Species Tree Estimation

Tandy Warnow

February 26, 2017

Species Tree Estimation
Tandy Warnow

Species tree estimation

Multiple challenges:

1. NP-hard problems on big datasets
2. Heterogeneity (different trees for different parts of genome)

Supertree estimation addresses the first problem, but the second problem is more challenging!

Causes for gene tree discordance

- Incomplete lineage sorting
- Gene duplication and loss
- Horizontal gene transfer
- Hybridization

And, of course, gene tree estimation error

Standard approaches

Can we use standard approaches, such as concatenation using maximum likelihood (CA-ML) or supertree methods to estimate species trees?

- CA-ML not statistically consistent (Roch and Steel, 2015)
- Standard supertree and consensus methods not statistically consistent
- Most frequent gene tree not statistically consistent

Simulations show that relative performance between methods depends on number of genes, number of species, amount and source of gene tree heterogeneity, and gene tree estimation error (e.g., Mirarab et al., 2014).

Species tree estimation using summary methods

- Input: set $\mathcal{T}$ of trees on subsets of S, species set
- Output: tree T on full set S, optimizing some criterion

Some species tree methods for ILS

Methods that are (or might be) statistically consistent in the presence of ILS:

- ASTRAL and BUCKy-pop, quartet-based methods
- ASTRID, NJst, GLASS, and other distance-based methods
- MP-EST, maximum pseudo-likelihood triplet-based
- SVDquartets and other site-based methods (not yet proved consistent)
- BEST and *BEAST, Bayesian methods

Note: Proofs of consistency for summary methods assume true gene trees.

ASTRAL: summary method

ASTRAL (Mirarab et al. 2014, Mirarab and Warnow 2015) computes a weight for every quartet tree, and then seeks a species tree that has the maximum total weight.

The weight of a quartet tree is the number of gene trees that induce the quartet tree.

This is NP-hard, but the constrained version can be solved using dynamic programming (just like FastRFS) in polynomial time.

The ASTRAL algorithm is a non-parametric method that is statistically consistent under the Multi-Species Coalescent model.

BUCKy: Bayesian summary method

The population tree in BUCKy (Larget et al. 2010) computes a quartet tree for every four species, and then seeks a tree that satisfies as many quartets as possible.

The calculation of quartet trees is based on a Bayesian distribution of gene trees for every gene.

SVDquartets: site-based tree inference

SVDquartets (Chifman and Kubatko 2014 and 2015) uses the singular-valued decomposition (SVD) to compute a tree on every quartet, and then seeks a tree that satisfies as many quartets as possible.

The calculation of quartet trees is based on the site patterns, not on estimated gene trees.

PAUP* (popular software package) provides this software.

Quartet amalgamation methods

Note the repeated use of quartet amalgamation methods, which construct a tree from a set of (possibly weighted) quartet trees.

NP-hard optimization problem.

Lots of established theory.

PTAS by Jiang et al. 2001

Distance-based methods

- ASTRID (Vachaspati and Warnow, 2015) and NJst (Liu and Yu, 2011): summary methods that compute the "average internode distance matrix" and run standard distance-based methods (FastME or neighbor joining) on the matrix.
- GLASS (Mossell and Roch, 2011) and related methods: summary methods that compute matrix of minimum pairwise distances, and run standard distance-based methods.
- METAL (Dasarathy et al., 2015): site-based method, computes the Jukes-Cantor distance matrix on the concatenated alignment and runs neighbor joining.

Bayesian methods

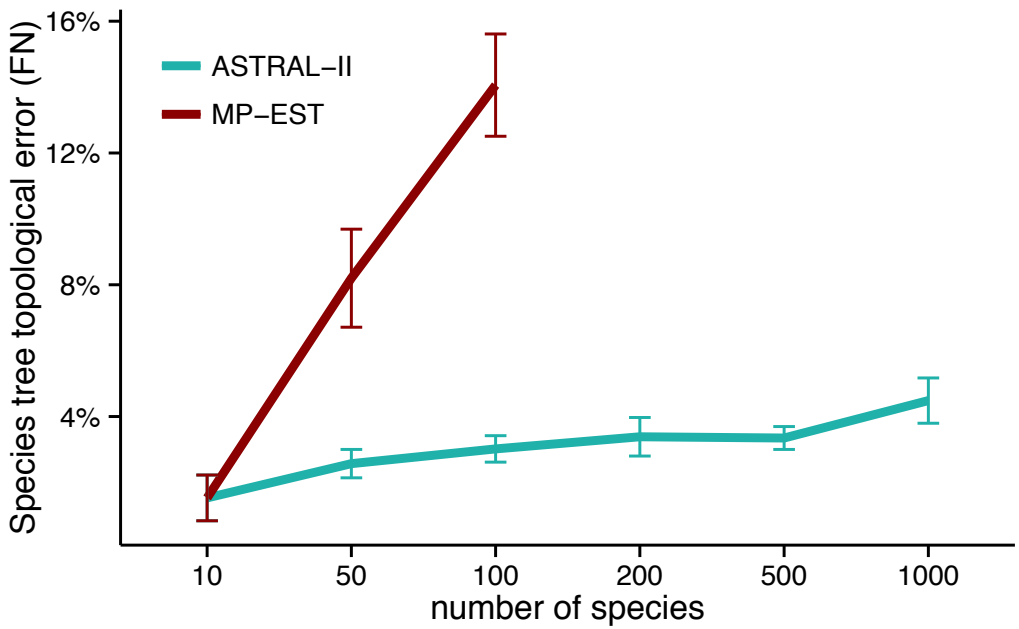
*BEAST (Heled and Drummond 2010) is a Bayesian method to co-estimate the species tree and gene trees from the input sequence alignments.

*BEAST is very computationally intensive, and can take weaks or months to analyze datasets with just 25 species and 50 genes.

See BBCA (Zimmermann et al. 2014), a technique to improve the convergence of *BEAST by random binning of genes.

ASTRAL vs. MP-EST

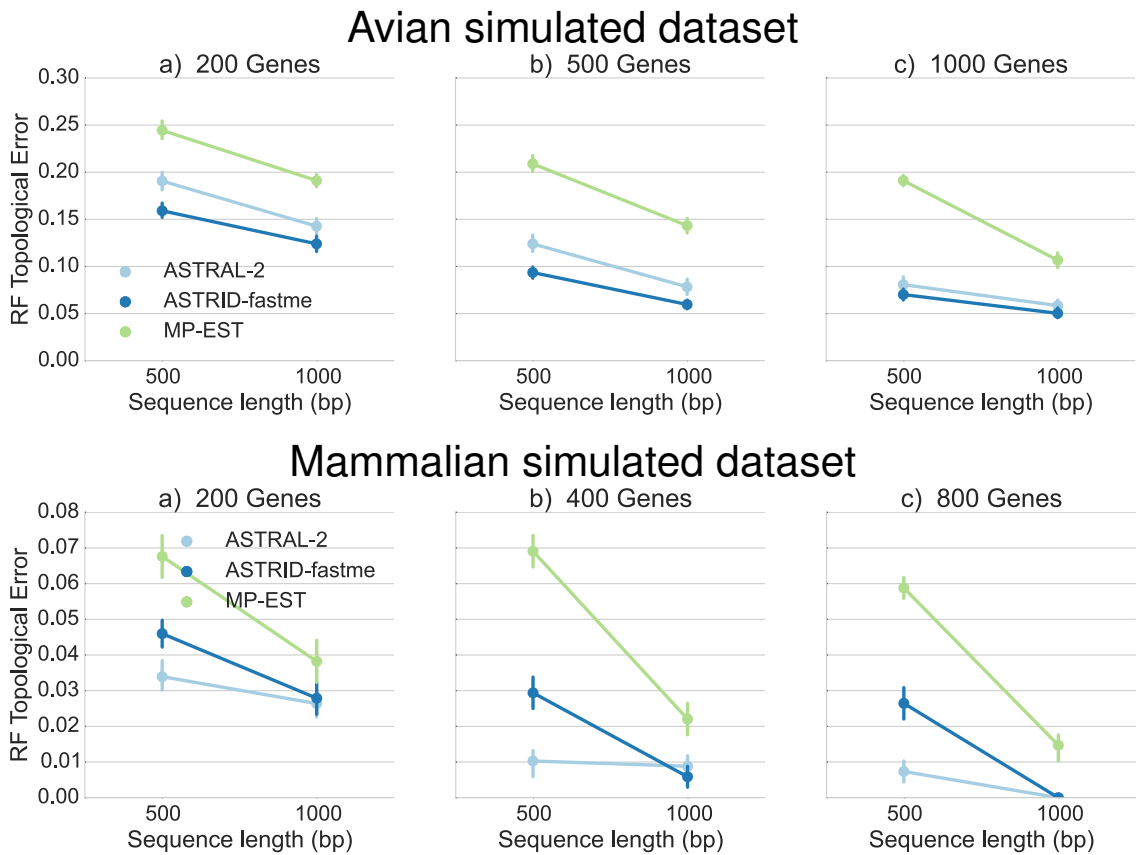
Tree accuracy when varying the number of species



1000 genes, “medium” levels of recent ILS

ASTRAL vs. ASTRID

Both ASTRAL and ASTRID substantially outperform MP-EST



Performance of species tree methods on supertree datasets

Method	500	500	500	500
Scaffold %	20	50	75	100
# Replicates	8	10	10	10
ASTRAL	15.3	14.8	12.7	11.2
ASTRAL-enhanced	14.8	14.1	12.6	11.2
ASTRID	26.0	50.1	45.4	**10.5**
MRL	15.4	14.3	12.1	11.2
MulRF	46.9	40.3	27.4	12.6
PluMiST	35.4	29.5	22.4	10.9
FastRFS-basic	14.5	14.3	12.4	11.1
FastRFS-enhanced	**14.3**	**13.9**	**12.0**	10.8

Table : Average supertree topology estimation error on simulated datasets.

Summary

When genes can differ due to ILS:

- Many methods are known to be statistically consistent for species tree estimation under the multi-species coalescent model.
- Relative performance varies and is impacted by (a) amount of ILS, (b) number of taxa, and (c) gene tree estimation error.
- Species tree methods can be used as supertree methods, and vice-versa.
- Limited theory yet established about guarantees in the presence of missing data and/or gene tree estimation error.

Other causes of gene tree discord

Other biological processes cause gene tree discord, such as

- Gene duplication and loss
- Horizontal gene transfer
- Hybridization

Accurate histories can require networks, and in general require ability to handle multiple sources of discord.

Possible research projects

1. Examine impact of alignment error on site-based methods
2. Improve quartet amalgamation methods
3. Develop simple PTAS for maximum quartet satisfiability
4. Establish consistency or inconsistency of various methods
5. Evaluate species tree estimation methods when number of loci is small
6. Develop and test mehtods for estimating trees from incomplete distance matrices
7. Find biological dataset and analyze using several methods;

See Projects chapter from textbook for more.