BATCH-SCAMPP: Scaling Phylogenetic Placement Methods to Place Many Sequences

Eleanor Wedell
Department of Computer Science, University of Illinois Urbana-Champaign, IL, USA

Chengze Shen
Department of Computer Science, University of Illinois Urbana-Champaign, IL, USA

Tandy Warnow
Department of Computer Science, University of Illinois Urbana-Champaign, IL, USA

Abstract

Phylogenetic placement is the problem of placing one or more query sequences into a phylogenetic "backbone" tree, which may be a maximum likelihood tree on a multiple sequence alignment for a single gene, a taxonomy with leaves labeled by sequences for a single gene [7], or a species tree [4]. When the backbone tree is a tree estimated on a single gene, the most accurate techniques for phylogenetic placement are likelihood-based, and can be computationally intensive when the backbone trees are large [3]. Phylogenetic placement into gene trees occurs when updating existing gene trees with newly observed sequences, but can also be applied in the "bulk" context, where many sequences are placed at the same time into the backbone tree. For example, phylogenetic placement can be used to taxonomically characterize shotgun sequencing reads generated for an environmental sample in metagenomic analysis [7, 2].

The two most well known maximum likelihood phylogenetic placement methods are pplacer [5] and EPA-ng [2]. Of these two, EPA-ng is optimized for scaling the number of query sequences and is capable of placing millions of sequences into phylogenetic trees of up to a few thousand sequences [2], and achieves sublinear runtime in the number of query sequences (see Figure 2 from [1]).

Previously we introduced the SCAMPP framework [8] to enable both pplacer and EPA-ng to perform phylogenetic placement into ultra-large backbone trees, and we demonstrated its utility for placing into backbone trees with up to 200,000 sequences. By using maximum likelihood methods pplacer or EPA-ng within the SCAMPP framework, the resulting placements are more accurate than with APPLES-2 [1], with the most notable accuracy improvement for fragmentary sequences, and are computationally similar for single query sequence placement [8]. However, SCAMPP was designed to incrementally update a large tree, one query sequence at a time, and was not optimized for the other uses of phylogenetic placement, where batch placement of many sequencing reads is required.

Here we introduce BATCH-SCAMPP, a technique that improves scalability in both dimensions: the number of query sequences being placed into the backbone tree and the size of the backbone tree. Furthermore, BATCH-SCAMPP is specifically designed to improve EPA-ng's scalability to large backbone trees. Although BATCH-SCAMPP is based on SCAMPP, it uses a substantially modified design in order to be able to take advantage of EPA-ng's ability to place many query sequences efficiently.

The BATCH-SCAMPP method operates by allowing the input set of query sequences to suggest and then vote on placement subtrees, thus enabling many query sequences to select the same placement subtree. We pair BATCH-SCAMPP with EPA-ng to explore the capability of this approach for scaling to many query sequences. We show that this combination of techniques (which we call BSCAMPP+EPA-ng, or BSCAMPP(e)) not only provides high accuracy and scalability to large backbone trees, matching that of SCAMPP used with EPA-ng (i.e., SCAMPP(e)), but also achieves the goal of scaling sublinearly in the number of query sequences. Moreover, it is much more scalable than EPA-ng and faster than SCAMPP+EPA-ng: when placing 10,000 sequences into a backbone tree of 50,000 leaves, EPA-ng is unable to run due to memory issues, SCAMPP+EPA-ng requires 1421 minutes, and BSCAMPP(e) places all sequences in 7 minutes (all given the same computational resources. Figure 1 gives an example of this performance advantage on the nt78 [6] simulated dataset.

23rd International Workshop on Algorithms in Bioinformatics (WABI 2023).
Editors: Djamal Belazzougui and Aïda Ouangraoua; Article No. 3; pp. 3:1–3:2

Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

2012 ACM Subject Classification Applied computing → Bioinformatics

Keywords and phrases Phylogenetic Placement, EPA-ng, Phylogenetics

Digital Object Identifier 10.4230/LIPIcs.WABI.2023.3

Category Abstract

Related Version *Full Version*: https://doi.org/10.1101/2022.10.26.513936

Supplementary Material *Software (Source Code)*: https://github.com/ewedell/BSCAMPP_code archived at swh:1:dir:94380a8834bd8587970f2bd229b5ba62c733ca86

Funding The authors acknowledge the financial support of the Department of Computer Science. EW was supported by a Siebel Scholars scholarship, a SURGE Fellowship, and a Wing Kai Cheng Fellowship. CS was supported by NSF grant 2006069 (to TW).

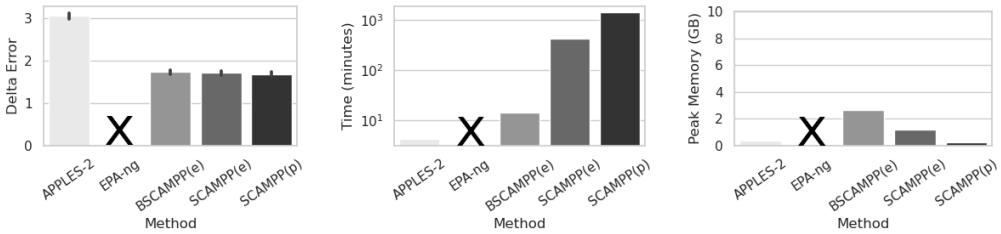

Figure 1 Results on the nt78 dataset (68,132 sequences in the backbone tree, placing 10,000 query sequences). Delta error measures placement error in number of edges, and is averaged across the query sequences. EPA-ng reported out-of-memory issues given 64 GB and 16 cores.

References

1 Metin Balaban, Yueyu Jiang, Daniel Roush, Qiyun Zhu, and Siavash Mirarab. Fast and accurate distance-based phylogenetic placement using divide and conquer. *Molecular Ecology Resources*, 22(3):1213–1227, 2022.

2 Pierre Barbera, Alexey M Kozlov, Lucas Czech, Benoit Morel, Diego Darriba, Tomáš Flouri, and Alexandros Stamatakis. EPA-ng: massively parallel evolutionary placement of genetic sequences. *Systematic Biology*, 68(2):365–369, 2019.

3 Gillian Chu and Tandy Warnow. SCAMPP+FastTree: improving scalability for likelihood-based phylogenetic placement. *Bioinformatics Advances*, 3(1), January 2023. vbad008. doi:10.1093/bioadv/vbad008.

4 Yueyu Jiang, Metin Balaban, Qiyun Zhu, and Siavash Mirarab. DEPP: deep learning enables extending species trees using single genes. *Systematic Biology*, 72(1):17–34, 2023.

5 Frederick A Matsen, Robin B Kodner, and E Virginia Armbrust. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC bioinformatics*, 11(1):538, 2010.

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.

7 Nidhi Shah, Erin K. Molloy, Mihai Pop, and Tandy Warnow. TIPP2: metagenomic taxonomic profiling using phylogenetic markers. *Bioinformatics*, 2021. doi:10.1093/bioinformatics/btab023.

8 Eleanor Wedell, Yirong Cai, and Tandy Warnow. SCAMPP: Scaling alignment-based phylogenetic placement to large trees. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pages 1417–1430, 2022. doi:10.1109/TCBB.2022.3170386.