BIOINFORMATICS

Vol. 00 no. 00
Pages 1–2

MAMMLE: phylogeny estimation based on multiobjective application-aware MUSCLE and maximum likelihood ensemble

Muhammad Ali Nayeem [1], Naser Anjum Samudro [1], M. Saifur Rahman [1] and M. Sohel Rahman [1,*]

[1]Department of CSE, BUET.

ABSTRACT

Motivation: Phylogenetic trees are often inferred from a multiple sequence alignment (MSA) where the tree accuracy is heavily impacted by the nature of estimated alignment. MUSCLE is a general-purpose MSA tool widely used for its high throughput and accuracy. Carefully equipping MUSCLE with multiple application-aware objectives positively impacts its capability to yield better trees.

Results: We introduce MAMMLE, a framework for inferring better phylogenetic trees from unaligned sequences by hybridizing MUSCLE with multiobjective optimization strategy and leveraging multiple Maximum Likelihood hypotheses. MAMMLE may offer a significant improvement (upto 27% in our experiments) in tree accuracy over MUSCLE.

Availability and implementation: MAMMLE is an Open Source tool available at `https://github.com/ali-nayeem/mammle`

Contact: ali_nayeem@cse.buet.ac.bd

1 INTRODUCTION

Maximum likelihood (ML) is a statistical method for inferring high quality phylogenetic trees. ML trees are estimated on multiple sequence alignments (MSA). The characteristics of the estimated MSA dramatically influences the tree accuracy. Besides phylogeny estimation, MSA has other important biological applications, such as, prediction of structure/function of new proteins, identification of conserved regions, etc. Thus an MSA tool that is aware of its intended usage (i.e., phylogeny estimation in our case) is expected to yield output of higher quality as opposed to general-purpose MSA tools (Nayeem *et al.* (2020)).

MUSCLE (Edgar (2004)) is one of the most widely-used MSA methods cited by around ten new papers every day (Edgar (2015)). It performs progressive alignment and then iteratively refines the estimated MSA based on the popular SP (sum-of-pairs) score as the objective function. Nayeem *et al.* (2020) developed a systematic method to identify application-aware MSA objective functions based on their correlation to the tree accuracy. It was subsequently shown, through extensive experiments, that optimizing those objectives by multiobjective (MO) techniques can yield high-quality ML trees. An MO approach treats all objectives (usually conflicting) equally and generates a set of non-dominated Pareto-optimal solutions that are equivalent in the context of conflicting objectives.

Here, we present MAMMLE, a framework through which we infuse the concept of MO application-awareness into MUSCLE by incorporating four application-aware objectives from Nayeem *et al.* (2020) within the iterative refinement phase thereof through an MO strategy. MAMMLE generates multiple alternative alignments and for each of them an ML tree is inferred. We take these multiple hypotheses into our advantage and develop an ensemble approach for producing a better phylogenetic tree. We present our overall approach for phylogeny estimation from unaligned sequences as a flexible framework whose components can potentially be modified, replaced or further refined by bioinformatics researchers and practitioners.

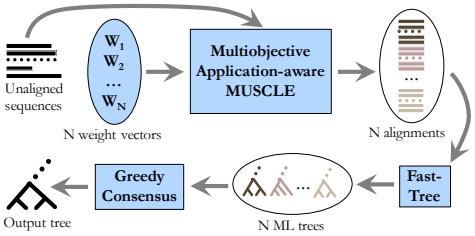


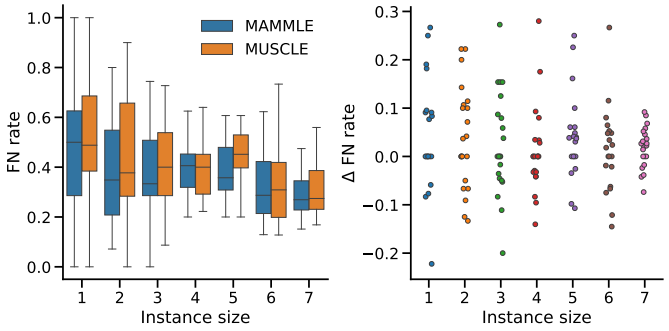
Fig. 1: Simplified workflow of MAMMLE framework.

2 METHODS

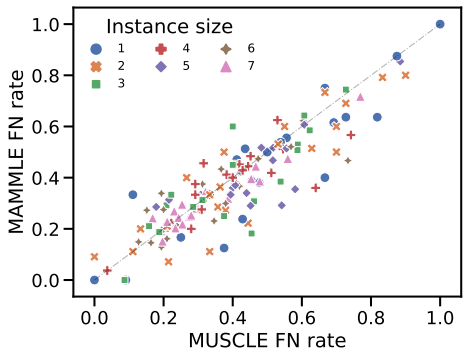
The phylogenetic reconstruction pipeline of MAMMLE is illustrate in Figure 1. The components open to modification are marked with a blue shade. It starts by feeding the unaligned sequences to the system, which simultaneously optimizes four application-aware objectives with the help of N well-distributed 4D weight vectors given as input and generate N (in this article $N = 30$) alternative alignments (please refer to Section S1 of the supplementary file for details). Next, MAMMLE infers a ML tree from each of the N alignments using FastTree (Price *et al.* (2010)). We prefer FastTree over other ML methods due to its speed. Finally, MAMMLE summarizes the N ML trees using a simple greedy consensus method of PAUP* (Phylogenetic Analysis Using PAUP) available at `https://paup.phylosolutions.com`. Any tree summarizing approach can be employed here. Note that

*to whom correspondence should be addressed

MAMMLE relieves the user from parameter tuning through its well-spaced weight vectors and thus we used default parameter settings of MUSCLE and FastTree.



(a) FN rate distribution (b) MAMMLE: FN rate improvement



(c) Individual FN rate: MUSCLE vs MAMMLE

Fig. 2: Close examination of MAMMLE vs MUSCLE.

3 RESULTS AND DISCUSSION

We compared MAMMLE with MUSCLE (with FastTree) on the most widely-used BAliBASE 3.0 benchmark (Thompson *et al.* (2005)) of protein families using the FN rate (percentage of true tree edges missing in the estimated tree) as the accuracy measure (see supplementary file for details). Following Mirarab *et al.* (2015), we avoid the instances with a lower number (i.e., below 11) of sequences and generate a reference tree for 147 considered instances by RAxML (Stamatakis (2014)) bootstrap analysis on the reference alignment.

Among the 147 instances, MAMMLE performs better in 74 cases and worse in 43 cases than MUSCLE (supplementary Table S2). We conduct the *one-sided* (i.e., null hypothesis: median FN rate of MUSCLE is better than MAMMLE) Wilcoxon signed ranks test at 95% confidence level where the null hypothesis is rejected with p-value=0.001. To get more insight, we divide the 147 instances into 7 levels of instance size, considering their number of sequences and average sequence length (supplementary Table S1), each level having 20-22 instances. Figure 2a depicts the FN rate distribution of MAMMLE and MUSCLE across different instance sizes. We see that MAMMLE helps to get better accuracy in all levels except 4 and 6. Notably, variance of both MAMMLE and MUSCLE reduces as the instance size increases; this can be attributed to the fact that ML approach performs better as the data size increases.

To complement Figure 2a, we visualize the improvement achieved by MAMMLE by plotting the individual FN rates and MUSCLE FN rate − MAMMLE FN rate for the instances of each level in Figure 2c (scatterplot) and Figure 2b (*stripplot*; adjusts the position of points having similar FN rate for the ease of perception) respectively. Points below (above) the diagonal line (horizontal line at ΔFN rate=0) in Figure 2c (Figure 2b) represent the instances where MAMMLE outperforms MUSCLE. Evidently, the maximum improvement registered by MAMMLE is around 27% across different levels (Figure 2b). Figure 2c offers further insights. MUSCLE performed worse mostly for the smaller instances (Levels 1-4), which are presumed to be difficult for the ML approach; and here MAMMLE outperforms MUSCLE. On the other hand, MUSCLE outperforms MAMMLE mostly for the larger instances where ML approach performs well anyway. Also, observe that MAMMLE outperforms MUSCLE mostly in the cases when MUSCLE FN rate > 0.6.

In summary, although, MAMMLE's improved cases are distributed across all levels and a wider range of FN rates, the improvement seems higher on the smaller instances, indicating that MAMMLE possesses the potential to yield better trees with limited data (e.g., to estimate gene tree on fewer/smaller gene sequences which is challenging for ML approach). However, the cases (45 out of 147) where MUSCLE is better (within 15%) demands further research effort to enhance the current ensemble method (i.e., greedy consensus).

ACKNOWLEDGMENT

The first author is supported by the ICT Doctoral Fellowship administered by ICT Division, Bangladesh.

REFERENCES

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, **32**(5), 1792–1797.

Edgar, R. C. (2015). MUSCLE website. `https://www.drive5.com/muscle`. Accessed: 2021-10-12.

Mirarab, S., Nguyen, N., Guo, S., Wang, L.-S., Kim, J., and Warnow, T. (2015). PASTA: ultra-large multiple sequence alignment for nucleotide and amino-acid sequences. *Journal of Computational Biology*, **22**(5), 377–386.

Nayeem, M. A., Bayzid, M. S., Rahman, A. H., Shahriyar, R., and Rahman, M. S. (2020). Multiobjective formulation of multiple sequence alignment for phylogeny inference. *IEEE Transactions on Cybernetics*.

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). Fasttree 2–approximately maximum-likelihood trees for large alignments. *PloS one*, **5**(3), e9490.

Stamatakis, A. (2014). Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**(9), 1312–1313.

Thompson, J. D., Koehl, P., Ripp, R., and Poch, O. (2005). Balibase 3.0: latest developments of the multiple sequence alignment benchmark. *Proteins: Structure, Function, and Bioinformatics*, **61**(1), 127–136.