New methods for estimating species trees from genome-scale data

Tandy Warnow
The University of Illinois
Species Tree Estimation

From the Tree of Life Website,
University of Arizona
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.

Courtesy James Degnan
Traditional approach: concatenation

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

- Mixed accuracy in simulations
  [Kubatko and Degnan, Systematic Biology, 2007]
  [Mirarab, et al., Systematic Biology, 2014]
This talk

- **Statistical binning** (Science 2014) and **Weighted Statistical Binning** (PLOS One 2015): improving gene trees
- **ASTRAL** (Bioinformatics 2014, 2015) and **ASTRID** (BMC Genomics 2014): coalescent-based species tree estimation methods that have high accuracy on large datasets (1000 species and genes)
- **BBCA** (improving *BEAST* scalability to large numbers of loci)
- **DCM-boosting** (improving scalability to large numbers of species)
Results on 11-taxon datasets with weak ILS

*BEAST MORE ACCURATE than summary methods, because *BEAST gets more accurate gene trees!

*BEAST more accurate than summary methods (MP-EST, BUCKy, etc)
CA-ML (concatenated analysis) most accurate

Datasets from Chung and Ané, 2011
Bayzid & Warnow, Bioinformatics 2013
Results on 11-taxon datasets with weak ILS

*BEAST more accurate than summary methods (MP-EST, BUCKy, etc)
CA-ML (concatenated analysis) most accurate

Datasets from Chung and Ané, 2011
Bayzid & Warnow, Bioinformatics 2013
Results on 11-taxon datasets with weak ILS

*BEAST* more accurate than summary methods (MP-EST, BUCKy, etc)

CA-ML (concatenated analysis) most accurate

Concatenation (RAxML) best of all methods on these data!
(However, for high enough ILS, concatenation is not as accurate as the best summary methods.)

Datasets from Chung and Ané, 2011
Bayzid & Warnow, Bioinformatics 2013
Avian Phylogenomics Project

• Approx. 50 species, whole genomes, 14,000 loci
• Jarvis, Mirarab, et al., Science 2014

Major challenges:
• Concatenation analysis took > 250 CPU years, and suggested a rapid radiation
• We observed massive gene tree heterogeneity consistent with incomplete lineage sorting
• Very poor resolution in the 14,000 gene trees (average bootstrap support 25%)
• Standard coalescent-based species tree estimation methods contradicted concatenation analysis and prior studies
Avian Phylogenomics Project

E Jarvis, HHMI

MTP Gilbert, Copenhagen

G Zhang, BGI

T. Warnow UT-Austin

S. Mirarab UT-Austin

Md. S. Bayzid, UT-Austin

Plus many many other people...

- Approx. 50 species, whole genomes, 14,000 loci

Solution: **Statistical Binning**

- Improves coalescent-based species tree estimation by improving gene trees (Mirarab, Bayzid, Boussau, and Warnow, *Science* 2014), see also weighted statistical binning (Bayzid et al., PLOS One 2015)

- Avian species tree estimated using **Statistical Binning with MP-EST** (Jarvis, Mirarab, et al., *Science* 2014)
Ideas behind statistical binning

• “Gene tree” error tends to decrease with the number of sites in the alignment

• Concatenation (even if not statistically consistent) tends to be reasonably accurate when there is not too much gene tree heterogeneity
The statistical binning pipeline for estimating species trees from gene trees. Loci are grouped into bins based on a statistical test for combinability, before estimating gene trees.

Note: Supergene trees computed using fully partitioned maximum likelihood. Vertex-coloring graph with balanced color classes is NP-hard; we used heuristic.
Theorem 2 (PLOS One, Bayzid et al. 2015): WSB pipelines are statistically consistent under GTR+MSC

Easy proof:
As the number of sites per locus increase

• All estimated gene trees converge to the true gene tree and have bootstrap support that converges to 1 (Steel 2014)
• For every bin, with probability converging to 1, the genes in the bin have the same tree topology
• Fully partitioned GTR ML analysis of each bin converges to a tree with the common topology of the genes in the bin

Hence as the number of sites per locus and number of loci both increase, WSB followed by a statistically consistent summary method will converge in probability to the true species tree. Q.E.D.
Datasets: 11-taxon strongILS datasets with 50 genes from Chung and Ané, Systematic Biology
Binning produces bins with approximate 5 to 7 genes each
Comparing Binned and Un-binned MP-EST on the Avian Dataset

Unbinned MP-EST strongly rejects Columbea, a major finding by Jarvis, Mirarab, et al.

Binned MP-EST is largely consistent with the ML concatenation analysis.

The trees presented in Science 2014 were the ML concatenation and Binned MP-EST
1KP: Thousand Transcriptome Project

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

**Challenges:**
- Massive gene tree heterogeneity consistent with ILS
- Could not use MP-EST due to missing data (many gene trees could not be rooted) and large number of species
1KP: Thousand Transcriptome Project

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

Solution:
- New coalescent-based method ASTRAL (Mirarab et al., Bioinformatics 2015)
- ASTRAL is statistically consistent, polynomial time, and uses unrooted gene trees.
• Estimates the species tree from unrooted gene trees by finding the species tree that has **the maximum quartet support, subject to input constraint set.**

• Theorem: ASTRAL is **statistically consistent** under the MSC, even when solved in constrained mode (drawing bipartitions from the input gene trees).

• The constrained version of ASTRAL runs in polynomial time

• Open source software at [https://github.com/smirarab](https://github.com/smirarab)

• Published in Bioinformatics 2014 and 2015

• Used in Wickett, Mirarab et al. (PNAS 2014)
Simulation study

- Variable parameters:
  - Number of species: 10 – 1000
  - Number of genes: 50 – 1000
  - Amount of ILS: low, medium, high
  - Deep versus recent speciation
  - 11 model conditions (50 replicas each) with heterogenous gene tree error
  - Compare to NJst, MP-EST, concatenation (CA-ML)
  - Evaluate accuracy using FN rate: the percentage of branches in the true tree that are missing from the estimated tree

Used SimPhy, Mallo and Posada, 2015
Tree accuracy when varying the number of species

1000 genes, “medium” levels of recent ILS
Tree accuracy when varying the number of species

1000 genes, “medium” levels of recent ILS
Tree accuracy when varying the number of species

1000 genes, “medium” levels of recent ILS
Running time when varying the number of species

1000 genes, “medium” levels of recent ILS
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
• ASTRID: Accurate species trees using internode distances
• Re-implementation of NJst, using a different method for computing the tree given the distance matrix (FastME instead of NJ)
• Statistically consistent under the MSC
• Vachaspati and Warnow BMC Genomics 2015
• Available at
  http://www.github.com/pranjalv123/ASTRID
Both ASTRAL and ASTRID substantially outperform MP-EST.
ASTRID is very fast

On the ASTRAL-2 dataset with 1000 taxa, 1000 genes, ASTRID-FastME takes 33 minutes, ASTRAL takes 12 hours.
Scaling methods to large datasets

• BBCA: combining random binning with *BEAST to enable scalability to large numbers of loci (Zimmermann et al., BMC Genomics 2014)

• Using divide-and-conquer to scale MP-EST (and other methods) to large numbers of taxa (Bayzid et al., BMC Genomics 2014)
Summary

• ASTRAL and ASTRID: highly accurate summary methods, both statistically consistent under the MSC

• Weighted statistical binning: statistically consistent pipelines, and improve coalescent-based methods in the presence of gene tree estimation error

• Concatenation: mixed performance. Current best coalescent-based methods better than concatenation if high enough ILS, and otherwise concatenation reasonable.

• Future methods may be reliably better than concatenation!
Acknowledgments

NSF grant DBI-1461364 (joint with Noah Rosenberg at Stanford and Luay Nakhleh at Rice): http://tandy.cs.illinois.edu/PhylogenomicsProject.html
Papers available at http://tandy.cs.illinois.edu/papers.html

Software
- **ASTRAL and statistical binning**: Available at https://github.com/smirarab
- **ASTRID**: Available at http://pranjalv123.github.io/ASTRID/

**Other Funding**: David Bruton Jr. Centennial Professorship, TACC (Texas Advanced Computing Center), Grainger Foundation, and HHMI (to SM)