DI

BI

A PARALLEL FRAMEWORK FOR SCALING PHYLOGENY ESTIMATION METHODS TO LARGE GENOMIC DATA SETS

Allocation: Exploratory/50 Knh
PI: William Gropp[1]
Co-PI: Erin Molloy[1]
Collaborator: Tandy Warnow[1]

[1]University of Illinois at Urbana–Champaign

EXECUTIVE SUMMARY

Phylogenetic trees are graphical models of evolution that can be used to study how species evolve and adapt to their environment. Recent advances in sequencing technology have resulted in an explosion of data and the creation of ultralarge data sets. Today, scientists need to estimate highly accurate phylogenetic trees from these data sets; however, this process is computationally challenging. Many of the best methods are not easily parallelizable and have large computational footprints (memory and running time).

The research team's main result is a parallel framework for scaling phylogeny estimation methods to large data sets while maintaining accuracy. By running their preferred method within the team's parallel framework, scientists will be able to estimate evolutionary trees using reduced computational resources. Thus, these computational tools will be useful for researchers attempting to build the Tree of Life, a scientific and computational Grand Challenge, as well as researchers in medicine, agriculture, and other domains.

RESEARCH CHALLENGE

Phylogenetic trees are graphical models of evolution that can be used to study evolutionary processes such as how species evolve and adapt to their environments. Such models are useful in a variety of applications, including the prediction of protein function or the classification of molecular sequences of unknown origins. Recent advances in sequencing technologies have resulted in an explosion of data, including genome-scale data for very large numbers of species. Today, scientists need to estimate highly accurate phylogenetic trees from such data sets; however, building these evolutionary trees from genome-scale data sets is challenging. For example, genome-scale data sets can be more heterogeneous owing to incomplete lineage sorting (ILS) and other biological processes that result in different regions of the genome having different evolutionary histories. In addition, many of the best methods for phylogeny estimation are heuristics for solving NP (nondeterministic polynomial-time)-hard optimization problems, and the solution space for these problems grows exponentially with the number of species. Parallelizing the existing heuristics does not typically reduce the computational effort required to search the solution space. Thus, even high-performance computing implementations can require many CPU years to run on large, heterogeneous data sets.

METHODS & CODES

Divide-and-conquer is a technique used by computer scientists to run methods on large data sets. The traditional divide-and-conquer approach for estimating large evolutionary trees operates by: (1) dividing the species into overlapping subsets, (2) estimating trees on each subset, and (3) combining subset trees together by solving the supertree problem. However, supertree estimation is also computationally challenging (the best methods are heuristics for solving NP-hard optimization problems and the solution space for these problems grows exponentially with the number of species). In order to overcome this challenge, the research team introduced a new divide-and-conquer approach that divides the species into pairwise disjoint subsets and combines subset trees together by solving the disjoint tree merger (DTM) problem, which can be solved in polynomial time. TreeMerge, the DTM method developed for this Blue Waters project, is a dramatic improvement on NJMerge, requiring far less memory and far less communication.

RESULTS IMPACT

The team implemented the divide-and-conquer approach (using TreeMerge to combine subset trees) as a parallel framework and performed a benchmarking study on Blue Waters. They compared running methods within the parallel framework (in order to estimate subset trees) versus running methods to estimate a tree on the full data set. The researchers found that running methods within the parallel framework dramatically reduced running time and maintained accuracy for two leading species tree estimation methods: ASTRAL-III [1] (Fig. 1) and RAxML [2] (not shown). Novel species tree methods are continually being developed, and many of these new methods are computationally intensive (for example, new Bayesian inference methods). Because researchers can specify that the method be run on subsets, the research team's parallel framework will be useful in the rapidly progressing field of computational biology and will further research efforts toward building the Tree a Life, a scientific and computational grand challenge.

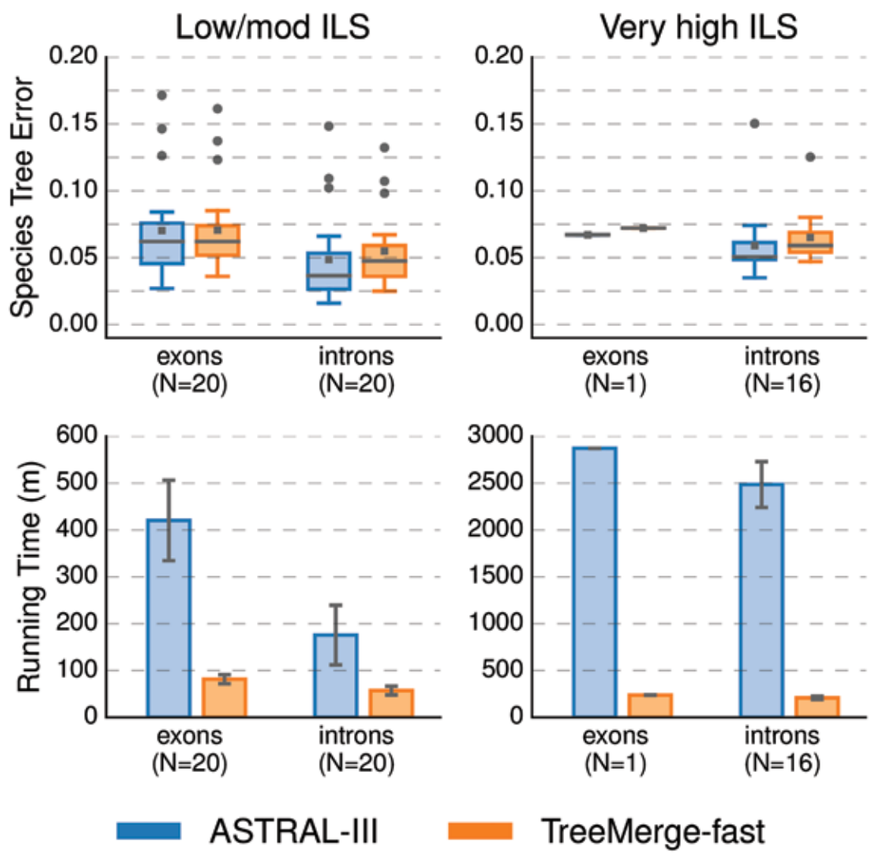

Figure 1: Running ASTRAL-III within the research team's parallel framework (TreeMerge) maintains accuracy (top row) and reduces running time (bottom row). Left and right columns show two model conditions: low/moderate and very high levels of ILS (incomplete lineage sorting). The *x*-axis indicates the sequence type (exon or intron) and the number *N* of replicates where ASTRAL-III completed.

WHY BLUE WATERS

Method development is an iterative process that requires testing any new method on many large data sets. This process also requires that the new method be compared in terms of accuracy to the best existing methods; running the existing methods on large data sets can take months of CPU time. Without the Blue Waters system, the research team would have been unable to efficiently develop and improve upon novel methods.

PUBLICATIONS & DATA SETS

E. Molloy and T. Warnow, "Statistically consistent divide-and-conquer pipelines for phylogeny estimation using NJMerge," *Algo. Mol. Biol.*, vol. 14, no. 1, article 14, Jul. 2019, doi: 10.1186/s13015-019-0151-x.

Data from statistically consistent divide-and-conquer pipelines for phylogeny estimation using NJMerge, Illinois Data Bank, 2019. [Online]. Available: https://doi.org/10.13012/B2IDB-0569467_V2

E. 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, July 2019, doi: 10.1093/bioinformatics/btz344.

Data from TreeMerge: A new method for improving the scalability of species tree estimation methods, Illinois Data Bank, 2019. [Online]. Available: https://doi.org/10.13012/B2IDB-9570561_V1