Mathematical and Computational Statistics
Challenges in Reconstructing Evolution

Tandy Warnow
The University of Illinois
Phylogeny
(evolutionary tree)

From the Tree of the Life Website,
University of Arizona
Phylogenomics

Phylogeny + genomics = genome-scale phylogeny estimation
• “Nothing in biology makes sense except in the light of evolution”

• “...... nothing in evolution makes sense except in the light of phylogeny ...”
Phylogenetic Inference

“Big Data”:
- Heterogeneous
- Large
- Noisy
- Error-ridden
- Streaming
- Model-misspecification

Approaches:
- NP-hard optimization problems and large datasets
- Statistical estimation under stochastic models of evolution
- Probabilistic analysis of algorithms
- Graph-theoretic divide-and-conquer
- Chordal graph theory
- Combinatorial optimization
This Talk

• Models of evolution, identifiability, statistical consistency
• Genome-scale phylogeny:
  – Incomplete lineage sorting and species tree estimation under the Multi-Species Coalescent model (MSC)
  – ASTRAL: non-parametric accurate and statistically consistent species tree estimation under the MSC
  – TreeMerge: scaling species tree methods to large datasets
• Discussion and Future directions
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.)
DNA Sequence Evolution (Idealized)

AAGACTT

-3 mil yrs -2 mil yrs -1 mil yrs
today

AAGGCCT

AGGGCAT

AGGGCAT TAGCCCA TAGACTT AGCACAA AGCGCTT

AAGACTT

TGGACTT

AGCACTT
Phylogeny Problem

AGGGCAT  TAGCCCA  TAGACTT  TGCACAA  TGCGCTT
Statistical Consistency/Identifiability

error

Data
Questions

• Is the model tree identifiable?
• Which estimation methods are statistically consistent under this model?
• How much data does the method need to estimate the model tree correctly (with high probability)?
• What are the computational issues?
Answers?

• We know a lot about which site evolution models are identifiable, and which methods are statistically consistent.

• We know a little bit about the sequence length requirements for standard methods.

*Take home message: need to limit (or not allow) heterogeneity to get model identifiability!*
Computational issues

• **Maximum likelihood** (ML): NP-hard, and tree-space grows exponentially with the number of leaves
• **Bayesian estimation**: need to run to convergence (may fail)
• Parallelism helps but is not enough
Genome-scale data?
Phylogeny + genomics = genome-scale phylogeny estimation
Gene tree discordance

Incomplete Lineage Sorting (ILS) is a dominant cause of gene tree heterogeneity
Gene Trees inside the Species Tree (Coalescent Process)

Gorilla and Orangutan are not siblings in the species tree, but they are in the gene tree.
Gorilla and Orangutan are not siblings in the species tree, but they are in the gene tree.
1KP: Thousand Transcriptome Project

- 103 plant transcriptomes, 400-800 single copy “genes”
- Next phase will be much bigger
- Wickett, Mirarab et al., *PNAS* 2014

**Major Challenge:**
- Massive gene tree heterogeneity consistent with ILS
Avian Phylogenomics Project

- Approx. 50 species, whole genomes
- 14,000 loci
- Multi-national team (100+ investigators)
- 8 papers published in special issue of Science 2014

Major challenge:
- Massive gene tree heterogeneity consistent with ILS.
Hierarchical Model: MSC+GTR

• Multi-locus data, generated by a hierarchical model
  – Species tree generates gene trees under MSC
  – Gene trees generate sequences under the Generalized Time Reversible (GTR) model
Gene evolution model

Species tree

Gorilla  Human  Chimp  Orangutan

Gene tree

Gene tree

Gene tree

Gene tree

Sequence evolution model

Sequence data (Alignments)

ACTGCACACCG
ACTGC–CCCCCG
AATGC–CCCCCG
–CTGCACACCG

CTGAGC–ATCG
CTGAGC–TCG
ATGAGC–TC–
CTGA–CAC–G

AGCAGC–ATCGTG
AGCAGC–TCGTG
AGCAGC–TC–TG
C–TA–CACGGTG

CAGGCACGCACGAA
AGC–CACGC–CATA
ATGGCACGC–C–TA
AGCTAC–CACGGAT

Sequence data (Alignments)
Gene evolution model

Sequence data (Alignments)

- ACTGCACACCG
- ACTGC--CCCCG
- AATGC--CCCCG
- --CTGCACACCGG
- CTGAGCATCG
- CTGAGC--TCG
- ATGAGC--TC-
- CTGA--CAC--G
- AGCAGCATCGTG
- AGCAGC--TCGTG
- AGCAGC--TC-TG
- C--TA--CACGGTG
- CAGGCACGCACGAA
- AGC--CACGC--CATA
- ATGGCACGC--C--TA
- AGCTAC--CACGGAT

Species tree

Chimp  Human  Orang.

Gene tree

Chimp  Human  Orang.

Gene tree

Chimp  Human  Orang.

Gene tree

Chimp  Human  Orang.

Sequence evolution model

Sequence data (Alignments)

2

Sequence data (Alignments)
Step 2: infer species trees

Step 1: infer gene trees (traditional methods)

Gene tree

ACTGCACACCG
ACTGC–CCCCG
AATGC–CCCCG
–CTGCACACCGG

Gene tree

CTGAGCATCG
CTGAGC–TCG
ATGAGC–TC–
CTGA–CAC–G

Gene tree

AGCAGCATCGTG
AGCAGC–TCGTG
AGCAGC–TC–TG
C–TA–CACGGTG

Gene tree

CAGGCACGCACGAA
AGGC–CAGGC–CAT
ATGGCAGGC–C–TA
AGCTAC–CAGGAT
Theorem (Allman et al.): Under the multi-species coalescent model, for any four taxa A, B, C, and D, the most probable unrooted gene tree on \( \{A,B,C,D\} \) is identical to the unrooted species tree induced on \( \{A,B,C,D\} \).
Quartet tree estimation

Theorem (Allman et al.): Under the multi-species coalescent model, for any four taxa A, B, C, and D, the most probable unrooted gene tree on \{A,B,C,D\} is identical to the unrooted species tree induced on \{A,B,C,D\}. 

![Diagram of quartet tree estimation]

\[\theta_1 = 70\% \quad \theta_2 = 15\% \quad \theta_3 = 15\%\]
Species tree estimation from unrooted gene trees

Corollary: Under the multi-species coalescent model, the species tree is identifiable from the gene tree distribution.
Species tree estimation from unrooted gene trees

Corollary: Under the multi-species coalescent model, the species tree is identifiable from the gene tree distribution.

Proof: For every four species, select most frequently observed tree as the species tree. Then combine quartet trees!
• Optimization Problem (NP-Hard):

Find the species tree with the maximum number of induced quartet trees shared with the collection of input gene trees

\[ \text{Score}(T) = \sum_{t \in T} |Q(T) \cap Q(t)| \]

• **Theorem**: Statistically consistent under the multi-species coalescent model when solved exactly
ASTRAL

• Statistically consistent under the MSC, and runs in polynomial time

• Solves constrained version of the NP-hard Maximum Quartet Support problem using dynamic programming
  – Input: Gene trees and set $X$ of allowed bipartitions
  – Output: Species tree $T$ that maximizes the quartet support criterion, subject to drawing its bipartitions from the set $X$
ASTRAL on biological datasets

- 1KP: 103 plant species, 400-800 genes
- Yang, et al. 96 Caryophyllales species, 1122 genes
- Dentinger, et al. 39 mushroom species, 208 genes
- Giarla and Esselstyn. 19 Philippine shrew species, 1112 genes
- Laumer, et al. 40 flatworm species, 516 genes
- Grover, et al. 8 cotton species, 52 genes
- Hosner, Braun, and Kimball. 28 quail species, 11 genes
- Simmons and Gatesy. 47 angiosperm species, 310 genes
- Prum et al. 198 avian species, 259 genes
Tree accuracy when varying the number of species

![Graph showing species tree topological error (FN) vs. number of species. The graph compares ASTRAL-II and MP-EST, with error rates increasing as the number of species increases. The x-axis represents the number of species ranging from 10 to 1000, and the y-axis represents the species tree topological error (FN) ranging from 0% to 16%. The graph includes error bars indicating variability.]

1000 genes, “medium” levels of recent ILS
ASTRAL – great, but...

- The good: ASTRAL is
  - increasingly used in practice
  - Statistically consistent given true gene trees
  - sometimes more accurate than concatenation, but impacted by gene tree estimation error
  - very fast for many datasets (faster than concatenation)

- The bad: ASTRAL can fail to complete on large enough datasets within reasonable time frames (days of computation)
The alternatives are worse

• Concatenation Analyses (e.g., using RAxML):
  – most commonly used method, not statistically consistent, sometimes more accurate than summary methods
  – computationally intensive (e.g., 250 CPU years for the Avian Phylogenomics project with only 48 species) and do not scale to large numbers of species

• Co-estimation of gene trees and species trees: too expensive

• Other statistically consistent methods: not as accurate as ASTRAL
Decompose species set into pairwise disjoint subsets.

Full species set

Build a tree on each subset using "Disjoint Tree Merger" method

Tree on full species set

Auxiliary Info (e.g., distance matrix)

Compute tree on entire set of species using "Disjoint Tree Merger" method
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](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
Impact of using TreeMerge with ASTRAL-III on 1000 species and 1000 genes

![Impact of using TreeMerge with ASTRAL-III on 1000 species and 1000 genes](image)
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
  – Statistically consistent under MSC+GTR when used with ASTRAL (or other statistically consistent base method)
Guide Tree Merger

- Smirnov and Warnow, RECOMB-Comparative Genomics
- Guide Tree Merger (GTM): Another Disjoint Tree Merger method, but does not allow blending
- 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
- connect subset trees by adding edges, so as to minimize distance to the given guide tree (polynomial time!)
ASTRAL+GTM: better than ASTRAL!
Summary

• The best tree estimation methods are computationally intensive, and tree-space grows exponentially
Summary

• The best tree estimation methods are computationally intensive, and tree-space grows exponentially

• Statistical consistency is important but not sufficient

• Divide-and-conquer improves scalability, maintains statistical consistency, and can maintain accuracy (or only lose a small amount)

• Divide-and-conquer is highly parallelizable
Summary

- The best tree estimation methods are computationally intensive, and tree-space grows exponentially.
- Statistical consistency is important but not sufficient.
- Parallel implementations of expensive methods are helpful but not enough.
- Divide-and-conquer improves scalability, maintains statistical consistency, and can maintain accuracy (or only lose a small amount).

- Divide-and-conquer is highly parallelizable.
Summary

• The best tree estimation methods are computationally intensive, and tree-space grows exponentially.
• Statistical consistency is important but not sufficient.
• Parallel implementations of expensive methods are helpful but not enough.
• **Divide-and-conquer improves scalability, maintains statistical consistency, and can maintain accuracy (or only lose a small amount).**
Summary

• The best tree estimation methods are computationally intensive, and tree-space grows exponentially
• Statistical consistency is important but not sufficient
• Parallel implementations of expensive methods are helpful but not enough
• Divide-and-conquer improves scalability, maintains statistical consistency, and can maintain accuracy (or only lose a small amount)
• Divide-and-conquer is highly parallelizable
Acknowledgments

Papers available at http://tandy.cs.illinois.edu/papers.html
Presentations available at http://tandy.cs.illinois.edu/talks.html

Funding: NSF (CCF 1535977 and also NSF Graduate Fellowship to Erin Molloy)
Supercomputers: Blue Waters and Campus Cluster, both supported by NCSA
Potential PhD projects

• Graph-theoretic and discrete algorithms
• Heuristics for NP-hard optimization problems
• Machine learning (mostly unsupervised or semi-supervised)
• Applications to microbiome studies, biological evolution, etc.
• High-performance computing
Necessary Skills

• Background in algorithm design and analysis
• Ability to program in Python, Java, C, etc.
• Ability to look at data
• Interest in having an impact on science
• Collaboration and communication skills
CS 581: Algorithmic Genomic Biology

- Topics: statistical and computational phylogenomics, statistical estimation of phylogenetic trees and multiple sequence alignments
- I teach this every year, and typically students do research projects for their course projects.
- Several of these projects are then published in journals!
- No biology background needed (seriously!)
- Next offering: Spring 2020

I am looking for new students – please contact me
Accuracy in the presence of HGT + ILS

200 Estimated Gene Trees

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\textsuperscript{8}, 1000 gene trees estimated from GTR simulated sequences using FastTree-2\textsuperscript{7}

\textsuperscript{7} Price, Dehal, Arkin 2015
\textsuperscript{8} Fletcher, Yang 2009

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