J. Appl. Environ. Biol. Sci., 4(8S)76-80, 2014

© 2014, TextRoad Publication

ISSN: 2090-4274
Journal of Applied Environmental
and Biological Sciences
www.textroad.com

Multiple Sequence Alignment Tools: Assessing Performance of the Underlying Algorithms

Masroor Ellahi Babar[1], Muhammad Tariq Pervez[2], Asif Nadeem[3], Tanveer Hussain[4] and Naeem Aslam[5]

[1]Department of Bioinformatics, Virtual University of Pakistan
[2]Department of Computer Science, Virtual University of Pakistan
[3]Institute of Biochemistry and Biotechnology, University of Veterinary and Animal Sciences, Lahore, Pakistan
[4]Department of Live Stock Production, University of Veterinary and Animal Sciences, Lahore, Pakistan
[5]Department of Computer Science, NFC Institute of Engineering & Technological Training, Multan, Pakistan

Received: September 1, 2014
Accepted: November 13, 2014

ABSTRACT

Multiple sequence alignments have primary role in several domains of modern molecular biology such as protein 3D structure/function prediction, phylogeny inference, molecular function, intermolecular interactions and many other common tasks in sequence analysis. Presently, many tools to construct multiple sequence alignments are available but none of them is accurate for all types of data sets. Several comparative studies have been conducted to report quality and efficiency of MSA tools but their focus was on the individual popular MSA tools. This study presents a comparative study of various groups of MSA tools. MSA tools were placed in four groups. First group had progressive consistency approach based MSA tools. Second group comprised progressive matrix approach based MSA tools. Third group consisted of Hidden Marko Model based MSA tool and fourth group had iterative divide and conquer approach based MSA tool. Results showed that SATe, which is an iterative divide and conquer approach based tool, outperformed all other MSA tools. However in the group of progressive consistency technique based MSA tools, ProbCons and MAFFT-L-INS-I were on the first and second positions. Among the progressive matrix based tools, Muscle is on the top.

KEYWORDS: Comparison, MSA SPS, CS.

INTRODUCTION

Multiple Sequence Alignments (MSAs) have primary role in several domains of modern molecular and bioinformatics such as protein 3D structure/function prediction, phylogeny inference [1-3] molecular function, intermolecular interactions [4-6] and many other common tasks in sequence analysis. A lot of innovative algorithms and techniques have been proposed to get better quality of MSAs, but, still none of the MSA tools can reconstruct accurate alignments for all types of data sets [7]. Several warehouses of MSAs such as SABmark [8], PREFAB [9] and BAliBASE [10] are available. These repositories of MSAs comprise highly quality manually refined alignments which may be used to measure performance of various MSA methods.

BAliBASE was first database of benchmark alignments specifically developed to investigate accuracy of MSA tool. It comprises various datasets which simulate real challenges faced during reconstruction process of multiple sequence alignments. BAliBASE is divided into various reference datasets. Reference 1 comprises equidistant sequences. This data set was subdivided based on the percent identity. Reference 2 comprises protein families with orphan sequences. Reference 3 is a set of divergent subfamilies. These subfamilies have groups with less then 20% identity. Reference 4 is a set of sequences with large N/C terminals. References 5 comprise sequences which have large internal insertions and deletions. Reference data set 6 comprises sequences with repeats. This was further subdivided into sequences with diverse residue similarity, input order and having additional domains. Reference 7 is a set of sequences with transmembrane regions. This was ordered into subgroups having highly conserved core blocks. Reference 8 comprises alignments with inverted domains. The latest addition to BAliBASE is Reference data set 9 [11] which is an assembly of protein families having linear motifs [12].

Several studies [12-14reported accuracy and efficiency of MSA tools but their focus was individual MSA tools. Secondly vary less effort was made to study the underlying algorithms [12].

This research work presents a comparative study of groups of MSA tools. We placed MSA tools into various groups based on the underlying algorithms. Four groups of MSA tools were developed. First group comprised five MSA tools such as Dialign-TX [15], T-Coffee [16], ProbCons [17], MAFFT-L-INS-I [18] and MAFFT-FTNS2 [18]. These MSA tools used progressive-consistency algorithm. Second group consisted of four MSA tools such as MultAlign [19], ClustalW [20], KAlign [21] and Muscle [22]. This group of MSA tools implemented progressive-matrix algorithm. Third group which consisted of Clustal Omega [23] implemented Hidden Markov Model approach. Fourth group which consisted of SATe [24] implemented iterative divide and conquer approach.

* **Corresponding Author:** Muhammad Tariq Pervez, Department of Computer Science, Virtual University of Pakistan. tariq_cp@hotmail.com

MATERIALS AND METHODS

Benchmark Dataset

Six reference test cases (RV11, RV12, RV20, RV30, RV40 and RV50) available in the version 3 of the BALiBASE (ftp://ftp-igbmc.u-strasbg.fr/pub/BAliBASE3).

Alignment Accuracy Assessment Procedure

Alignment accuracy was measured by comparing an alignment generated by an MSA tool with a benchmark alignment and calculating SPS and CS. The sum of pair score is computed by adding up the correctly aligned sequences. It determines the capability of MSA methods to align some, if not all, of the sequences in an MSA. Column score measures the capacity of MSA tools to align all of the sequences correctly.

MSA Tools

MSA tools were selected based on their underlying algorithms. Study of groups of MSA tools was conducted. The selected MSA tools, their versions, the underlying algorithms and the links to download them is provided in Table 1.

Computing Machine

All MSA programs were run on a computing machine having Core i7 3.34 GHz processor, 8 GB RAM and Fedora OS.

RESULTS

Sequences and Benchmark Alignments

Six reference test cases were downloaded from the home page of BAliBASE. These test cases comprised of benchmark alignments and the corresponding sequences. The sequences were aligned by each MSA tool and compared with the benchmark alignments. The quality of alignments, constructed by the MSA methods, was measured by the two most popular scores i.e. sum of pairs (SPS) and column score (CS).

MSA Algorithm Quality Evaluation

Overall alignment accuracy of the MSA tools was measured using SPS and CS (Figure 1). Results showed that SATe, which implemented the iterative divide and conquer (IDC) approach, outperformed all the MSA tools of all the groups. In the group of MSA tools which have implemented progressive consistency approach, ProbCons generated the most accurate alignments. ProbCons is also on the second position among the all groups of MSA tools. In the group of MSA tools, which have implemented progressive matrix approach, Muscle is on the top while performance of ClustalW is the poorest. Clustal Omega which have implemented HMM performs better then many MSA tools of the group of consistency (Dialign-TX, MAFFT-FTNS2) and matrix (ClustalW and MultAlign).

MSA Algorithm Efficiency Evaluation

Results showed that, in the group of PC approach based MSA tools, ProbCons and MAFFT-FTNS2 are the slowest and fastest tools respectively. T-Coffee is the second slowest tool. In the group of PM approach based MSA tools, MultAlign and KAlign are the slowest and the fastest tools. However, there is a little bit between efficiency difference between KAlign and Muscle. SATe is 109% and 90% more efficient than ProbCons and T-Coffee respectively. Figure 2 shows time spent in seconds by each MSA method.

Table 1. Summary of MSA tools used in this study

Major algorithm	MSA tools	Version	Download link
Progressive-Consistency	Dialign-TX	1.0.2	**http://dialign-tx.gobics.de/download**
	T-Coffee	10.00.r1613	**http://www.tcoffee.org/**
	ProbCons	1.12	**http://probcons.stanford.edu/download.html**
	MAFFT-L-INS-I	7.0	**http://mafft.cbrc.jp/alignment/software/**
	MAFFT-FTNS2	7.0	**http://mafft.cbrc.jp/alignment/software/**
Progressive-Matrix	MultAlign		**http://multalin.toulouse.inra.fr/multalin**
	ClustalW	2.0.10	**http://www.clustal.org/download/current/**
	KAlign	2.0	**http://msa.sbc.su.se/cgi-bin/msa.cgi**
	Muscle	3.8.31	**http://www.drive5.com/MUSCLE/downloads.htm**
Hidden Markov Model	Clustal Omega	1.2.0	**http://www.clustal.org/omega/**
Iterative divide and conquer	SATe	2.2.7	**http://phylo.bio.ku.edu/software/sate/sate.html**

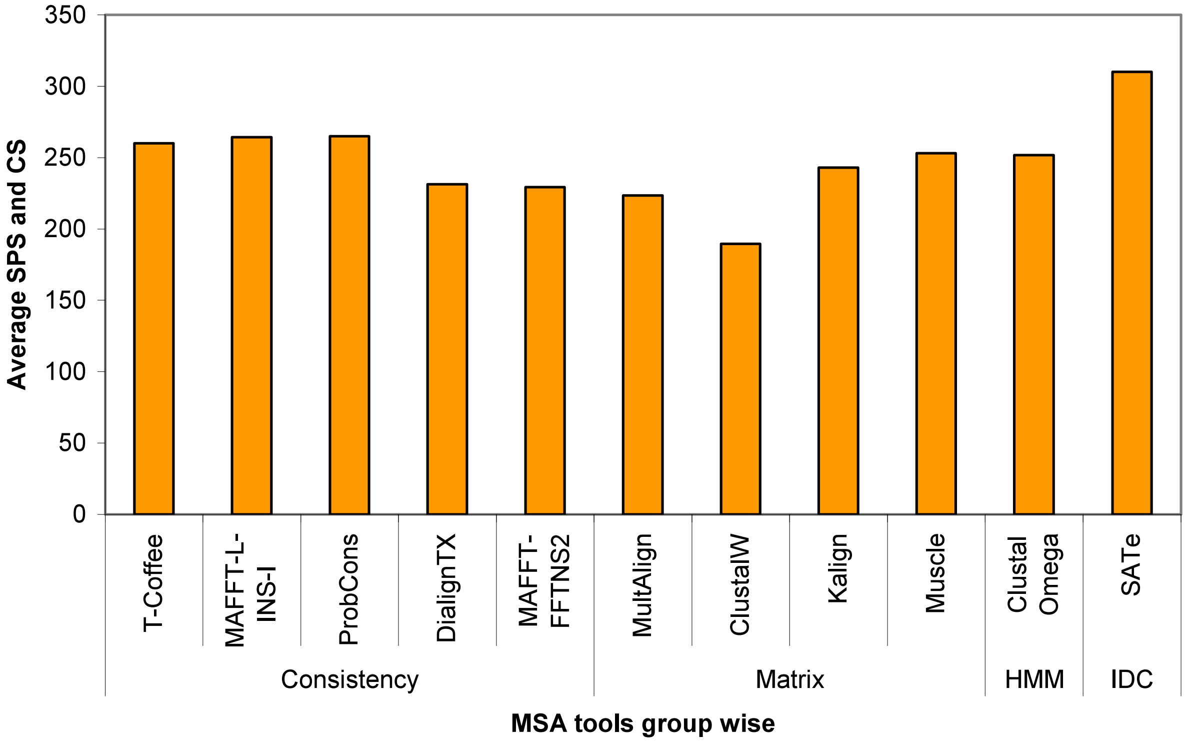

Figure 1. MSA algorithm quality evaluation. IDC approach implemented by SATe outperformed all the MSA tools of all the groups. ProbCons, which belongs to the group of PC, was on the second position among all the tools of all groups. Performance of Clustal Omega was better than many MSA tools of other groups. ClustalW, which belongs to the PM group, generated least quality alignments.

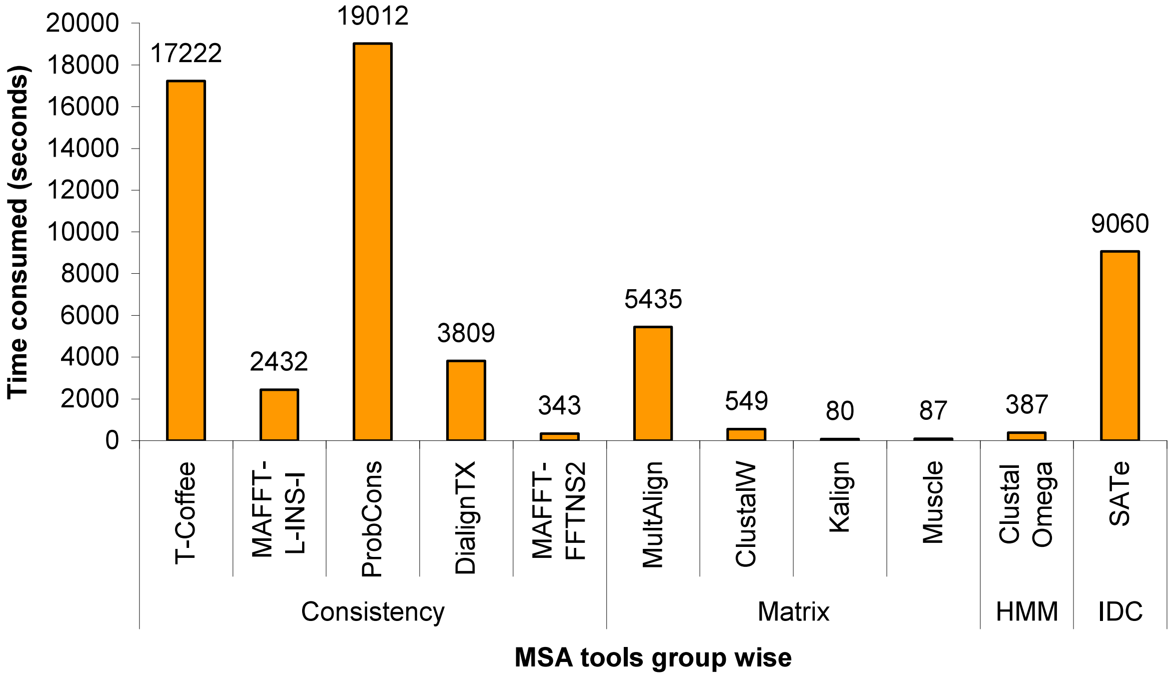

Figure 2. Efficiency comparison among the groups of MSA tools. In the group of PC approach based MSA tools ProbCons the slowest tool. In the group of PM approach based MSA tools, MultAlign is the slowest tool. SATe is many times faster than T-Coffee and ProbCons, its quality competitors

Conclusion

Several studies for comparison of MSA tools are available. All of them report that none of the MSA tool is accurate for all types of data sets. Most of the studies selected MSA tools based on their popularity. Furthermore, they studied individual MSA tools. This study presents comparison of groups of MSA tools based on the underlying algorithms.

Results showed that overall, MSA tools developed using consistency approach are more accurate and MSA tools developed using matrix approach are faster. However SATe which have used iterative divide and conquer approach is the fastest tool and it is very efficient than many MSA tools of the group of consistency based approach.

REFERENCES

1. Kim, J. and Ma J 2011. PSAR: measuring multiple sequence alignment reliability by probabilistic sampling. Nucl. Acids Res. 39 (15): 6359-6368.

2. Siepel, A. Bejerano, G. Pedersen, J.S. Hinrichs, A.S. Hou, M. Rosenbloom, K. Clawson, H. Spieth, J. Hillier, L.W. Richards, S. Weinstock, G.M. Wilson, R.K. Gibbs, R.A. Kent, W.J. Miller, W. and Haussler, D. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Research. 15(8):1034-1050.

3. Roskin, K.M. Diekhans, M. and Haussler, D. 2003. Scoring Two-Species Local Alignments to Try to Statistically Separate Neutrally Evolving from Selected DNA Segments. Proceedings of the seventh annual international conference on Computational molecular biology ACM Press. 257-266.

4. Levasseur, A. Pontarotti, P. Poch, O. and Thompson, J.D. 2008. Strategies for reliable exploitation of evolutionary concepts in high throughput biology. Evol Bioinform Online 4: 121–137.

5. Wong, K.M. Suchard, M.A. and Huelsenbeck, J.P. 2008. Alignment uncertainty and genomic analysis. Science 319: 473–476.

6. Lo¨ytynoja, A. and Goldman, N. 2008. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science 320:1632–1635.

7. Thompson, J.D. Linard, B. Lecompte, O. and Poch, O. 2011. A Comprehensive Benchmark Study of Multiple Sequence Alignment Methods: Current Challenges and Future Perspectives. PLoS ONE 6(3): e18093.

8. Walle, I.V. Lasters, I. Wyns, L. 2005. SABmark-a benchmark for sequence alignment that covers the entire known fold space. Bioinformatics 21(7):1267-8.

9. Edgar, R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Research 32(5), 1792-97.

10. Thompson, J. Koehl. P. Ripp, R. and Poch, O. 2005. BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. Proteins 61:127-36.

11. Perrodou, E. Chica, C. Poch, O. Gibson, T.J. and Thompson, J.D. 2008. A new protein linear motif benchmark for multiple sequence alignment software. BMC Bioinforma, 9:213.

12. Pais, F.S.M. Patrícia, C.R. Guilherme, O. and Roney, S.C. 2014. Assessing the efficiency of multiple sequence alignment programs. Algorithms for Molecular Biology, 9: 4.

13. Thompson, J.D. Linard, B. Lecompte, O. and Poch, O. 2011. A Comprehensive Benchmark Study of Multiple Sequence Alignment Methods: Current Challenges and Future Perspectives. PLoS ONE 6(3): e18093.

14. Nuin, P.A. Wang, Z. and Tillier, E.R. 2006. The accuracy of several multiple sequence alignment programs for proteins. BMC Bioinformatics;7:471.

15. Subramanian, A.R. Kaufmann, M. nad Morgenstern, B. 2008. DIALIGN-TX: greedy and progressive approaches for segment-based multiple sequence alignment algorithms for Molecular Biology 3:6.

16. Notredame, C. Higginsm D.G. and Heringam J. 2000 T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol 302(1):205-217

17. Do, C.B. Mahabhashyam, M.S. Brudno, M. and Batzoglou, S. 2005. ProbCons: Probabilistic consistency-based multiple sequence alignment. Genome Res. 15:330–340.

18. Katoh, K. and Standley, D. M. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular biology and evolution, 30(4), 772-780.

19. CORPET, F. 1988. Multiple sequence alignment with hierarchical clustering" Nucl. Acids Res 16 (22),:10881-10890

20. Thompson, J.D. and Higgins, D.G. and Gibson, T.J. 1994. 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(22):4673–4680.

21. Lassmann, T. Frings. O.S. and Sonnhammer, E.L. 2009. Kalign2: high-performance multiple alignment of protein and nucleotide sequences allowing external features. Nucleic Acids Res 37: 858–865.

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

23. Sievers, F. Wilm, A. Dineen, D.G. Gibson, T.J. Karplus, K. Li, W. Lopez, R. McWilliam, H. Remmert, M. Söding, J. Thompson, J.D. and Higgins, D.G. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology 7:539.

24. Liu, K. Warnow, T. Holder, M.T. Nelesen, S. Yu, J. Stamatakis, A. and Linder, C.R. 2012. SATé-II: Very Fast and Accurate Simultaneous Estimation of Multiple Sequence Alignments and Phylogenetic Trees. Systematic Biology. 61(1):90-106