Performance Evaluation of Leading Protein Multiple Sequence Alignment Methods

Arunima Mishra, B. K. Tripathi, S. S. Soam

Abstract: Protein Multiple sequence alignment (MSA) is a process, that helps in alignment of more than two protein sequences to establish an evolutionary relationship between the sequences. As part of Protein MSA, the biological sequences are aligned in a way to identify maximum similarities. Over time the sequencing technologies are becoming more sophisticated and hence the volume of biological data generated is increasing at an enormous rate. This increase in volume of data poses a challenge to the existing methods used to perform effective MSA as with the increase in data volume the computational complexities also increases and the speed to process decreases. The accuracy of MSA is another factor critically important as many bioinformatics inferences are dependent on the output of MSA. This paper elaborates on the existing state of the art methods of protein MSA and performs a comparison of four leading methods namely MAFFT, Clustal Omega, MUSCLE and ProbCons based on the speed and accuracy of these methods. BAliBASE version 3.0 (BAliBASE is a repository of manually refined multiple sequence alignments) has been used as a benchmark database and accuracy of alignment methods is computed through the two widely used criteria named Sum of pair score (SPscore) and total column score (TCscore). We also recorded the execution time for each method in order to compute the execution speed.

Keywords: Multiple Sequence Alignment, Execution speed, Sum of pair score, Total column score.

I. INTRODUCTION

There are two types of methods to perform sequence alignment. One is the Pairwise sequence alignment (PSA) method and the other is multiple sequence alignment (MSA) method. PSA is a procedure to expose the evolutionary relationship between two sequences of DNA, RNA or protein, by finding out the maximum similarity between the sequences. It was first done by Needleman and Wunsch (1970) [1] leveraging the dynamic programming algorithm that can provide globally optimum alignment of two full-length protein sequences. Smith and Waterman (1981) [2] introduced another algorithm with a different scoring scheme where in place of global alignment optimum local alignment of sub-sequences are used. There are many PSA methods such as BLAST [3], AlignMe[4], EMBOSS [5] and PSI-BLAST[6].MSA is a well-known method of alignment of three or more biological sequences. Multiple sequence alignment is a very intricate problem, therefore, computation of exact MSA is only feasible for the very small number of sequences which is not practical in real situations. Dynamic programming as used in pairwise sequence method is impractical for a large number of sequences while performing MSA and therefore the heuristic algorithms with approximate approaches [7] have been proved more successful. Generally, various biological sequences are organized into a two-dimensional array such that the residues in each column are homologous or having the same functionality. Many MSA methods were developed over the period of time such as CLUSTALW [8], webPRANK[9], PAGAN[10],Clustal Omega[11], MUSCLE [12,13],T-Coffee[14], MAFFT[15,16], PRRP[17], ProbCons[18] , PROMALS [19], PROMALS3D[20],GLprobs[21],MSAprobs[22], and NAST [23] etc. MSA methods are generally applied for the identification of family of a new protein, locating DNA regulatory elements such as binding sites, protein structure prediction, and phylogenetic analysis. The assumption behind the MSA that actually help in performing these functions is that all the biological sequences that are aligned share evolutionary relationship. The reliability of alignment results plays a key role in such analyses, but it is seen that results generated by various methods of alignment are pretty different. [24]. Several studies are earlier done on the assessment of MSA methods by taking BAliBASE [25], HOMSTRAD [26] as benchmark datasets. The output of these studies pointed towards the fact that none of the existing MSA method was suitable for all type of datasets. [27,28,29], Each MSA method has its own advantages and limitations. BAliBASE is a repository of manually refined correct alignments of conserved residues with 3D structural super-positions. HOMSTRAD is also a benchmark database which provided protein sequences and their structure information, extracted from various protein databases that stored structures and families of protein like- PDB [30], Pfam [31] and SCOP [32], etc. Some studies are also done using simulated databases but the main disadvantage of the simulated sequences is that if the settings made for simulation are more similar to any alignment method than the other, they can produce the results in favour of that method [33].

Revised Manuscript Received on October 30, 2019.

* Correspondence Author
Arunima Mishra*, computer science & engineering, Dr APJ Abdul Kalam Technical University, Lucknow, India. Email: arunimamishra02@gmail.com
Bipin Kumar Tripathi, computer science & engineering, Rajkiya Engineering college , Bijnor, India. Email: abkt.iitk@gmail.com
Sudhir Singh Soam, computer science & engineering, Institute of Engineering and Technology, Lucknow, India. Email: sssoam@gmail.com

We have used BAliBASE v3.0 as benchmark database which contained specially constructed reference sequence sets of protein and corresponding sets of alignments which are meant with the help of 3-dimensional superpositions of protein structures and then refined manually to ensure the accuracy of the MSA. Apart from the benchmark dataset, a good means for comparison also required and hence we chose two broadly accepted score sum-of-pair score (SPscore) and total -column score (TCscore) for the assessment of accuracy of four alignment methods. SPscore shows the level to which an alignment program is able to align the sequences in a most accurate manner into an alignment, instead TCscore access the capability of the alignment program to align all the residues correctly in each column. SPscore is helpful to assess the alignment of sequences where the data set is divergent, whereas the TCscore is more accurate with an alignment of sequences that are closely related. In the event when the datasets are closely related with few very divergent or orphan sequences and if any tool aligns only closely related sequences correctly and misalign the divergent or orphan sequences, will get the high SPscore which is not the correct representation. In such scenarios, the TCscore gives more meaningful result as it is able to distinguish between the alignment programs which could align the highly divergent or orphan sequences correctly.

To perform analysis using MSA, it is imperative to use the right method as right choice of a method can lead to the correct results and vice versa, therefore, it seemed essential to conduct a comparative study of various MSA methods on the speed and accuracy of the results, which will help even the non-specialist biologists to select the right method . This study focuses on the critical issues like accuracy and speed of four popular MSA methods MAFFT v7, ClustalΩ, MUSCLE and ProbCons by the systematic comparison and evaluation. Accuracy and speed are the standard parameters used for evaluation of methods. The main algorithms and availability of MSA methods are mentioned in table 1.

Table 1: Description of MSA methods chosen for evaluation

Tools name/version	Main Algorithm	Link
Clustal Omega/1.2.4	Progressive method followed by HMM	http://www.clustal.org/omega/
ProbCons/1.12	Markov model based Progressive alignment	http://probcons.stanford.edu/download.html
MAFFT v7	Progressive alignment with Iterative refinement	http://mafft.cbrc.jp/alignment/software/
MUSCLE/3.8.31	Iterative	http://drive5.com/MUSCLE/downloads

II. STATE OF ART METHODS OF MSA

In literature, different MSA algorithms have been proposed, based on different methods. For example, progressive methods, Iterative methods, Hidden Markov models, Phylogeny aware methods, incorporation of 3D structure of protein and Optimization techniques for the refinement of the solution obtained by any other method of MSA.

A. Progressive alignment

Da-Fei Feng and Doolittle [34] developed a heuristic approach for multiple sequence alignment in 1987 called progressive alignment. This approach performs multiple sequence alignments in two stages, first stage builds a guide tree by any clustering method like neighbour-joining [35] or UPGMA [36], second step constructs the final MSA by adding sequences with the help of guide tree i.e. adding most similar sequences first followed by the most distantly related sequences. Progressive alignments do not guarantee for globally optimal results. but are very fast and efficient enough when a large number of protein sequences are there. The main disadvantage of this method is that if errors are made at any stage, it cannot be rectified later on and propagated through to the final result. Clustal W is the most used progressive alignment-based tools.

B. Iterative- Progressive alignment Method

Iterative-Progressive alignment in contrast to the progressive alignment method repeatedly performs dynamic programming to realign the already aligned sequences and at the same time keeps adding new sequences to the MSA. This method due to iterative dynamic programming, improves the accuracy of the alignment and is able to correct the errors made at the initial stage. The limitation of this method is that Iterative methods cannot perform alignments for a large number of sequences [37]. Key iterative alignment methods include PRRP, MUSCLE, and Dialign[38].

C. Hidden Markov model

Hidden Markov model is a graphical model that use probability to predict a sequence of unknown variables from the set of observed variables. It is extensively used in multiple sequence alignment. This model assigns the probability of various combinations of alignments with matches, mismatches, and gaps at different positions to arrive at the possible MSAs. it provides as an output a single highest scoring MSA and also other possible alignments that can be analyzed further for other biological implications.

The accuracy of the output is dependent upon the posterior probabilities of residues. By using this model, the computing speed is improved for the sequences having overlapping regions. Examples of Hidden Markov model are ProbCons, GL prob and MSAprobs etc.

D. Phylogeny aware method:

Phylogeny aware methods of sequence alignment gather the evolutionary information of the sequences and use it for distinguishing gaps occurrences caused by insertions or deletions in the course of alignment and treat them differently, eventually this correct handling of gaps results into good alignments from evolutionary point of view. Nevertheless, they run slowly compared to progressive and iterative methods. A few examples are PRANK [39], webPRANK and PAGAN.

E. Optimization method

There are two optimization techniques primarily used in MSA "the genetic algorithm (GA)" and "simulated annealing". Genetic algorithms in multiple sequence alignment simulates the evolutionary process. It first generates a possible set of MSAs as population and thereafter repeatedly exchanges the fragments of MSAs with the insertion of gaps of varying sizes. These random insertion and fragment exchange give rise to the diversity in the next generation of the population of MSA. A general objective of this process is to maximize the value of the objective function, like the sum of pair or total column score. SAGA [40], RAGA [41] and PASA [42] etc are MSA methods which are implemented with GA. Simulated annealing is a method analogous to the term for a metal to be cool and anneal in thermodynamics. This method improves the quality of existing MSA which is produced by any other method by rearranging the alignments in order to maximize the objective function like sum of pair score or total column score MSASA [43] (Multiple Sequence Alignment by Simulated Annealing) program is an example of this technique.

III. EVALUATION CRITERIA & BENCHMARK DATABASE:

In this paper, we've chosen the SPscore and TCscore as two parameters for the assessment of the accuracy of the four alignment methods.

The SPscore is the ratio of sum-of-pair of, test alignment and reference alignment. Sum-of-pair is calculated by adding the values obtained from the structure-based matrix for all the possible pairs of symbols in each column. The value of SPscore lies between 0 to 1 ('1' if the test alignments are identical to reference alignments and '0' if it does not match at all). Thus, the high value of SPscore shows better alignment.

SPscore = sum-of-pair of test alignment/sum-of-pair of reference alignment

Sum-of-pair = $\sum_{j=1}^{n-1}\sum_{k=j+1}^{n} S(j,\ k)$ + Gap-Penalty

Here 'n' is the number of sequences in the alignment and S(j,k) is the score for jth and kth sequence, where score for match/mismatch of residues while comparing two sequences, is obtained using structure-based substitution matrix, like BLOSSUM[44] or PAM[45] with addition of the Gap-Penalty. The Gap-Penalty is a fine which is incurred due to one or more successive gaps in the alignment. There are primarily two types of Gap-Penalties, one is constant Gap-Penalty where a constant negative value is given for each gap in the alignment and the second is affine Gap-Penalty which can be shown as-

$$A_{f} = X+Y.d$$

where X is a fine for the opening of the gap and 'Y' is the fine for extending the gaps and d is the number of continuous gaps.

The summation of scores for each possible pair of sequences gives the SPscore of the full MSA.

The TCscore is obtained by the division of the number of columns (where symbols are correctly aligned with respect to the reference alignment) with the total number of columns in the MSA. It is a binary function that returns the value as '1' if the column in test MSA (Ti) matches with the column in reference MSA (Ri) and '0' if it does not match.

$$S = \sum_{i=1}^{d} \begin{cases} 1 & \text{if } Ti = Ri \\ 0 & \text{Otherwise} \end{cases}$$

Here i ranges from 1 to d (d is the length of multiple sequence alignment) and therefore TCscore of the MSA given as-

$$TCscore = S/d$$

We have used a C language program bali_score which is a part of BAliBASE v3.0 and is available at http://www.lbgi.fr/balibase/ for the calculation of the SP score and TC score.

BAliBASEv3.0 is a benchmark database specially constructed for the benchmarking of MSA methods. Details of reference datasets are listed in Table 2. It is divided into 5 reference sets of sequences and their respective multiple sequence alignments. Reference set 1 comprises of equidistant sequences with highly different and medium divergent sequences (RV11 <20% identity and RV12 is having 20-40 % identity). Reference 2 contains sequences belong to the number of families with few very divergent orphan sequences. A sequence in a group is called an orphan sequence if it has <20 identity with respect to the other sequences while all other sequences are having >40% identity with each other. Reference 3 comprises groups of sequences having <25 % of similarity between the groups. This reference set was created to demonstrate the capability of the MSA methods to align the different groups of sequences having approximately equal distances with each other, into a single MSA. Reference 4 comprise sets of sequences that share>20% identity and having N/C-terminal extensions on the other hand reference 5 also has >20% identity with internal insertions. All the reference sets comprise full-length protein sequences.

Table 2: Description of reference sets in benchmark database Balibase3.0.

Ref Number	Name	Number of Reference sequence sets	Description
1	RV11	38	<20% identity
	RV12	44	20-40% identity
2	RV20	41	>40 % identity with orphans
3	RV30	30	Different families with <25% identity
4	RV40	49	N/C-terminal large extension
5	RV50	16	Internal insertions

IV. METHOD OF EVALUATION:

To perform the comparative study of selected MSA tools, below steps were performed:

A. Data extraction

we downloaded six datasets from BAliBASE v3.0 that contained a total of 218 reference sequence sets and their reference alignments in msf format. Description of reference sequence set is given in table 2.

B. Generation of test alignments

The 218 reference sequence sets obtained in step 1 were run on the four methods namely Clustal Omega, MAFFT, ProbCons, MUSCLE. We selected the default parameters and protein alignment while executing all the four methods, only MAFFT ran on auto mode. In auto mode, MAFFT chooses the most suitable method based on the size of the dataset. For small datasets, MAFFT chooses more accurate mode (L-INS-i) and for small datasets, it chooses high in speed but less accurate mode (FFT-NS-2). Each method generated 218 test alignments and a total of 872 (218X4) test alignments. We have recorded the time of execution for each MSA of all 218 test alignments.

C. Accuracy evaluation

SPscore and TCscore are the two parameters which has been used to evaluate the accuracy of each MSA method out of the four chosen methods. 218 test alignments and the same number of reference alignments are provided as input to the bali_score program which is a part of BAliBASE3.0 the resultant SP score and TCscore were produced for each test alignment generated by respective methods. High value of SP score and TCscore shows better alignment and vice versa. Fig1 depicts the framework for evaluation method.

D. Comparison and analysis

Analysis is performed by comparing each method on the basis of accuracy and speed where the accuracy is measured based on the SPscore and TCscore and speed is measured on the basis of time recorded while executing protein sequences in respective methods.

E. Hardware specification

2.2-GHz i3-2330M processor with 6 GB RAM.

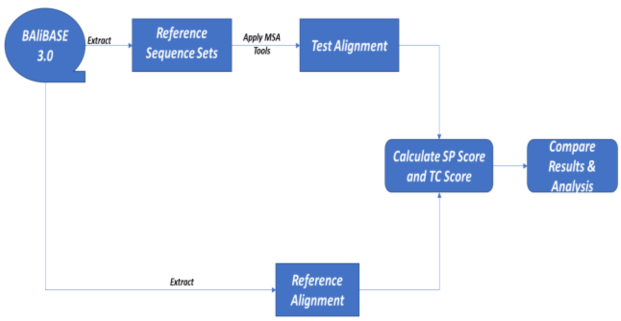

Fig 1: Framework of comparative study of MSA tools

III. RESULTS AND DISCUSSIONS

We performed multiple sequence alignments using 4 MSA methods on the six benchmark sequence sets. Results of SPscore and TCscore showed (see table 3) that for RV11 (sequences lies in twilight zone,<20% of similarity) and RV12(20 to 40 % identity) ProbCons have slightly higher values for SP score and TC score than MAFFT, MAFFT comes next in accuracy for both of the reference sequence sets but is five times faster for RV11 and two times faster for RV12 Sequence set than ProbCons.The value of SP score and TCscore is highest for RV20 for all tools due to the equidistant sequences with 20-40% identity, for RV20 and RV30 MAFFT performed better according to SPscore and TCscore and also six-time and four times faster than ProbCons respectively. RV40 sequences sharing minimum 20% residue identity and high N/C-terminal extensions SP score is almost the same for Clustal Omega, MAFFT, and PobCons but TC score is high for MAFFT among all and Clustal Omega is fastest among three. RV50 sequences are sharing a minimum of 20% residue identity and internal insertions, MAFFT and ProbCons have the same higher values of SP scores but MAFFT has slightly higher value for TCscore than ProbCons fig4 shows that MAFFT is slower than ProbCons in this case. Clustal Omega is the fastest method among all. MUSCLE comes next to Clustal Omega in terms of speed, fig 2, fig3 and fig4 shows the average values of SPscore, TCscore and time respectively for each method across the data sets.

Table 3: Average values of the SPscore, TCscore and execution time for each tool across the benchmark datasets.

Tool Name	Parameters	Sequence Sets					
		RV11	RV12	RV20	RV30	RV40	RV50
ClustalO	SPscore	0.45284	0.82670	0.77210	0.68167	0.75682	0.68175
	TCscore	0.24737	0.68318	0.30561	0.37300	0.42816	0.41729
	Time	0.00027	0.00055	0.00002	0.00004	0.00002	0.00002
MAFFT	SPscore	0.50184	0.85336	0.80183	0.69330	0.76073	0.75760
	TCscore	0.28526	0.71523	0.31732	0.36933	0.43694	0.42771
	Time	0.00086	0.00106	0.00004	0.00033	0.00074	0.00075
MUSCLE	SPscore	0.44263	0.82714	0.76629	0.64700	0.72600	0.72458
	TCscore	0.22868	0.67909	0.24854	0.24600	0.33837	0.34479
	Time	0.00048	0.00056	0.00004	0.00009	0.00004	0.00004
ProbCons	SPscore	0.52368	0.85877	0.78571	0.67077	0.75943	0.75700
	TCscore	0.29605	0.73455	0.29122	0.34333	0.40449	0.39521
	Time	0.00369	0.00197	0.00425	0.00333	0.00045	0.00045

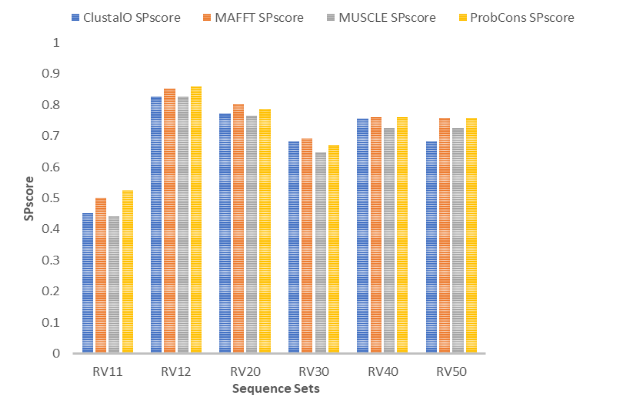

Fig 2: Shows average value of SPscore for each tool across 6 benchmark datasets

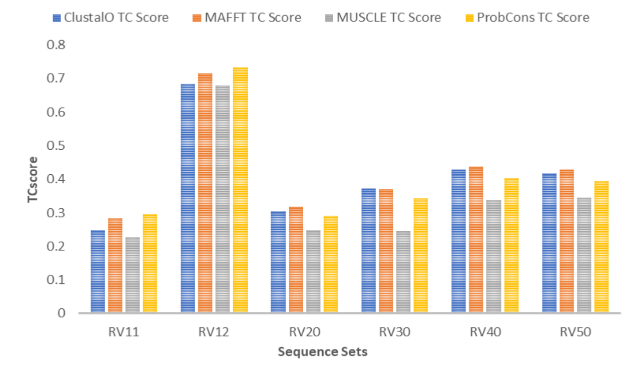

Fig 3: Shows average value of TCscore for each tool across 6 benchmark datasets

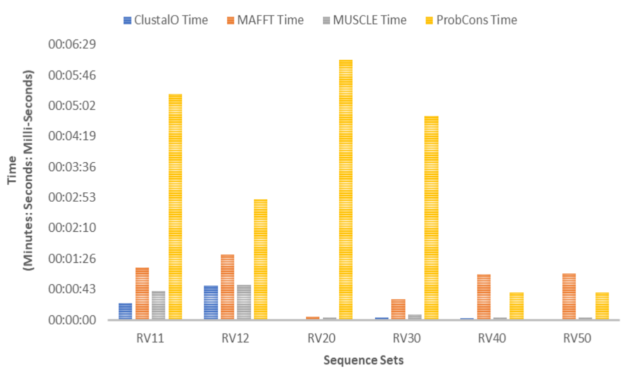

Fig 4: Shows average value of Time for each tool across 6 benchmark datasets

VI. CONCLUSION

Multiple sequence alignment is an extensively applied method and is the first step in the analysis of many bioinformatics functions including secondary structure prediction of protein, identification of functional sites, phylogenetic analysis, and sequence database searching, etc. None of the sequence alignment methods gives the exact solution, therefore, the selection of the MSA method is difficult for biologists and naive users. In this research paper we have performed an evaluation of leading protein multiple sequence alignment methods with the help of benchmark dataset, to test the performance on the basis of their sum of pair score, total column score and time consumed by each tool across the datasets, results showed that overall performance of MAFFT is better among all, it gives accurate and fast result in most of the cases. Overall accuracy for ProbCons over the six datasets is good but its average execution time is high. Clustal Omega is the fastest method but it compromises accuracy as discovered in the analysis.

REFERENCES

1. Needleman, S.B. & Wunsch, C. D., “A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins”,1970, J.Mol. Biol., vol. 48, pp. 443-453.
2. Smith, T. F. & Waterman, M. S., "Identification of common molecular subsequences",1981, J. Mol. Biol., vol. 147, pp 195-197.
3. Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden TL, “NCBI BLAST: a better web interface”, Nucleic Acids Res. 2008;36: W5–9.
4. Marcus Stamm, René Staritzbichler, Kamil Khafizov, and Lucy R. Forrest, “AlignMe—a membrane protein sequence alignment web server”, Nucleic Acids Res. 2014 Jul 1; 42(Web Server issue): W246–W251.
5. Rice P, Longden I, Bleasby A,” EMBOSS: the European molecular biology open software suite”, Trends Genet. 2000; 16:276–7.
6. Altschul SF1, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ, “Gapped BLAST and PSI-BLAST**: a** new generation of protein database search programs”, Nucleic Acids Res. 1997 Sep 1;25(17):3389-402.
7. C. Kemena and C. Notredame, “Upcoming challenges for multiple sequence alignment methods in the high-throughput era,” Bioinformatics 2009, vol.25, no.19, pp.2455–2465.
8. Thompson JD, Higgins DG, Gibson,”. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties, and weight matrix choice”, Nucleic Acids Res.1994;22:4673–80.
9. Ari Loytynoja, Nick Goldman, “webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser”, BMC Bioinformatics 2010, 11:579.
10. Ari Loytynoja1, Albert J. Vilella1 and Nick Goldman, “Accurate extension of multiple sequence alignments using a phylogeny-aware graph algorithm”,2012, Bioinformatics, Vol. 28, 1684–1691.
11. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al.,” Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega”. Mol Syst Biol. 2011; 7:539.
12. Edgar RC,” MUSCLE: multiple sequence alignment with high accuracy and high throughput”, Nucleic Acids Res. 2004;32:1792–7.
13. Edgar RC,” MUSCLE: a multiple sequence alignment method with reduced time and space complexity”, BMC Bioinformatics. 2004; 5:113.
14. Notredame C, Higgins DG, Heringa J,”. T-coffee: a novel method for fast and accurate multiple sequence alignment”, J Mol Biol. 2000;302:205–17.
15. Katoh K, Standley DM,” MAFFT multiple sequence alignment software version 7: improvements in performance and usability”, Mol Biol Evol. 2013; 30:772–80.
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., 2002;30:3059–66.
17. Gotoh O,” Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments”, J Mol Biol 1996, 264: 823–838.
18. Do CB, Mahabhashyam MS, Brudno M, Batzoglou S.,” ProbCons: Probabilistic consistency-based multiple sequence alignment”, Genome Res. 2005;15:330–40.
19. Pei J, Grishin NV,” PROMALS: towards accurate multiple sequence alignments of distantly related proteins”, Bioinformatics. 2007,23:802–8.
20. Pei J, Kim BH, Grishin NV.,” PROMALS3D: a tool for multiple protein sequence and structure alignments”, Nucleic Acids Res. 2008;36:2295–300.
21. Ye Y, Cheung, DW, Wang W, Yiu S m, Zhan Q, Lam T W, Ting HF, “GLProbs: Aligning Multiple Sequences Adaptively”, IEEE/ACM Trans Comput Biol Bioinform. 2015
22. Gonzalez-Dominguez J, Liu Y, Tourino J, Schmidt B, “MSAProbs-MPI: parallel multiple sequence aligner for distributed-memory systems”, Bioinformatics 2016.
23. DeSantis TZ, Hugenholtz P, Keller K, Brodie EL, Larsen N, Piceno YM, et al., “NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA gene”, Nucleic Acids Res. 2006;34: W394–9.
24. Wong KM, Suchard MA, Huelsenbeck JP, “Alignment uncertainty and genomic analysis”, Science. 2008; 319:473–6.
25. Thompson JD, Koehl P, Ripp R, Poch O, “ BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark”, Proteins. 2005 ;61(1):127-36.
26. Stebbings LA, Mizuguchi K, “HOMSTRAD: recent developments of the Homologous Protein Structure Alignment Database”, Nucleic Acids Res. 2004;32(Database issue): D203-7.
27. Yingying Wang, Hongyan Wu and Yunpeng Cai, “A benchmark study of sequence alignment methods for protein clustering”, BMC Bioinformatics 2018, 19(Suppl 19):529.
28. Fabiano Sviatopolk-Mirsky Pais, Patrícia de Cássia Ruy, Guilherme Oliveira, and Roney Santos Coimbra, “Assessing the efficiency of multiple sequence alignment programs”, Algorithms for Molecular Biology 2014, 9:4.
29. Julie D. Thompson, Frédéric Plewniak and Olivier Poch, “A comprehensive comparison of multiple sequence alignment programs”, 1999, Nucleic Acids Research, Vol. 27, No. 13, 2682–2690.
30. Berman H, Henrick K, Nakamura H., “Announcing the worldwide protein data Bank”, Nat Struct Biol. 2003;10:980.
31. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR, et al., “The Pfam protein families database”, Nucleic Acids Res. 2008;36: D281–8.
32. Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJ, Chothia C, et al., “Data growth and its impact on the SCOP database: new developments”, Nucleic Acids Res. 2008;36: D419–25.
33. Iantorno S, Gori K, Goldman N, Gil M, Dessimoz, “Who watches the watchmen? An appraisal of benchmarks for multiple sequence alignment”, In: Russell DJ, ed. Multiple Sequence Alignment Methods. Vol 1069. Clifton, NY: Humana Press; 2014:59–73.
34. Feng DF, Doolittle RF, “Progressive sequence alignment as a prerequisite to correct phylogenetic tree",1987, J Mol Evol. 25 (4): 351–360.

35. Saitu N, Nei M, " The neighbor-joining method: a new method for reconstructing phylogenetic trees", Mol Biol Evol. 1987 Jul;4(4):406-25
36. I.GronauandS.Moran, "Optimal implementations of UPGMA and other common clustering algorithms", Information Processing Letters, vol.104,no.6,pp.205–210,2007.
37. F. Sievers, A. Wilm, D. Dineen et al., "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega", Molecular Systems Biology, vol. 7, article 539, 2011.
38. Morgenstern B, Frech K, Dress A, Werner T.," DIALIGN: finding local similarities by multiple sequence alignment", Bioinformatics. 1998; 14:290–4.
39. Ari Loytynoja, " Phylogeny-aware alignment with PRANK", **m**ultiple Sequence Alignment Methods pp 155-170.
40. C Notredame, D G Higgins, "SAGA: sequence alignment by genetic algorithm" Nucleic Acid Research 1996.
41. C. Notredame, E A O Brian and D G Higgins "RAGA: RNA sequence alignment by genetic algorithm", Nucleic Acids Res. 1997 Nov 15; 25(22): 4570–4580.
42. Narayan Behra at el "Higher accuracy protein multiple sequence alignments by genetic algorithm", ICCS 2017.
43. Jin Kim, Sakti Pramanik, Moon Jung Chung, "Multiple sequence alignment using simulated annealing", Bioinformatics, Volume 10, Issue 4, July 1994.
44. Dayhoff, M. O.; Schwartz, R. M.; Orcutt, B. C., "A model of evolutionary change in proteins". Atlas of Protein Sequence andStructure,1978, Vol. 5, Suppl. 3 National Biomedical Research Foundation, Washington D.C. U.S.A, pp 345-352.
45. Henikoff, S; Henikoff, JG, "Amino acid substitution matrices from protein blocks". PNAS, 1992, vol. 89, pp 10915-10919

AUTHORS PROFILE

Arunima Mishra is pursuing her PhD degree in machine learning area from Dr APJ Abdul Kalam Technical University, Lucknow, India and completed her M Tech degree from Bharath University, Chennai, India. Her research interests include machine learning, bioinformatics, Data mining, and programming methodologies. She has published many research papers in these areas.

.

Dr Bipin Kumar Tripathi completed his PhD degree in computational intelligence from IIT Kanpur, India, and MTech degree in computer science and engineering from IIT Delhi, India. Dr Tripathi is currently serving as a director of Rajkiya Engineering College, Bijnor, U.P., India. His research interests include high-dimensional neurocomputing, computational neuroscience, intelligent system design, machine learning and computer vision focused on biometrics, and 3D imaging. He has published several research papers in these areas in many peer-reviewed journals including IEEE Transaction, Elsevier, Springer, and other international conferences.

Dr Sudhir Singh Soam has received his PhD degree in computational immunology and bioinformatics from Dr APJ Abdul Kalam Technical University, Lucknow, India. He is currently working in Institute of Engineering and Technology, Lucknow. His research interest includes artificial immune system algorithms, machine learning in bioinformatics, data mining. He has published several research papers in these areas in many peer-reviewed journals.