Entering the genomic era of the Tree of Life

Tandy Warnow
The University of Illinois
Species Tree

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

Phylogeny + genomics = genome-scale phylogeny estimation
Opening remarks: this is an exciting time!

• My goal: Phylogenomics for all!
  • without a supercomputer or large compute farm!

• I predict tremendous advances over the next 3-5 years for species tree estimation in the presence of GDL+ILS+HGT:
  • theoretical foundations (what is consistent, new statistical models)
  • development of new methods (with strong empirical and theoretical performance)
  • advances in scalable methods (thousands of species and tens of thousands of genes)
Phylogeny Estimation

AGGGCAT  TAGCCCA  TAGACTT  TGCACAA  TGCAGCTT

U  V  W  X  Y

Diagram showing the phylogenetic relationships among different sequences.
Phylogeny estimation: CS and Statistics

- Assume DNA sequences are generated on an unknown model tree, and try to infer the tree from the observed sequences seen at the leaves.

NP-hard optimization problems
Large datasets
Years of CPU time for standard methods

This research combines mathematics, statistics, and computer science:
  - Algorithm design, proofs, implementation, simulations and testing
Identifiability, Consistency, and Sample Complexity

Assume DNA sequences are generated on an unknown model tree, and try to infer the tree from the observed sequences seen at the leaves

• Is the model tree identifiable?
• Is a given method statistically consistent?
• How much data does the method need to be accurate with high probability? (aka “sample complexity”)


Is method M statistically consistent under model G?

Question answered by mathematical proof

Error in species tree inferred by method M

Amount of data generated under model G and then given to method M as input
Genome-scale data?

error

Length 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)
This talk

• Species tree estimation when ILS is the only “problem”
  • Proof of identifiability and statistically consistent methods (and those which aren’t)
  • What works in practice – and how to improve estimation on large datasets

• Species tree estimation when GDL is present
  • Tremendous advances in the last two years

• What next?
Gene tree discordance

Multiple causes for discord, including
- Incomplete Lineage Sorting (ILS),
- Gene Duplication and Loss (GDL), and
- Horizontal Gene Transfer (HGT)
MSC+GTR Hierarchical Model

1. Gene trees evolve within the species tree (under the Multi-Species Coalescent model)
2. Sequences evolve down the gene trees (under GTR model)
1KP: Thousand Transcriptome Project

- 2014 *PNAS* study: 103 plant transcriptomes, 400-800 single copy “genes”
- 2019 *Nature* study: much larger!

**Major Challenges:**
- Multi-copy genes omitted (9500 -> 400)
- Massive gene tree heterogeneity consistent with ILS
Avian Phylogenomics Project

Erich Jarvis, HHMI
MTP Gilbert, Copenhagen
Guojie Zhang, BGI
Siavash Mirarab, Texas
Tandy Warnow, Texas and UIUC

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

Major challenge:
Multi-copy genes omitted
Massive gene tree heterogeneity consistent with ILS.
Four major approaches

- **Concatenation using Maximum likelihood (CA-ML):** NP-hard, and not consistent in the presence of ILS
- **Summary methods (e.g., ASTRAL, ASTRID, MP-EST):** consistent, and much faster than CA-ML, but impacted by gene tree estimation error
- **Site-based methods (e.g., SVDquartets, SVDquest):** consistent but not very computationally efficient
- **Co-estimation methods (StarBEAST and StarBEAST2):** consistent, very accurate, but very expensive
Four major approaches

- **Concatenation using Maximum likelihood (CA-ML):** NP-hard, and not consistent in the presence of ILS

- **Summary methods (e.g., ASTRAL, ASTRID, MP-EST):** consistent, and much faster than CA-ML, but impacted by gene tree estimation error

- **Site-based methods (e.g., SVDquartets, SVDquest):** consistent but not very computationally efficient

- **Co-estimation methods (StarBEAST and StarBEAST2):** consistent, very accurate, but very expensive
Traditional approach: concatenation

- Statistically **inconsistent** and can even be positively misleading (proved for unpartitioned maximum likelihood) [Roch and Steel, Theo. Pop. Gen., 2014]

Four major approaches

- **Concatenation using Maximum likelihood** (CA-ML): NP-hard, and not consistent in the presence of ILS
- **Summary methods** (e.g., ASTRAL, ASTRID, MP-EST): consistent, and much faster than CA-ML, but impacted by gene tree estimation error
- **Site-based methods** (e.g., SVDquartets, SVDquest): consistent but not very computationally efficient
- **Co-estimation methods** (StarBEAST and StarBEAST2): consistent, very accurate, but very expensive
Main competing approaches

Analyze separately

Summary Method

Concatenation

gene 1  gene 2 …  gene k

Species
MSC+GTR Hierarchical Model

1. Gene trees evolve within the species tree (e.g., under the Multi-Species Coalescent model)
2. Sequences evolve down the gene trees (e.g., under GTR model, possibly with indels)
Summary Method Protocol

Step 1: infer gene trees (traditional methods)

Step 2: infer species trees

1. Given gene sequence alignments, compute gene trees
2. Given gene trees, combine into species tree

Faster than concatenation, and can be parallelized
ASTRAL

[Mirarab, et al., ECCB/Bioinformatics, 2014]

• Optimization Problem (NP-Hard):

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

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

• **Theorem**: Statistically consistent under the multi-species coalescent model when solved exactly
ASTRAL
[Mirarab, et al., ECCB/Bioinformatics, 2014]

- 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

Theorem: Under MSC, most probable quartet tree is the species tree

Set of quartet trees induced by \( T \)

all input gene trees

a gene tree
• Optimization Problem (NP-Hard):

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

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

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

ASTRAL uses dynamic programming to solve a constrained version of this problem, and is provably statistically consistent
ASTRAL on biological datasets

- 1KP: 103 plant species, 400-800 genes
- Yang, et al. 96 Caryophyllales species, 1122 genes
- Dentingter, et al. 39 mushroom species, 208 genes
- Giarla and Esselstyn. 19 Philippine shrew species, 1112 genes
- Launer, 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
ASTRAL – pros and cons

• The good: ASTRAL is
  • Very popular statistically consistent method for species tree estimation
  • Very fast for many datasets (much faster than concatenation)

• The mixed:
  • Concatenation can be more accurate under some conditions

• The bad:
  • ASTRAL can fail to complete on large enough datasets within reasonable time frames (days of computation)
Decompose species set into pairwise disjoint subsets.

Full species set

Build a tree on each subset

Tree on full species set

Compute tree on entire set of species using “Disjoint Tree Merger” method

Note: use most accurate method on subsets, and treat as absolute constraints

Auxiliary Info (e.g., distance matrix)
Divide-and-Conquer Species Tree Estimation

Decompose species set into \textit{pairwise disjoint} subsets.

Build a tree on each subset.

Compute tree on entire set of species using “Disjoint Tree Merger” method.

Note: use most accurate method on subsets, and treat as \textit{absolute constraints}.

Auxiliary Info (e.g., distance matrix)

Tree on full species set

Guide Tree Merger

ASTRAL
GTM-ASTRAL vs ASTRAL

GTM-pipeline vs ASTRAL (seconds) on 1000 species (after gene trees estimated):

- 10 genes: 98 vs 8,656
- 25 genes: 175 vs 5,539
- 1000 genes: 7,949 vs 153,046
Summary of DTM pipeline results

• Divide-and-conquer pipelines using DTMs maintain statistical consistency and run in polynomial time
• Used with ASTRAL or RAxML, DTM pipelines reduce running time and memory usage, and produce species trees of comparable or better accuracy
• Runtimes are generally only slightly more expensive than the starting tree
• Embarassingly parallel
Impact of gene tree estimation error

• Basic questions:
  • How do methods perform when there is low signal per gene?
  • Can statistical consistency still be guaranteed?
Statistical consistency (two variants)

• Type 1: the number of loci and number of sites per gene both increase.

• Type 2: the number of loci increases, but the number of sites per gene is bounded by a fixed value.

Roch, Nute, and Warnow (Syst Biol 2019)

• Summary methods and fully partitioned ML can be positively misleading in the Type 2 setting.

• Proof is basically an elaboration on the Felsenstein Zone argument.
Impact of Gene Tree Estimation Error
(from Molloy and Warnow 2017)

Summary Methods

- Site-based Method
  - Error is fraction of bipartitions that are not recovered

- GTEE: gene tree estimation error
- CA-ML: concatenation with ML
- 26 species, 1000 genes, with high ILS (anomaly zone)

Note: Summary methods better than CA-ML for low GTEE, then worse!
ILS-based species tree estimation (Summary)

• Last few years have seen a revolution in methods and theory
• Several methods have excellent accuracy – but choice of method depends on data
• Statistical consistency is complicated
• Empirical accuracy (biological and simulated data) also complicated:
• Scaling to large datasets an issue
This talk

• Species tree estimation when ILS is the only "problem"
  • Proof of identifiability and statistically consistent methods (and those which aren’t)
  • What works in practice – and how to improve estimation on large datasets

• Species tree estimation when GDL is present
  • Tremendous advances in the last two years

• What next?
1KP: Thousand Transcriptome Project

- 2014 *PNAS* study: 103 plant transcriptomes, 400-800 single copy “genes”
- 2019 *Nature* study: much larger!

**Major Challenges:**
- Multi-copy genes omitted (9500 -> 400)
- Massive gene tree heterogeneity consistent with ILS
The species tree has one duplication (at the root), which produces a gene family tree that has two copies of the species tree!

Multi-copy trees: MUL-trees
Problem: Given set of MUL-trees, infer the species tree

(a) Species tree $T^*$

(b) Gene tree $M_1$ with one duplication.

(c) Gene tree $M_2$ with one duplication and two losses.

(d) Gene tree with one duplication and three losses.

Note: no orthology detection
Species tree estimation under GDL

• Concatenation: requires restriction to single-copy genes (throws out data) OR knowledge of orthology (not reliable)
• Summary methods: gene tree parsimony (e.g., DupTree and iGTP) and supertree methods adapted to MUL-trees (e.g., MulRF and FastMulRFS), also guenomo (Bayesian method).
• Bayesian co-estimation of gene trees and species trees (e.g., PhylDog)
Species tree estimation under GDL

• Concatenation: requires restriction to single-copy genes (throws out data) OR knowledge of orthology (not reliable)

• Summary methods: gene tree parsimony (e.g., DupTree and iGTP) and supertree methods adapted to MUL-trees (e.g., MulRF and FastMulRFS), also guenomo (Bayesian method).

• Bayesian co-estimation of gene trees and species trees (e.g., PhylDog)

Note:

Nothing proven to be statistically consistent under GDL... until 2019
Theorem (Legried, Molloy, Warnow, and Roch, 2019): **ASTRAL-multi** is statistically consistent under GDL and runs in polynomial time.

Theorem (Molloy and Warnow, 2019): **FastMulRFS** is statistically consistent under a generic duplication-only or loss-only model, and runs in polynomial time.

Note: Both methods use dynamic programming to solve NP-hard discrete optimization problems within constrained search space in polynomial time.

Theorem: Under GDL, most probable quartet tree is the species tree.
But...ASTRAL-multi is not as accurate as other methods!

Results on 100-species datasets with moderate GDL, moderately high ILS, and high GTEE
ASTRAL-Pro: Quartet-Based Species-Tree Inference despite Paralogy

Chao Zhang, Celine Scornavacca, Erin K Molloy, Siavash Mirarab

*Molecular Biology and Evolution,* Volume 37, Issue 11, November 2020, Pages 3292–3307, [https://doi.org/10.1093/molbev/msaa139](https://doi.org/10.1093/molbev/msaa139)

*Published:* 04 September 2020
ASTRAL-Pro was designed to specifically address GDL:
1. Root and “tag” each gene tree (nodes are identified as duplications or speciations)
2. Modify weighting on quartet trees to reflect speciation
Accuracy: no important differences on these large datasets.

Speed: ASTRID-multi fastest, but ASTRAL-Pro nearly same. FastMuRFCS slowest (max about 7 hours).

1000 species and 1000 gene trees estimated from 100bp alignments (approx. 44% mean gene tree error), AD=20%, duplication rate of $5.0 \times 10^{-10}$, and loss/dup = 1.

Summary for GDL

• ASTRAL-Pro and ASTRID-multi
  • Very accurate (more accurate than ASTRAL-multi) – both hard to beat!
  • Can run on very large datasets (1000 species, 1000 genes)
  • Does not require orthology detection
  • Guaranteed polynomial time

• But
  • ASTRAL-Pro can be slow on large datasets with high rates of GDL, and proof of consistency depends on correct “rooting and tagging”
  • ASTRID-multi: no proof of consistency under any GDL model

• Note:
  • Same tricks used for ASTRAL (e.g., GTM and TreeMerge, FASTRAL) can be used on ASTRAL-Pro to improve scalability
This talk

• Species tree estimation when ILS is the only "problem"
  • Proof of identifiability and statistically consistent methods (and those which aren’t)
  • What works in practice – and how to improve estimation on large datasets

• Species tree estimation when GDL is present
  • Tremendous advances in the last two years

• What next?
What about HGT?

• HGT also makes heterogeneous gene trees
• Under some assumptions of random HGT operating, it may be possible to define the “underlying” species tree.
• Can the underlying species tree be inferred from the gene trees, or via concatenation?

• Statistical consistency of quartet-based methods established by:
  • Roch & Snir, JCB 2013
  • Daskalakis & Roch, arXiv 2015
Accuracy in the presence of HGT + ILS

Theorem (Snir and Roch): Under bounded but random HGT, most probable quartet tree is the species tree.

Figure 2 Mean Robinson-Foulds error rate on datasets with 10 genes. We show mean RF error rates for summary methods applied to estimated gene trees as well as for an unpartitioned maximum likelihood concatenation analysis. Error bars indicate standard error; 50 replicates per dataset.

Davidson et al., RECOMB-CG, BMC Genomics 2015
Quartet tree robustness

• Recurring theme: the most probable quartet tree is the species tree for:
  • ILS (Multi-species coalescent)
  • ILS+GDL (DLCOAL)
  • Random HGT

• Hence, quartet-based species tree estimation (from gene trees) is often statistically consistent, if performed properly.

• Interestingly, it is also fairly robust
Scaling to large datasets

• Scaling to large datasets an issue – affects accuracy and runtime/memory usage
  • Number of species (affects all methods)
  • Number of loci (affects *BEAST and ASTRAL)
  • Amount of gene tree heterogeneity (affects runtime for ASTRAL)

• Techniques
  • GTM/TreeMerge for scaling ASTRAL and RAxML to large numbers of species
  • FASTRAL (not shown): scales ASTRAL to large numbers of loci
  • BBCA (not shown): scales *BEAST to large numbers of loci
Summary

• Excellent methods for different conditions!
  • ASTRAL-Pro: excellent for GDL+ILS
  • SpeciesRax: shows promise for GDL+HGT
  • Concatenation good for many conditions
  • Disjoint Tree Mergers: Enable scalability to large datasets

• Lots of theory needs to be established
  • Statistical consistency?
  • Sample complexity?
  • Robustness (missing data, gene tree estimation error)?
Closing remarks: this is an exciting time!

- I predict tremendous advances over the next 3-5 years for species tree estimation in the presence of GDL+ILS+HGT:
  - theoretical foundations (what is consistent, new statistical models)
  - development of new methods (with strong empirical and theoretical performance)
  - Advances in scalable methods (thousands of species and tens of thousands of genes)

- My goal:
  - Phylogenomics for all! (without a supercomputer or large compute farm!)
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

Write to me: warnow@Illinois.edu
Selection of papers

• Willson et al., ALCOB (2021) (simulation study for GDL)
• Morel et al. (2021). bioRxiv DOI: 10.1101/2021.03.29.437460 (SpeciesRax)
• Markin & Eulenstein, Bioinformatics (2021). (Consistency of ASTRAL-multi under DLCOAL)
• Zhang et al., MBE (2020) doi:10.1093/molbev/msaa139 (ASTRAL-Pro)
• Molloy and Warnow, Bioinf (2020) 36.Supplement_1: i57-i65 (FastMulRFS)