ALGORITHMS FOR LARGE-SCALE EVOLUTIONARY TREE CONSTRUCTION: IMPROVING SCALABILITY AND ACCURACY THROUGH DIVIDE-AND-CONQUER

Allocation: Illinois/150 Knh
PI: Tandy Warnow[1]
Co-PIs: William Gropp[1], Erin Molloy[1], Pranjal Vachaspati[1]

[1]University of Illinois at Urbana–Champaign

EXECUTIVE SUMMARY

Evolutionary trees are used to advance the understanding of how life evolved on Earth, how species adapt to their environments, and to predict the structure and function of proteins. However, despite large numbers of whole genomes and increasing amounts of biomolecular sequence data available for use, the inference of a Tree of Life is beyond the reach of current methods, as even relatively small data sets can require many CPU years for analysis. This project aimed to develop new algorithms for large-scale evolutionary tree estimation, focusing on conditions with large numbers of species and/or whole genomes. Specific contributions of this work include new algorithmic strategies that greatly improve the scalability and accuracy of powerful statistical methods so they can be used to construct phylogenies on ultralarge data sets of importance to biologists.

FS

BI

RESEARCH CHALLENGE

Biologists use evolutionary trees to improve the understanding of how species evolve and adapt to their environments, to predict protein structure and function, to explore the early origins of life and how humans moved across the globe, and the like. The advances in whole genome assembly have suggested that the accurate inference of a Tree of Life may be achievable. However, despite large numbers of whole genomes and increasing amounts of biomolecular sequence data available for use, the inference of a Tree of Life is much more challenging than was expected. The main issues impeding this are: (1) the inference of these evolutionary trees is computationally challenging, as the most accurate methods are based on attempts to solve hard optimization problems (such as maximum likelihood) and current optimization methods do not scale to large data sets with good accuracy; and (2) standard approaches to phylogeny estimation, which make strong assumptions about the homogeneity of the statistical process underlying the molecular sequence data, do not have good accuracy in the presence of heterogeneity across genomes and across time, and yet substantial heterogeneity is now well established [1–3]. While some methods have been developed to enable phylogeny estimation in the presence of heterogeneity across genomes, these methods are computationally intensive, even on just moderately large data sets. Hence, current approaches to phylogeny estimation either do not provide good accuracy on large data sets or cannot even run on large data sets within reasonable timeframes.

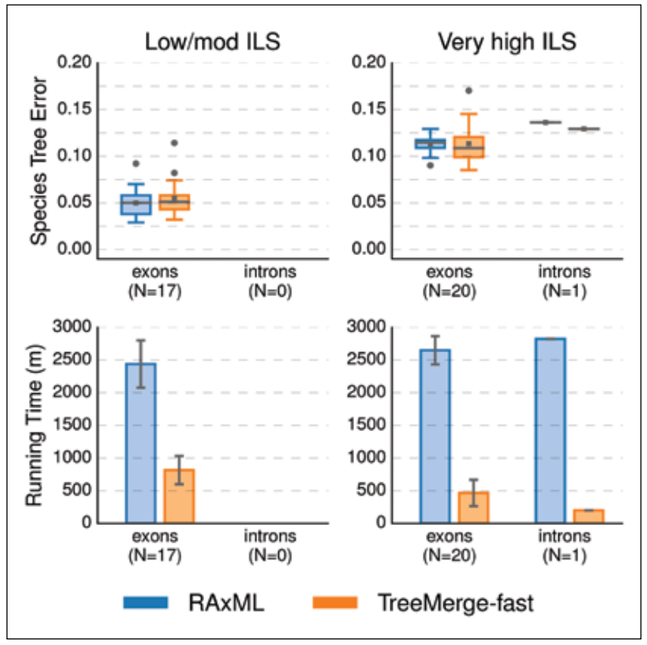

Figure 1: The result of using TreeMerge (a divide-and-conquer strategy) with the leading maximum likelihood method, RAxML, in species tree construction on 1,000 species with 1,000 genes when there is gene tree heterogeneity owing to incomplete lineage sorting [5]. The team explored results for two types of genes: exons and introns. The number of replicates for which RAxML returns a tree is given by N; when run by itself, RAxML cannot complete on some data sets within 48 hours on Blue Waters, but when run within the divide-and-conquer framework, it completes on all data sets. ("ILS" refers to incomplete lineage sorting.)

METHODS & CODES

The research team developed a divide-and-conquer framework that can be used with any method for constructing trees. The input is a set of species represented, for example, by sets of gene trees or DNA sequences. In the first step, the set of species is divided into disjoint subsets. Then, trees are constructed on each of the subsets using a selected method for phylogeny estimation. Finally, the trees are merged together using a distance matrix that is computed on the species. The key algorithmic innovation is the ability to merge trees together using the distance matrix. For this project, the team developed three such methods: NJMerge [4], TreeMerge [5], and INC [6]. The team used the divide-and-conquer strategies with the leading current methods for phylogeny estimation, including the maximum likelihood method RAxML [7] and ASTRAL [8,9], which estimate species trees by combining gene trees.

RESULTS & IMPACT

The major contributions of this project are methods for large-scale evolutionary tree estimation that are capable of constructing highly accurate trees on data sets that are currently beyond the reach of existing methods. Most importantly, the divide-and-conquer strategies, TreeMerge and NJMerge, are enabling highly accurate species tree estimation using genome-scale data on tens of thousands of species, where current methods fail to even run (owing to memory or time limitations) on these data sets. Using divide-and-conquer has reduced the running time of current methods on ultralarge data sets from multiple days to a few hours, thus dramatically advancing the capability of biologists to make biological discoveries. Specifically, this research has shown that TreeMerge and NJMerge can enable highly accurate species trees on tens of thousands of species from whole genome data sets without the need for supercomputers; this will transform research for evolutionary biologists. Furthermore, the algorithmic approach is quite general and should be useful for other problems where trees (even if not evolutionary trees) are constructed.

WHY BLUE WATERS

The design of new algorithms for phylogeny estimation is an iterative process in which algorithmic strategies are explored and tested, and results are then used to improve the algorithm design. Since each analysis can be computationally intensive, this process requires large resources. The use of Blue Waters enabled this team to explore the design space effectively, and to produce new statistical and computational methods with outstanding accuracy and scalability to large and ultralarge data sets.

PUBLICATIONS & DATA SETS

E. K. Molloy and T. Warnow, "NJMerge: A generic technique for scaling phylogeny estimation methods and its application to species trees," in *Comparative Genomics: RECOMB-CG 2018, Lecture Notes in Computer Science*, Magog–Orford, QC, Canada, Oct. 9–12, 2018, pp. 260–276.

E. K. Molloy and T. Warnow, "TreeMerge: A new method for improving the scalability of species tree estimation methods," *Bioinformatics*, vol. 35, no. 14, pp. i417–i426, Jul. 2019.

T. Le, A. Sy, E. K. Molloy, Q. Zhang, S. Rao, and T. Warnow, "Using INC within divide-and-conquer phylogeny estimation," in *Proc. Int. Conf. Algorithms for Comput. Biol.*, Berkeley, CA, U.S.A., May 28–30, 2019, pp. 167–178.