DI

ML

CONSTRUCTING LARGE EVOLUTIONARY TREES ON SUPERCOMPUTERS

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

[1]University of Illinois at Urbana-Champaign

EXECUTIVE SUMMARY

The organization of molecular sequences into evolutionary trees, called phylogenies, enables researches to study the evolution of bacteria and viruses that cause disease outbreaks and to identify previously unrecognized microbial organisms found in environmental samples, such as the soil or the human gut. The current and leading approaches to phylogenetic inference require two steps: first, a multiple sequence alignment is estimated and then a tree is estimated from the alignment. This two-phase approach is not scalable. We used Blue Waters to design and test a novel approach that bypasses (1) alignment estimation on the full dataset, (2) maximum likelihood tree estimation on the full dataset, and (3) supertree estimation. This work is a major advancement towards constructing the Tree of Life using supercomputers.

RESEARCH CHALLENGE

The organization of molecular sequences into evolutionary trees, called phylogenies, enables researches to study the evolution of bacteria and viruses that cause disease outbreaks and identify previously unrecognized microbial organisms found in the human gut [1]. The current and leading approaches to phylogenetic inference require two steps: first, a multiple sequence alignment is estimated and then a tree is estimated from the alignment. A multiple sequence alignment is an n x l matrix, where n is the number of sequences and l is the alignment length. Building alignments on large numbers of heterogeneous sequences is computationally intensive, and the leading methods produce very long alignments that require large amounts of storage [2]. Maximum likelihood methods are widely accepted as the gold standard for phylogenetic inference on single gene data sets. Building maximum likelihood trees on large numbers of sequences is also computationally challenging given that the number of possible tree topologies increases exponentially with the number of sequences and many numerical parameters must be optimized for each candidate tree topology. Most parallel codes for estimating phylogenetic trees are implemented to handle long alignments but not large numbers of sequences [3, 4]. DACTAL [5], a divide-and-conquer method, handles large numbers of sequences by dividing them into overlapping subsets, constructing trees on subsets, and then merges the trees into a "supertree" using heuristics for NP-hard optimization problems. These heuristic methods for constructing supertrees do not scale to ultra-large datasets. Hence, we worked to design, prototype, and test a parallel algorithm for constructing evolutionary trees on large numbers of sequences.

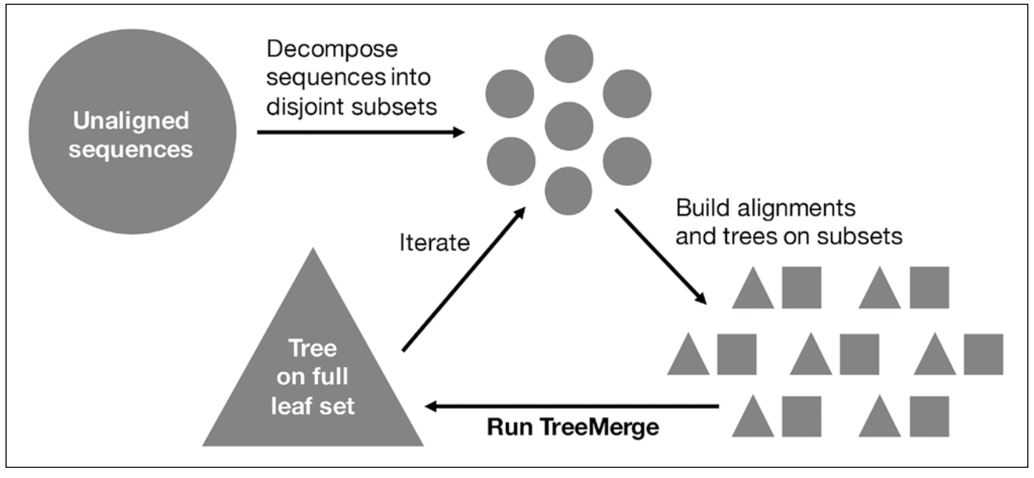

Figure 1: Our approach (TERADACTAL) divides unaligned sequences into disjoint subsets (circles), builds alignments (squares) and trees (triangles) on each subset, and then merges the subset trees together in polynomial time with a highly accurate technique called *TreeMerge*. This process can iterate by decomposing the tree on the full dataset into subsets.

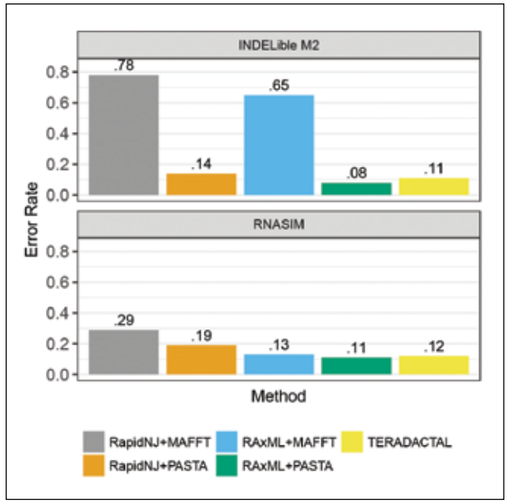

Figure 2: We compared TERADACTAL to methods that compute an alignment (using MAFFT [4] or PASTA [2]) and then estimate a tree (using RapidNJ [5] or RAxML [6]). TERADACTAL achieved an error rate within 1–3% of the leading two-phase method for the two model conditions. Bars are averages over ten replicate datasets.

METHODS & CODES

Our approach, called "TERADACTAL," divides unaligned sequences into disjoint (rather than overlapping) subsets, builds alignments and trees on each subset, and merges the subset trees together using a highly accurate and polynomial time technique we developed called *TreeMerge*. *TreeMerge* builds a minimum spanning tree on the subsets, and the edges in the minimum spanning tree indicate pairs of subset trees to be merged together using constrained Neighbor-Joining. Because the topologies of these merged trees do not conflict with the original subset trees, all of subset trees can be merged together into a tree on the full dataset using the minimum spanning tree.

TERADACTAL can iterate using the tree on the full dataset to create the next subset decomposition. We prototyped TERADACTAL in Python, and it is freely available on Github (https://github.com/ekmolloy/teradactal-prototype).

RESULTS & IMPACT

We used the Blue Waters supercomputer to perform a simulation study comparing TERADACTAL to several two-phase methods, including three alignment methods (only two shown: MAFFT [4] or PASTA [2]) and four tree estimation methods (only two shown: RapidNJ [9] and RAxML [8]). RapidNJ is distance-based method that uses polynomial time to construct a tree from a distance matrix, whereas RAxML [4] is one of the most (if not the most) widely used maximum likelihood codes for phylogenetic inference with over 11,000 citations. We found that TERADACTAL achieves similar error rates (within 1-3%) of the best methods tested. Hence, TERADACTAL achieves similar error rates to the leading two-phase methods but is highly parallel and can handle large number of sequences by avoiding (1) alignment estimation on the full dataset, (2) maximum likelihood tree estimation on the full dataset, and (3) supertree estimation.

This work is a major advancement toward constructing the Tree of Life using supercomputers.

WHY BLUE WATERS

We used Blue Waters to demonstrate that existing parallel codes (e.g., PASTA and RAxML) could not run on datasets with one million sequences on Blue Waters. We also used Blue Waters to extensively test the TERADACAL prototype and compare the TERADACTAL prototype to the leading and popular two-phase methods. Specifically, we performed a large simulation study requiring over 36,000 node hours. These analyses were completed in under a month but would have required over a year to run on our laboratory's four nodes on the campus cluster.

PUBLICATIONS & DATA SETS

Molloy, E.K., and T. Warnow, TERADACTAL: A scalable Divide-And-Conquer approach for constructing large phylogenetic Trees (almost) without Alignments. In progress.