COMBINING MULTIPLE TREE REARRANGEMENT OPERATORS FOR EFFICIENT PARSIMONY INFERENCE USING REINFORCEMENT LEARNING APPROACH

Students: Dung Tien Huynh, Tuan Quoc Vu, Dung Viet Nguyen,
Science Instructor: Dr. Diep Thi Hoang

UNIVERSITY OF ENGINEERING AND TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

21020007@vnu.edu.vn, 21020033@vnu.edu.vn, 21020043@vnu.edu.vn, diepht@vnu.edu.vn

INTRODUCTION

- In bioinformatics, the phylogenetic tree is a fundamental concept that traces the evolutionary history of a group of species back to a common ancestor. This study specifically addresses the challenge of parsimony phylogenetic bootstrapping, which entails reconstructing the phylogenetic tree through analysis of biological sequences and gauging the reliability of the partitions within the tree.
- As sequencing technology advances, researchers can generate vast amounts of data for phylogenetic analyses, making it essential to investigate more efficient techniques for phylogenetic bootstrapping.

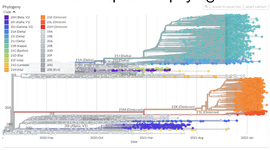

Figure 1 – Phylogenetic tree of Sars-CoV-2 constructed from millions of samples.

- We aim to improve the efficiency of MPBoot (HOANG et al., 2018), a widely used method for parsimony bootstrapping, by integrating tree bisection and reconnection (TBR) transformation. Our MPBoot-TBR algorithm utilizes optimization techniques to search for TBR transformations and hill climbs using TBR, improving sampling efficiency. Additionally, we introduce the MPBoot-ACO algorithm, which uses ant colony optimization to combine hill climbing of three transformations, including TBR, to enhance program performance.

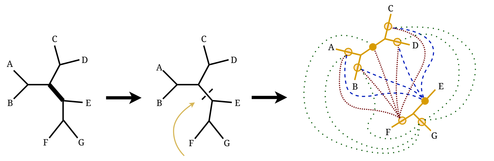

Figure 2 – Tree bisection and reconnection (TBR) tree transformation.

ABOUT MPBOOT

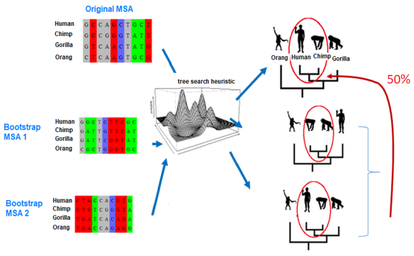

Figure 3 – Summarizing bootstrap results.

- MPBoot speeds up over the standard bootstrap (SBS) (FELSENSTEIN, 1985) with its approximation method. Each bipartition's reliability is calculated by its frequency on the bootstrap tree set - the best trees on pseudoreplicates of the original multiple sequence alignment. Unlike SBS, MPBoot requires only one tree search on the original MSA since it applies resampling parsimony score (REPS) to reuse the original MP column scores for bootstrap MP scores. MPBoot approximates the bootstrap tree set once the original tree search completes.
- MPBoot currently relies solely on SPR for tree rearrangement. Incorporating TBR, a more advanced operation, and combining it with NNI and SPR would potentially improve the MP tree and bootstrap tree set.

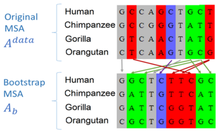

Figure 4 – REPS: Resampling Estimated Parsimony Scores .

PROPOSED METHOD 1: MPBoot-TBR

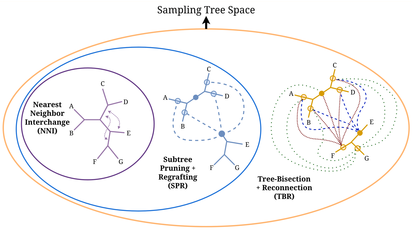

Figure 5 – Sampling tree space of NNI, SPR, and TBR transformation illustrated.

Due to the enhanced tree search space, TBR is time-consuming. To address this, we suggest a method for faster TBR move evaluation that avoids the unnecessary recalculation of MP scores for unchanged subtrees.

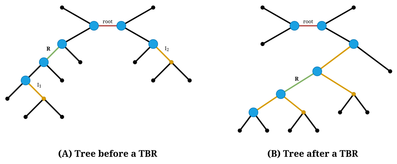

Figure 6 – Highlighted nodes indicate the nodes that require recalculation.

MPBoot-TBR is proposed as a substitute for SPR hill-climbing, with two TBR search strategies: MPBoot-TBR-best and MPBoot-TBR-better. The search continues while MP score is improved, with unsuccessful iterations tracked to determine search termination.

PROPOSED METHOD 2: MPBoot-ACO

MPBoot-ACO is recommended as a better alternative to hill-climbing algorithms that only use SPR or TBR by combining ACO with NNI, SPR, and TBR operations.

The algorithm selects the most appropriate operation among NNI, SPR, and TBR, rather than the most intensive one, to maintain accuracy while improving computational speed.

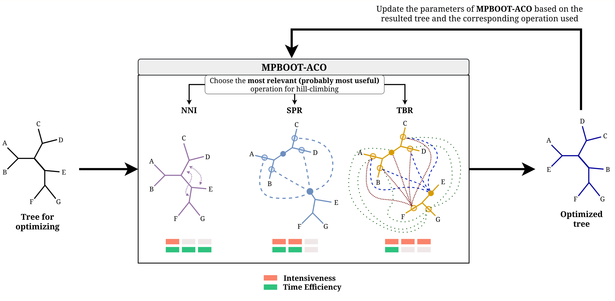

Figure 7 – MPBoot-ACO algorithm.

EXPERIMENTAL EVALUATION

- We compared four variations of MPBoot-TBR with the original MPBoot and also evaluated MPBoot-ACO against MPBoot-TBR-best (abbreviated as MPBoot-TBR) and the original MPBoot. The assessment was conducted on the high-performance computing system of VNU University of Engineering and Technology, using both biological and simulation data.
- MP Scores, Bootstrap Accuracy, and Computation Time were the evaluation metrics used to compare the algorithms.

Dataset	Sub-dataset	Data type	#MSAs	#taxa	#sites
Yule-Harding	YH1	DNA	200	100	500
	YH2		200	200	1000
	YH3		200	500	1000
	YH4	Protein	200	100	300
	YH5		200	200	500
TreeBASE	dna	DNA	70	201-767	976-61199
	prot	Protein	45	50-194	126-22426

Table 1 – Summary of the datasets.

- Evaluation results show that MPBoot-TBR is better than MPBoot in MP score and is equivalent to MPBoot in terms of bootstrap accuracy. Notably, MPBoot-ACO outperforms both MPBoot-TBR and the original MPBoot in MP score and has an advantage over MPBoot-TBR in computation time.

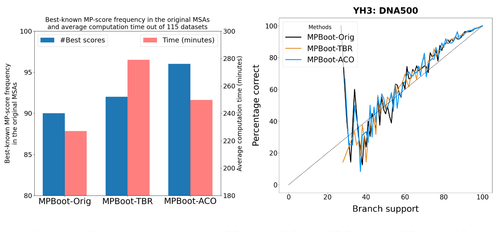

Table 2 – Evaluations result of MPBoot, MPBoot-TBR, and MPBoot-ACO

CONCLUSIONS

- We have proposed the MPBoot-TBR (with two different search strategies) and the MPBoot-ACO algorithm.
- The experimental results indicate that the performance in MP scores is improved with MPBoot-TBR and even further improved with MPBoot-ACO, while maintaining the same level of bootstrap accuracy as the original MPBoot. Moreover, MPBoot-ACO exhibits a faster processing time than MPBoot-TBR.
- Our implementation of the proposed algorithms is based on the open-source software MPBoot, available on Github for MPBoot-TBR and MPBoot-ACO.

RELATED PUBLICATION

- T. D. Huynh, Q. T. Vu, V. D. Nguyen and D. T. Hoang, "Employing tree bisection and reconnection rearrangement for parsimony inference in MPBoot," **2022 14th International Conference on Knowledge and Systems Engineering (KSE)**, Nha Trang, Vietnam, 2022, pp. 1-6, Oct. 2022. ISSN 2694-4804. DOI: https://doi.org/10.1109/KSE56063.2022.9953773.

REFERENCES

FELSENSTEIN, J. Confidence limits on phylogenies : an approach using the bootstrap. **Evolution**, Society for the Study of Evolution, Wiley, v. 39, n. 4, p. 783–791, 1985. ISSN 00143820, 15585646. DOI: https://doi.org/10.1111/j.1558-5646.1985.tb00420.x.

HOANG, D. T. et al. MPBoot: fast phylogenetic maximum parsimony tree inference and bootstrap approximation. **BMC Evolutionary Biology**, v. 18, n. 1, p. 11, Feb. 2018. ISSN 1471-2148. DOI: https://doi.org/10.1186/s12862-018-1131-3.

Address: E3 building, 144 Xuan Thuy road, Cau Giay district, Hanoi, Vietnam.
Tel.: +84-24-37547 461.
Website: http://www.vnu.edu.vn