Phylogenomics, Multiple Sequence Alignment, and Metagenomics

Tandy Warnow
University of Illinois at Urbana-Champaign
Phylogeny (evolutionary tree)

From the Tree of the Life Website,
University of Arizona
Phylogeny + genomics = genome-scale phylogeny estimation.
Multiple Sequence Alignment (MSA): a scientific grand challenge

Novel techniques needed for scalability and accuracy

NP-hard problems and large datasets
Current methods do not provide good accuracy
Few methods can analyze even moderately large datasets

Many important applications besides phylogenetic estimation

---

1 Frontiers in Massive Data Analysis, National Academies Press, 2013
Estimating the Tree of Life

Basic Biology:
How did life evolve?

Applications of phylogenies and multiple sequence alignments to:
protein structure and function
population genetics
human migrations
metagenomics

Figure from https://en.wikipedia.org/wiki/Common_descent
Species sequenced by year

- Viruses
- Eukaryotes
- Prokaryotes

Number of species vs. Year

Muir, 2016
Estimating the Tree of Life

Large datasets!
Millions of species
thousands of genes

NP-hard optimization problems
Exact solutions infeasible
Approximation algorithms
Heuristics
Multiple optima

High Performance Computing:
necessary
but not sufficient

Figure from
https://en.wikipedia.org/wiki/Common_descent
Computational Phylogenetics (2005)

Current methods can use months to estimate trees on 1000 DNA sequences

Our objective:
More accurate trees and alignments on 500,000 sequences in under a week

Courtesy of the Tree of Life web project, tolweb.org
Computational Phylogenetics (2018)

1997-2001: Distance-based phylogenetic tree estimation from polynomial length sequences

2012: Computing accurate trees (almost) without multiple sequence alignments

2009-2015: Co-estimation of multiple sequence alignments and gene trees, now on 1,000,000 sequences in under two weeks

2014-2015: Species tree estimation from whole genomes in the presence of massive gene tree heterogeneity

2016-2017: Scaling methods to very large heterogeneous datasets using novel machine learning and supertree methods.

Courtesy of the Tree of Life web project, tolweb.org
Metagenomic taxonomic identification and phylogenetic profiling

Metagenomics, Venter et al., Exploring the Sargasso Sea: Scientists Discover One Million New Genes in Ocean Microbes
Basic Questions

1. What is this fragment? (Classify each fragment as well as possible.)

2. What is the taxonomic distribution in the dataset? (Note: helpful to use marker genes.)

3. What are the organisms in this metagenomic sample doing together?
Some of Our Methods and Software


- **ASTRAL** (species tree estimation from multi-locus data), S. Mirarab et al., Bioinformatics (2014), 30(17): i541-i548.


All software available in open-source form on github. See [http://tandy.cs.illinois.edu/software.html](http://tandy.cs.illinois.edu/software.html) for more information.