BLUE WATERS
SUSTAINED PETASCALE COMPUTING

Allocation: Exploratory/50 knh
PI: William Gropp
Co Pi's: Tandy Warnow, Erin Molloy
University of Illinois at Urbana-Champaign
Biology, Chemistry & Health

CONSTRUCTING LARGE EVOLUTIONARY TREES ON SUPERCOMPUTERS

Research Challenge

Evolutionary trees (called phylogenies) are important for addressing many questions in basic biology and applied research (for example, phylogenies can be used to identify microorganisms living in the human gut). However, phylogeny estimation from genetic data is computationally challenging. The leading approaches to phylogeny estimation have two steps: 1) a multiple sequence alignment is estimated from the input genetic data and 2) a Maximum Likelihood tree is estimated from the alignment. Both steps do not scale to genetic datasets with large numbers of unaligned sequences.

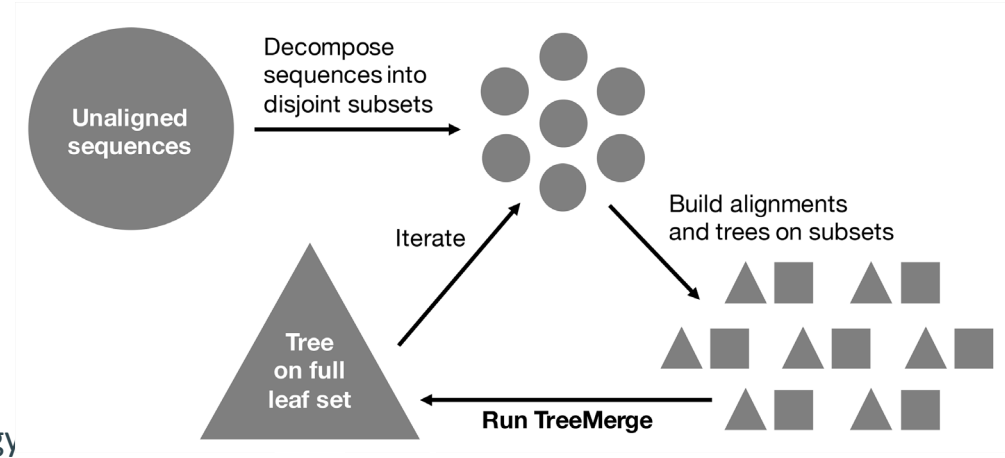

Figure 1: **TERADACTAL** divides unaligned sequences into **disjoint subsets** (circles), builds alignments (squares) and trees (triangles) on each subset, and then merges the subset trees together in polynomial time using a novel technique, called **TreeMerge**. This process can iterate by decomposing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

Methods & Codes

The team's approach, called **TERADACTAL**, is shown in Figure 1. The divide-and-conquer approach enables alignments and trees to be estimated on subsets in parallel; however, the innovation is a newly developed technique, called TreeMerge, that combines subset trees in polynomial time without sacrificing accuracy. Because TreeMerge combines subset trees in a pair-wise fashion, the conquer phase can also be performed in parallel. TERADACTAL is freely available on Github (https://github.com/ekmolloy/teradactal-prototype).

Results & Impact

TERADACTAL achieved similar error rates (within 1-3%) of the best two-phase methods tested (e.g., RAxML given the true alignment). Thus, TERADACTAL can achieve similar error rates to the leading two-phase methods but is highly parallelizable and avoids computationally challenging tasks, such as, alignment estimation on the full dataset, Maximum Likelihood tree estimation on the full dataset, and supertree estimation. **This work is a major advancement toward constructing the Tree of Life using supercomputers.**

Why Blue Waters

Blue Waters was used to demonstrate that existing parallel codes (e.g., PASTA and RAxML) could not effectively run on datasets with one million sequences on Blue Waters, to extensively test the TERADACAL prototype, and to compare the TERADACTAL prototype to the leading two-phase methods. Specifically, the team 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 other available systems.