APPLES: Fast Distance-Based Phylogenetic Placement

Metin Balaban[1], Shahab Sarmashghi[2], and Siavash Mirarab[2(✉)]

[1] Bioinformatics and Systems Biology, UC San Diego, La Jolla, CA 92093, USA
[2] Electrical and Computer Engineering, UC San Diego, La Jolla, CA 92093, USA
smirarab@ucsd.edu

Keywords: Phylogenetic placement · Distance-based methods · Genome-skimming

Extended Abstract

Methods for inferring phylogenetic trees from very large datasets exist, yet, large-scale tree reconstructions still require significant resources. New species are continually being sequenced, and as a result, even large trees can become outdated. Reconstructing the tree *de novo* each time new sequences become available is not practical. An alternative approach is phylogenetic placement where new sequence(s) are simply added to an existing *backbone* tree. Phylogenetic placement has applications other than updating trees, including sample identification, where the goal is to detect the identity of given *query* sequences of unknown origins. This problem arises [3] in the study of mixed environmental samples that make up much of the microbiome literature. Sample identification is also the essence of barcoding and meta-barcoding, methods used often in biodiversity studies.

Maximum Likelihood (ML) methods of phylogenetic placement are now available and in wide use (e.g., [4] and EPA(-ng) [2]). The ML approach is computationally demanding, and in particular requires large amounts of memory, and therefore, is limited in the size of the backbone tree it can use. More fundamentally, existing placement tools take as input alignments of assembled sequences for the backbone set, even when queries allowed to be unassembled reads. This reliance on assembled sequences makes them unsuitable for alignment and assembly-free scenarios. For example, sample identification using genome-skimming is fast becoming cost-effective. Methods like Skmer [5] (introduced in RECOMB 2018) can be used to infer k-mer-based estimates of phylogenetic distance from genome skims, and these distances can potentially be used for placement on phylogenetic trees. However, existing methods cannot be used for this purpose.

Distance-based phylogenetics has a rich methodological history, and yet, there are no existing tools for distance-based phylogenetic placement. Such methods, if developed, can be scalable to ultra-large backbone trees. Moreover,

L. J. Cowen (Ed.): RECOMB 2019, LNBI 11467, pp. 287–288, 2019.
https://doi.org/10.1007/978-3-030-17083-7

distance-based methods only need distances, not assembled sequences, and therefore, can be used for sample identification from reads in an assembly-free and alignment-free fashion.

We have developed a new method for distance-based phylogenetic placement called APPLES (Accurate Phylogenetic Placement using LEast Squares). APPLES finds the placement of a query sequence that minimizes the least square error of phylogenetic distances with respect to sequence distances. It can also operate on the minimum evolution principle, or a hybrid of minimum evolution and least square error. Using dynamic programming, APPLES is able to perform placement in time and memory that both scale linearly with the size of the backbone tree.

We have performed extensive studies on simulated and real datasets to evaluate APPLES. Our results show that in the alignment-based scenario, APPLES is much faster than ML tools, uses much less memory, and is very close to ML in the accuracy. Moreover, APPLES can handle much larger backbone trees (we have tested up to 200,000 leaves), and has *increased* accuracy when the backbone trees become larger and more densely sampled. In contrast, ML methods cannot handle backbones with several thousand species. For assembly-free scenarios, we study three genome skimming datasets of insects and show that APPLES applied to Skmer distances can accurately identify genome skim samples using coverage below 1X [1]. APPLES is open-source and freely available at https://github.com/balabanmetin/apples.

References

1. Balaban, M., Sarmashghi, S., Mirarab, S.: Apples: Fast distance-based phylogenetic placement. bioRxiv (2018). https://doi.org/10.1101/475566. https://www.biorxiv.org/content/early/2018/11/23/475566
2. Barbera, P., et al.: EPA-ng: massively parallel evolutionary placement of genetic sequences. BioRxiv, 291658 (2018)
3. Janssen, S., et al.: Phylogenetic placement of exact amplicon sequences improves associations with clinical information. mSystems **3**(3), 00021–18 (2018). https://doi.org/10.1128/mSystems.00021-18. http://msystems.asm.org/lookup/doi/10.1128/mSystems.00021-18
4. Matsen, F.A., Kodner, R.B., Armbrust, E.V.: pplacer: linear time maximum-likelihood and bayesian phylogenetic placement of sequences onto a fixed reference tree. BMC Bioinf. **11**(1), 538 (2010)
5. Sarmashghi, S., Bohmann, K., Gilbert, M.T.P., Bafna, V., Mirarab, S.: Assembly-free and alignment-free sample identification using genome skims. Genome Biology (abstract appeared at RECOMB 2018) (2018, in press). https://doi.org/10.1101/230409. https://www.biorxiv.org/content/early/2018/04/02/230409