Scaling Species Tree Estimation to Large Datasets

Tandy Warnow
The University of Illinois
Phylogeny + genomics = genome-scale phylogeny estimation.
Phyogenomic Pipeline

• Assemble and annotate genomes (e.g., determine orthologs)
• Compute multiple sequence alignments of individual loci
• Construct gene trees
• Construct species tree
• Perform post-tree analyses (e.g., estimate dates, infer selection, etc.)
Phylogenomic Pipeline

- Assemble and annotate genomes (e.g., determine orthologs)
- Compute multiple sequence alignments of individual loci
- Construct gene trees
- Construct species tree
- Perform post-tree analyses (e.g., estimate dates, infer selection, etc.)
Phyogenomic Pipeline

• Assemble and annotate genomes (e.g., determine orthologs)
• Compute multiple sequence alignments of individual loci
• Construct gene trees
• Construct species tree
• Perform post-tree analyses (e.g., estimate dates, infer selection, etc.)
PASTA and UPP
scaling MSA to 1,000,000 sequences

• PASTA (Mirarab et al., J. Computational Biology 2014) – an improvement over SATe (Liu et al., Science 2009)
  • PASTA available at https://github.com/smirarab/
  • PASTA+BAli-Phy at http://github.com/MGNute/pasta

• UPP (Nguyen et al., Genome Biology 2015) – uses PASTA for backbone, inserts sequences into the backbone using ensembles of HMMs, handles fragmentary sequences well
  – https://github.com/smirarab/sepp
Phyogenomic Pipeline

• Assemble and annotate genomes (e.g., determine orthologs)
• Compute multiple sequence alignments of individual loci
• **Construct gene trees**
• Construct species tree
• Perform post-tree analyses (e.g., estimate dates, infer selection, etc.)
Phyogenomic Pipeline

• Assemble and annotate genomes (e.g., determine orthologs)
• Compute multiple sequence alignments of individual loci
• Construct gene trees
• **Construct species tree**
• Perform post-tree analyses (e.g., estimate dates, infer selection, etc.)
I’ll use the term “gene” to refer to “c-genes”: recombination-free orthologous stretches of the genome.
Gene tree discordance

Multiple causes for discord, including
• Incomplete Lineage Sorting (ILS),
• Gene Duplication and Loss (GDL), and
• Horizontal Gene Transfer (HGT)
Gene tree discordance

Multiple causes for discord, including
- Incomplete Lineage Sorting (ILS),
- Gene Duplication and Loss (GDL), and
- Horizontal Gene Transfer (HGT)
Main competing approaches

- Gene 1
- Gene 2
- ... gene k

Species

Concatenation

Analyze separately

Summary Method
ASTRAL

• Mirarab and Warnow, Bioinformatics 2014
• https://github.com/smirarab/ASTRAL

Algorithmic approach:

• Given set of gene trees, find the species tree that agrees with the maximum number of quartet trees within a constrained search space.
• Polynomial time and statistically consistent in the presence of ILS.
ASTRID

• ASTRID: Accurate species trees using internode distances, Vachaspati and Warnow, RECOMB-CG 2015 and BMC Genomics 2015
• Github site: https://github.com/pranjalv123/ASTRID

Algorithmic design:
• Nearly the same as NJst (Liu and Yu, 2010)- computes a matrix of average leaf-to-leaf topological distances, and then computes a tree using FastME (more accurate than Neighbor Joining and faster, too).
• Polynomial time and statistically consistent in the presence of ILS.
Both ASTRAL and ASTRID substantially outperform MP-EST.
On the ASTRAL-2 dataset with 1000 taxa, 1000 genes, ASTRID-FastME takes 33 minutes, ASTRAL takes 12 hours.
Impact of Gene Tree Estimation Error
(from Molloy and Warnow 2017)

Summary Methods

Error is fraction of bipartitions that are not recovered

Note: Summary methods better than CA-ML for low GTEE, then worse!
Main competing approaches

- Analyze separately
- Concatenation
- Summary Method
Large datasets are challenging!

- RAxML is expensive!
  - 250 CPU years for the Avian Phylogenomics project with only 48 species
  - Not guaranteed to find global optimum, in any event
  - And concatenation analyses are not statistically consistent

- ASTRAL can also fail to complete in reasonable times for very challenging datasets (very high ILS, many taxa), though not as problematic as RAxML
Our Approach

• Divide-and-conquer pipeline to:
  – Improve scalability,
  – reduce running time, and
  – maintain accuracy for species tree estimation on large datasets
Divide-and-Conquer Pipeline

1. Decompose species set into *pairwise disjoint* subsets.
2. Build a tree on each subset.
3. Compute tree on entire set of species using "Disjoint Tree Merger" method.

Additional information: (e.g., distance matrix)
TreeMerge

- Molloy and Warnow, ISMB 2019 and Bioinformatics, Volume 35, Issue 14, July 2019, Pages i417–i426
- TreeMerge is a direct improvement to NJMerge (which it uses on pairs of trees)
- Github site: https://github.com/ekmolloy/treemerge

Algorithmic strategy:
- divide species set into disjoint subsets,
- compute species trees on the subsets using selected species tree method, and
- merge subset trees using a distance-based method
DTMs Merge Subset Trees

Notes:
• Subset trees are requirements (constraint trees)
• Blending is permitted!
TreeMerge study

• Datasets: 1000 species and 1000 genes (exons and introns), two levels of ILS (low/mod and very high)

• Base methods: ASTRAL and with RAxML

• Criteria: Impact on base method
  – How many replicates complete?
  – Species tree error
  – Running time?
Impact of using TreeMerge with RAxML on 1000 species and 1000 genes
Impact of using TreeMerge with ASTRAL-III on 1000 species and 1000 genes
Summary for TreeMerge

• **TreeMerge**
  – Improves scalability (ASTRAL and RAxML can fail to complete in 48 hours, but not when run with TreeMerge)
  – Reduces running time, sometimes dramatically
  – Maintains accuracy
Other DTMs

- **NJMerge**: Molloy and Warnow, RECOMB-CG 2018 and Alg Mol Biol (to appear)
- **Constrained-INC**: Zhang, Rao, and Warnow, WABI 2018, Le et al., AlCoB 2019, and submitted
- **GTM (Guide Tree Merger)**: Smirnov and Warnow, RECOMB-CG 2019 and BMC Genomics (submitted)
  - Similar accuracy to TreeMerge in most cases, but faster
  - Does not permit blending
Summary

Recommendations for large and ultra-large datasets:

- MSA: PASTA and UPP
- Gene trees: ML methods, such as RAxML
- Species trees: ASTRAL, ASTRID, or concatenation using ML
- Ultra-large datasets? Use a Disjoint Tree Merger method, such as TreeMerge

Downloadable software:

- GTM: https://github.com/vlasmirnov/GTM
- TreeMerge: https://github.com/ekmolloy/treemerge
- PASTA: https://github.com/smirarab/pasta
- UPP: https://github.com/smirarab/sepp
- ASTRAL: https://github.com/smirarab/ASTRAL
- ASTRID: https://github.com/pranjalv123/ASTRID-1
Acknowledgments

Software: all open source, see http://tandy.cs.illinois.edu/software.html

NSF grants DBI-1461364, ABI-1458652, CCF 1535977
NSF graduate fellowships to Pranjal Vachaspati and Erin Molloy
HHMI graduate fellowship to Siavash Mirarab
Computing done at Blue Waters (part of NCSA)

Papers: http://tandy.cs.illinois.edu/papers.html
Accuracy in the presence of HGT + ILS

Data: Fixed, moderate ILS rate, 50 replicates per HGT rates (1)-(6), 1 model species tree per replicate on 51 taxa, 1000 true gene trees, simulated 1000 bp gene sequences using INDELible \(^8\), 1000 gene trees estimated from GTR simulated sequences using FastTree-2\(^7\)

---

\(^7\) Price, Dehal, Arkin 2015

\(^8\) Fletcher, Yang 2009

Davidson et al., RECOMB-CG, BMC Genomics 2015