and each of length L takes complexity of $O(L^N)$. So the optimal solution is can be by produced using the dynamic programming , but the problem is that it takes exponential time to solve. In progressive alignment technique, complexity is reduced to $O(N^2)$. Now at present days, many protein families sequences lengths greater than 50,000. Hence it is very hard to solve within reasonable time. To solve this problem some faster algorithms are introduced like K-align algorithm. But when the sequence lengths are small, K-align is extremely fast. In MUSCLE, after performing MSA on 3000 sequences, the program becomes extremely slow. T-Coffee takes the most time to complete the alignments, but the results obtained have a high level of accuracy. MAFFT gives the optimal solution of all the given sequences, it gives an optimal solution without taking much computational time. If users try to find the best possible solution in minimum time, then K-align is the best option out of the five programs. When MAFFT algorithm is compared with other algorithms, MAFFT works efficiently, as it generates results from huge data in a shorter period of time. On the other case, in order to construct MSA for sensitive data, the best solution is T-Coffee or MAFFT, as both of them give the more similar accurate result. The advantage of ClustalW algorithm is that there is no limitation on sequence length any number of sequence length can be used, but the disadvantage of ClustalW is it gives less accurate or scalable than other modern programs. Advantage of MUSCLE algorithm is that also like ClustalW there is no limitation on the number sequences , it is faster than ClustalW and produce more accurate than Clustal and useful when amount of data is huge. disadvantages of MUSCLE algorithm are it acceptable format as input sequences is limited to only FASTA format. Advantages of T-Coffee algorithm is, it is useful when high accuracy and quality of the alignment high is needed. There are also many other useful features that T-Coffee able to do, compare with the other MSA tools, but the disadvantage is, it use limited number of input sequences that can be-aligned when compared to other algorithms and takes more CPU time to produce the MSA results.

V. 5. CONCLUSION

The goal of this study is to compare different MSA algorithms. In order to select the proper algorithm which gives the best result as per user require. To get a better result user can choose a MSA algorithm on the basis of his objective. By comparing MSA alignment algorithms results were compared ClustalW is a good alignment algorithm but it is less accurate and scalable when compared to other programs. T-Coffee is used when high accuracy is required and the sequence that can be-aligned is limited. MUSCLE used to improve accuracy using iterative approach, But when the sequence number is very large, iterations are reduced to get the results in reasonable time. MAFFT achieves the highest alignment quality score, whereas K-align alignment quality is reduced to attain results in the best computational time. Now at present time, to get a best MSA solution by combined the strengths of these different algorithms.

VI. ACKNOWLEDGMENT

We are very much thankful to Rahul Karmakar sir,because of his continuous support, patience,motivation and immense knowledge, useful guidance, advice, encouragement and suggestions during this study.

REFERENCES:

[1] A. Sedaghatinia, R.B. Atan R B, K.T. Arifin, Masrah, "*Comparison and Evaluation of Multiple Sequence Alignment Tools In Bininformatics*" ,IJCSNS International Journal of Computer Science and Network Security, Vol.9 No.7, July 2009.

[2] U. Manzoora, S. Shahidb, B. Zafara, *"A comparative analysis of multiple sequence alignments for biological data*", Pakistan. Bio-Medical Materials and Engineering 26 (2015) S1781–S1789.

[3] T. Lassmann, E.L. Sonnhammer, "*Kalign – an accurate and fast multiple sequence alignment algorithm*", BMC Bioinformatics 61-9 (2005),.

[4] H. Carillo, D. Lipman, "*The multiple sequence alignment problem in biology*", SIAM Journal of Applied Mathematics, Vol 48, 1988, pp. 1073-1082.

[5] E.V. Koonin, "Darwinian evolution in the light of genomics", Nucleic Acids Research 37 (2009), 1011–1034

[6] A. Layeb, M. Selmane, E.B. Elhoucine, "*A new greedy randomized adaptive search procedure for multi-objective RNA structural alignment*", International Journal in Foundations of Computer Science & Technology (IJFCST)),9-24, 3 (2013.

[7] V.K. Sohpal, A. Singh, A. Dey, "*Optimization of substitution matrix for sequence alignment of major capsid proteins of human herpes simplex virus*", International Journal of BioAutomation 15 (2012), 277-284.

[8] T. Lassmann, E.L. Sonnhammer, "*Kalign – an accurate and fast multiple sequence alignment algorithm*", BMC Bioinformatics 6 (2005), 1-9.

[9] R.M. Potter, "*Constructing Phylogenetic Trees using Multiple Sequence Alignment*",University of Washington (2008).

[10] A. Ray, B. Kartikeyan, S. Garg, "*Towards Deriving an Optimal Approach for Denoising of RISAT-1 SAR Data Using Wavelet Transform*" ,International Journal of Computer Sciences and Engineering journal Volume-4, Issue-10, 2016.

[11] M. Iain, E. Wallace, "*Evaluation of Iterative Algorithms for Multiple Sequence Alignment*",Bioinformatics Oxford Journal, Vol 21 No 8 2005.

[12] Y. Chen, P. Hou, B. Manderick, "*An ensemble self-training protein interaction article classifier*", Bio-Medical Materials and Engineering 24 (2014), 1323–1332.

[13] C. Notredame,D.G. Higgins, J. Heringa, "*T-Coffee: a novel method for fast and accurate multiple sequence alignment*", J Mol Biol, 302:205-217. ,2000.
[14] Katoh K, Miyata T,Kuma K and Toh H., "*Improvement in the accuracy of multiple sequence alignment program MAFFT*", Genome Informatics 16 , 22-33,2005
[15] R.C. Edgar, "*MUSCLE: multiple sequence alignment with high accuracy and high throughput*". Nucleic Acids Research, 32:1792-1797, 2004.
[16] M. Kumar, "*A Comparative Study on the Alignment Quality of Multiple Protein Sequences*" J. Pharm. Sci. & Res., Vol. 7(6), 2015, 314-318.

Authors Profile

Mr. Rahul Karmakar pursed M.Tech in Computer Science and Engineering in 2006 and Currently work as a Assistant Professor in Burdwan University.

Mr. Timir Kumar Sadhu pursed Bachelor of Science from University of Burdwan,Burdwan in 2015 and currently pursuing Master of Science from Burdwan University.

Mr. Avijit Hazra pursed Bachelor of Science from University of Sidhu Kanhu Birsa University,Purulia in 2015 and currently pursuing Master of Science from Burdwan University.

Mr. Susanta Sahana pursed Bachelor of Computer Application from University of Burdwan,Burdwan in 2015 and currently pursuing Master of Science from Burdwan University.

Mrs. Sumana Karmakar Sahana pursed Bachelor of Computer science from University of Burdwan,Burdwan in 2015 and currently pursuing Master of Science from Burdwan University.

