Multiple Sequence Alignment and Phylogeny Reconstruction

Vishal Rana

October 1, 2020

Introduction

Mutiple sequence alignment is often a crucial first step in phylogeny reconstruction from a set of unaligned sequences. In this report we comapre the performance of two multiple sequence alignment software, MAFFT [2] and MUSCLE [1], on some synthetic as well as biological datasets by comparing their multiple sequence alignment to reference alignments available using sum-of-pairs false positives (SPFP) and sum-of-pair false negatives (SPFN) as metrics. We also compute phylogeny trees from these estimated alignments using FastTree and compare them to trees computed from reference alignments, using fasle positive (FP) and false negative (FN) rates as metrics.

Experiment design and expectations

Datasets:

The analysis was performed on datasets 1000M1 and 1000M4 from Datasets link. Both contain 20 replicates with 1000 taxa each. They differ in their rates of substitutions and hence in the depths of the simulated trees. We expect the methods to perform worse on 1000M1 dataset because of higher rates of substitutions. Both datasets have medium length gaps in the reference alignments as described in [3]. Additionally we also did the analysis for biological replicate 16S.M [3].

Multiple sequence alignment:

We used two different tools for sequence alignment. FastSP was used to calculate SPFN and SPFP. FastTree was used for tree estimation.

MUSCLE:

MUSCLE [1] starts off by finding a similarity score based on k-mers counting and using this similarity to construct a tree intially. Next, it uses this tree to further build a better alignment progressively. Then it obtains the final alignment by optimizing for sum-of-pair scores for all the pairs of sequences to be aligned. We chose this method because of prior experience in working with it.

MAFFT:

MAFFT [2] also starts by finding a similarity matrix and building a guide tree based on it. It then iterates between alignment and reconstructing guide treee. Then it performs an optimzation over a weighted sum of pair score and a consistency score. Consistency score measures consistency between multiple and pair-wise alignemnts. Since this takes into account consistency score in addition to the sum of pair score, it is expected to perform better than MUSCLE. We chose this software because the documentation describing the various modes to run it in and the underlying algorithms was easily accessible.

FastSP:

FastSP [4] calculates the SPFN and SPFP. SPFN is the fraction of homologus pairs in the reference alignment that are not present in the estimated multiple sequence alignment. SPFP is the fraction of homologus pairs in the estimated alignment that are not present in the reference multiple sequence alignement. Lower values of SPFN and SPFP indicate better performance of the alignment tool. We expect the scores to be lower for MAFFT than MUSCLE as MAFFT is expected to perform better and lower on dataset 1000M4 than 1000M1 as 1000M1 has a higher rate of substitutions and hence a higher diameter of the tree which makes estimation harder.

FastTree:

FastTree [6], [5] is a maximum likelihood based tree estimation method that uses nearest neighbour interchanges to iterate and improve the tree. This method was chosen beacuse it gives good performance and takes much less compute time compared to other options analyzed previously.

Comparing Trees:

We use false positive and false negative rates to evaluate the tree estimation. False negatives are the edges in the reference tree that are not present in the estimated tree while false positives are edges in the estimated tree that are not present in the reference tree. Normalizing the number of FPs and FNs by n-3 gives the respective rates where n is the number of edges in the tree. Robinson-Foulds number is another metic that can be used and is given as $\frac{FP+FN}{2n-6}$.

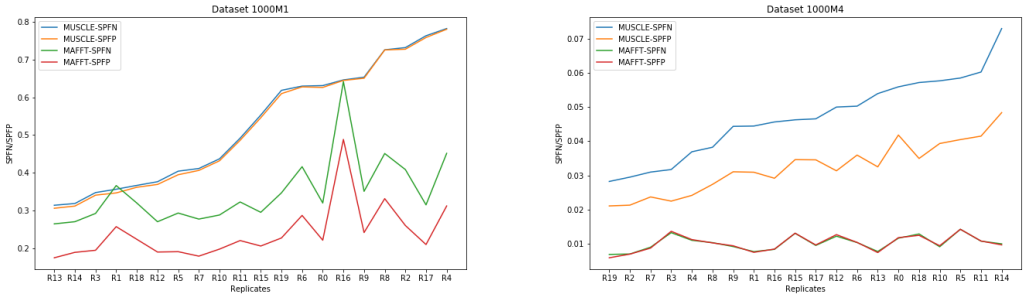
Results

We begin by reporting the mean values of various metric we used for the analysis 1, 2. The variance for various categories are indicated in parentheses.

As expected both the methods give better results on 1000M4 dataset than on 1000M1, this is because of higher rate of evolution in 1000M1. MAFFT performs better than

Table 1: Mean and variance for SPFN and SPFP calculated on the datasets 1000M1 and 1000M4 using MUSCLE and MAFFT

	1000M1		1000M4	
Method	SPFN	SPFP	SPFN	SPFP
MUSCLE	0.53 (2.4e-2)	0.52 (2.6e-2)	0.047 (1.3e-4)	0.032 (5.5e-5)
MAFFT	0.35 (7.8e-3)	0.24 (5.0e-3)	0.010 (4.4e-6)	0.010 (5.1e-6)



(a) 1000M1 (b) 1000M4

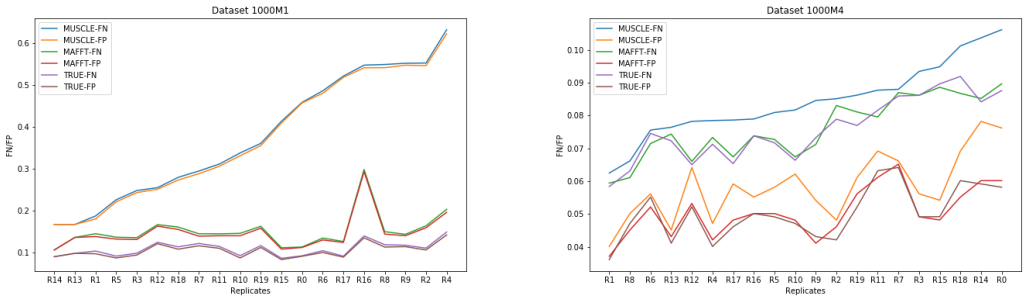
Figure 1: SPFN and SPFP for 1000M1 and 1000M4 datasets for all 20 replicates

MUSLCE. The only replicates where their preformance comes close to each other are R1 and R16 in 1000M1. It is interesting to note that SPFN and SPFP for MUSCLE on 1000M1 are very close for all replicates while the same is true for MAFFT on 1000M4 instead.

For the tree reconstruction part, multiple sequence alignment returned by MAFFT gives better results than MUSCLE. This is expected since a better alignment should translate to a better phylogeny reconstruction. The performace of MAFFT is close to tree constructed on the true reference alignment. On dataset 1000M1, the FP and FN rates are very close to each other for all methods of tree reconstruction which indicates that the estimated tree is almost fully resolved (since the reference tree is fully resolved). The same is not true for 1000M4 however. Surprisingly, tree using MAFFT gives results

Table 2: Mean and variance for FN and FP for datasets 1000M1 and 1000M4 using FastTree

	1000M1			1000M4		
Method	FN	FP	RF	FN	FP	RF
MUSCLE	0.38 (2e-2)	0.37 (2e-2)	0.37 (2e-2)	0.084 (1e-4)	0.059 (1e-4)	0.071 (1e-4)
MAFFT	0.15 (2e-3)	0.15 (2e-3)	0.15 (2e-3)	0.076 (8e-5)	0.051 (5e-5)	0.063 (5e-5)
Reference Alignment	0.11 (3e-4)	0.10 (2e-4)	0.11 (3e-4)	0.076 (9e-5)	0.050 (6e-5)	0.063 (6e-5)



(a) 1000M1 (b) 1000M4

Figure 2: FN and FP for 1000M1 and 1000M4 datasets for all 20 replicates

better than even the tree on reference alignment for some replicates in 1000M4. This can only be explained as an outcome of the tree reconstruction method used (FastTree in our case). Another observation that was made was that the total number of correctly aligned columns was very low for all the experiments.

On the biological dataset 16S.M MAFFT has SPFN 0.18 and SPFP 0.20 while MUSCLE gives SPFN 0.30 and SPFP 0.28. MAFFT performs better here as well. For tree reconstruction, since the reference tree is not fully resolved (many branches collapsed due to insufficent bootstrap support), it leads to a large number of false positives. It is therefore more instructive to look at the false negative rate instead. FN rate for MUSCLE is 0.09, for MAFFT is 0.03 and for Reference alignment is 0.005. This follows the same trend where MAFFT performs better than MUSCLE.

MAFFT takes much longer to run compared to MUSCLE. For this reason the number of iterations for MAFFT was limited to 5. It was noticed that this does not lead to any significant loss in performance as most drastic improvement happens in the first few iterations itself.

Supplementary material

Version numbers and command used for various software are given below.

MUSCLE:

Version 3.8.31. Command: muscle3.8.31_i86linux64 -in infile -out outfile

MAFFT:

Version 7.471. Command: mafft –localpair –maxiterate 5 –thread 8 infile > outfile

FastSP:

Version Github. Command: java -jar FastSP.jar -r ref.aln -e estimated.aln

FastTree:

Version 2.1.11. Command: FastTree -gtr -nt inputfile > outputfile

TreeCompare:

treecompare.py (Courtesy Erin Molly)

References

[1] Robert C Edgar. Muscle: a multiple sequence alignment method with reduced time and space complexity. *BMC bioinformatics*, 5(1):113, 2004.

[2] Kazutaka Katoh and Daron M Standley. Mafft multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, 30(4):772–780, 2013.

[3] Kevin Liu, Sindhu Raghavan, Serita Nelesen, C Randal Linder, and Tandy Warnow. Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. *Science*, 324(5934):1561–1564, 2009.

[4] Siavash Mirarab and Tandy Warnow. Fastsp: linear time calculation of alignment accuracy. *Bioinformatics*, 27(23):3250–3258, 2011.

[5] Morgan N Price, Paramvir S Dehal, and Adam P Arkin. Fasttree: computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular biology and evolution*, 26(7):1641–1650, 2009.

[6] Morgan N Price, Paramvir S Dehal, and Adam P Arkin. Fasttree 2–approximately maximum-likelihood trees for large alignments. *PloS one*, 5(3):e9490, 2010.