Supertree Estimation

Tandy Warnow

September 8, 2022

Supertree Estimation
Tandy Warnow

Today's material:

- Purpose of supertree methods
- Review of Tree Compatibility (Chapter 3)
- Intro to supertree methods (Chapter 7)

Supertree estimation

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

Notes:

1. check whether the input trees are rooted or unrooted (different problems).
2. the input $\mathcal{T}$ is called a *profile*

Uses of supertree estimation

1. Traditional: Combining trees computed by different researchers, on different groups of species
2. Also: Divide-and-conquer strategies

Tree Compatibility

A set $\mathcal{T}$ of trees is said to be **compatible** if there is a supertree T that is compatible with each tree in $\mathcal{T}$.
If so, then T is called a *compatibility supertree* for $\mathcal{T}$.

Tree Compatibility problem: are the trees in $\mathcal{T}$ compatible?

Question to class: what algorithms do you know for solving tree compatibility?

For each algorithm, what assumptions are made about the input?

Tree Compatibility

A set $\mathcal{T}$ of trees is said to be **compatible** if there is a supertree that induces each tree in $\mathcal{T}$.

Tree Compatibility problem: are the trees in $\mathcal{T}$ compatible?

- If all the trees are rooted, then the problem can be solved using Aho, Sagiv, Szymanski, and Ullman (Section 3.3 from textbook) in polynomial time.
- If the trees are unrooted, the problem is NP-hard (Section 3.5 from textbook).

Because unrooted tree compatibility is NP-hard, it is trivial to show that most optimization problems for supertree construction from unrooted source trees are NP-hard.

But they are still hard for supertree construction from rooted source trees!

Matrix Representation with Parsimony (MRP)

MRP: Represent every tree in $\mathcal{T}$ with a $0,1$-matrix (one column for every edge), concatenate the matrices, and then solve maximum parsimony

This is the most well known supertree approach among biologists.

Question: suppose the input trees are compatible. What does MRP return?

Other supertree optimization problems

- Robinson-Foulds Supertree: Find binary T minimizing the Robinson-Foulds distance to the trees in $\mathcal{T}$
- Maximum Quartet Support Supertree: Find binary T minimizing the quartet distance to the trees in $\mathcal{T}$
- Matrix Representation with Likelihood (MRL): Same matrix, but then solve CFN maximum likelihood
- Distance-based supertrees: Compute a matrix M of average leaf-to-leaf distances between species, and find an additive matrix close to M (minimizing some criterion)

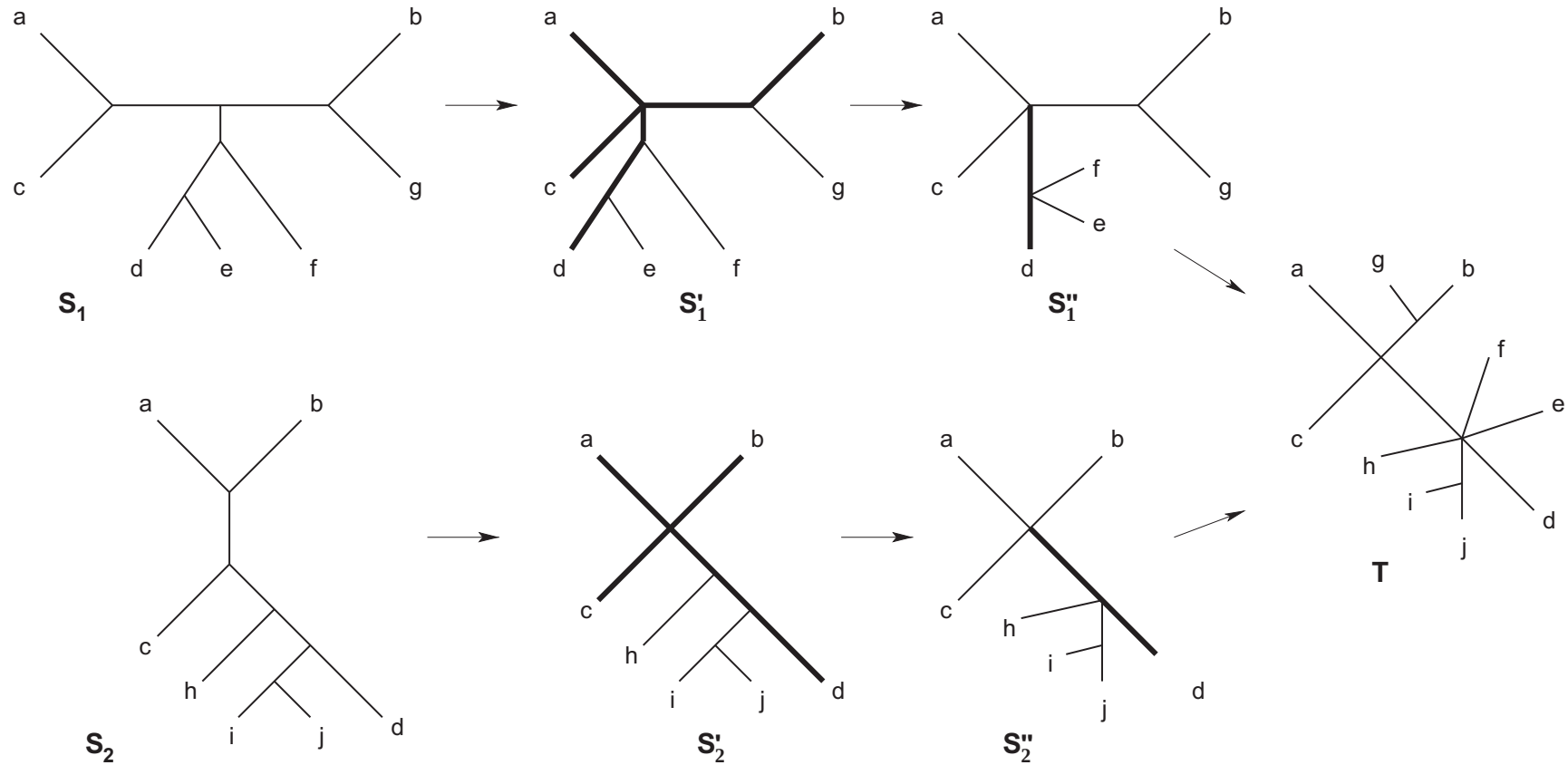
All these problems are NP-hard, and some of these optimization problems create very large inputs.

SuperFine: a supertree method "booster"

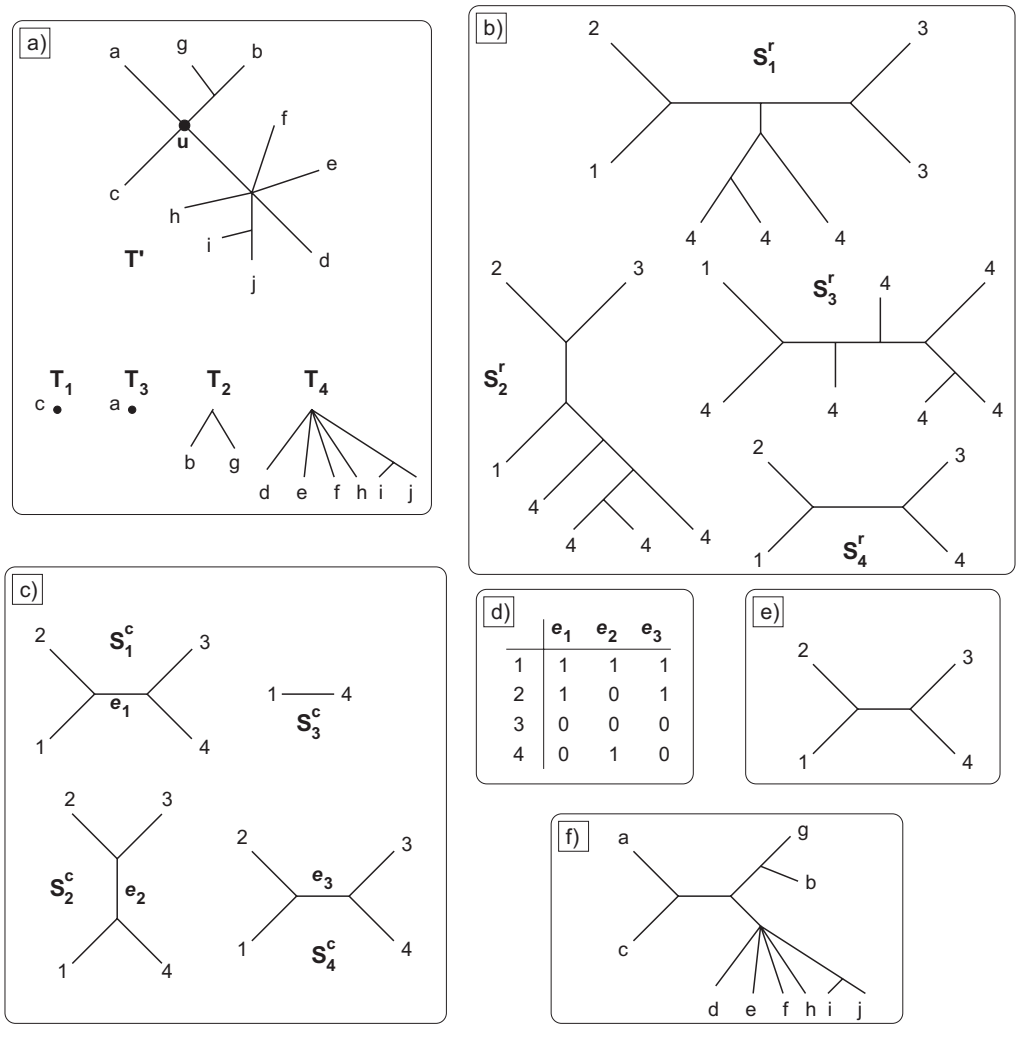
SuperFine (Swenson et al. 2012) is a meta-method for improving the speed and accuracy of supertree methods:

- Compute a constraint tree using the Strict Consensus Merger
- For each polytomy v (node of degree $d > 3$):
  - Compute an encoding of the source trees into trees with at most d leaves
  - Run preferred supertree method on the new source trees, obtaining $t(v)$
  - Refine polytomy v with the computed supertree $t(v)$

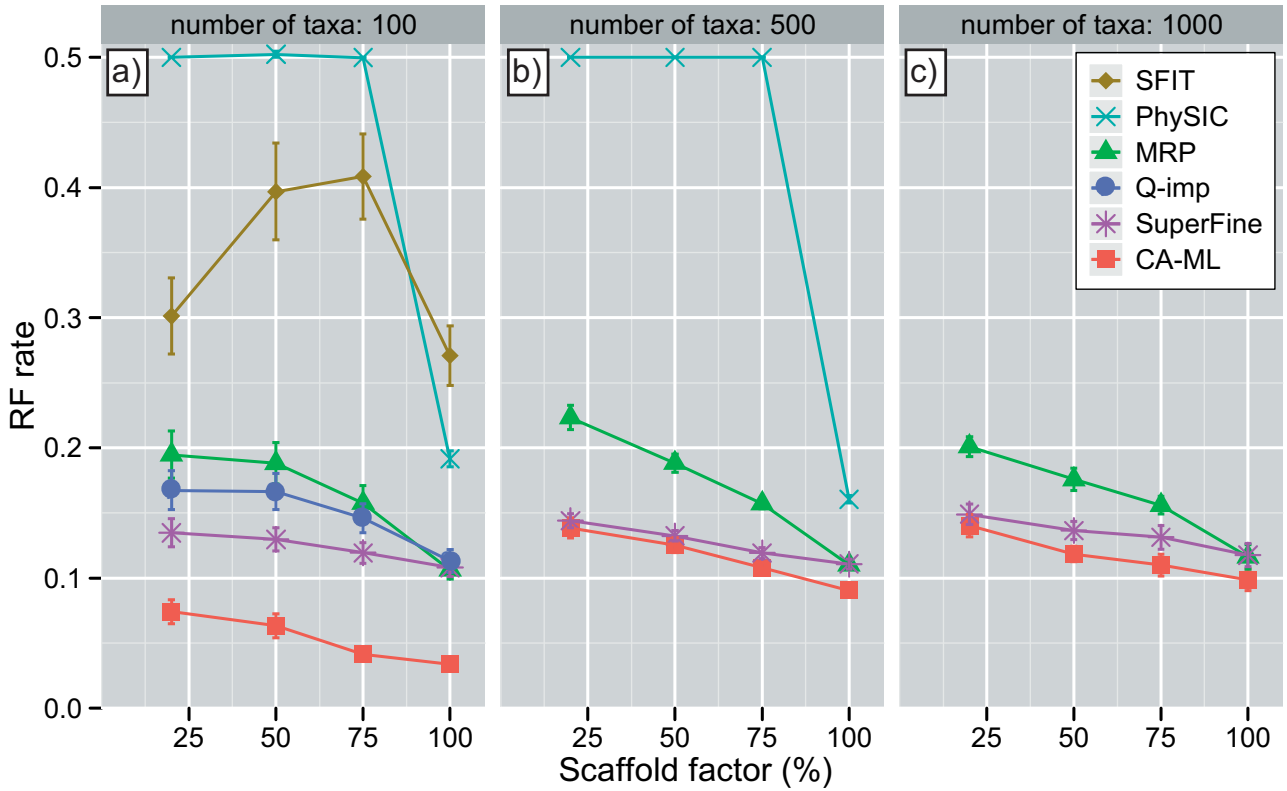
Phase 1: Strict Consensus Merger (SCM)



Phase 2: Refining the SCM tree using MRP



SuperFine+MRP vs MRP, CA-ML, and other methods



Robinson-Foulds Supertrees

Finding the supertree T that minimizes the Robinson-Foulds (RF) distance to the source trees is the Robinson-Foulds Supertree problem.

MulRF (Chaudhary et al., 2014) and PluMiST (Kupczok, 2011) are two methods for Robinson-Foulds Supertrees.

Robinson-Foulds Supertrees are (sort of) approximations to the Maximum Likelihood Supertree (Steel and Rodrigo, 2008) problem (see Bryant and Steel 2009 for more).

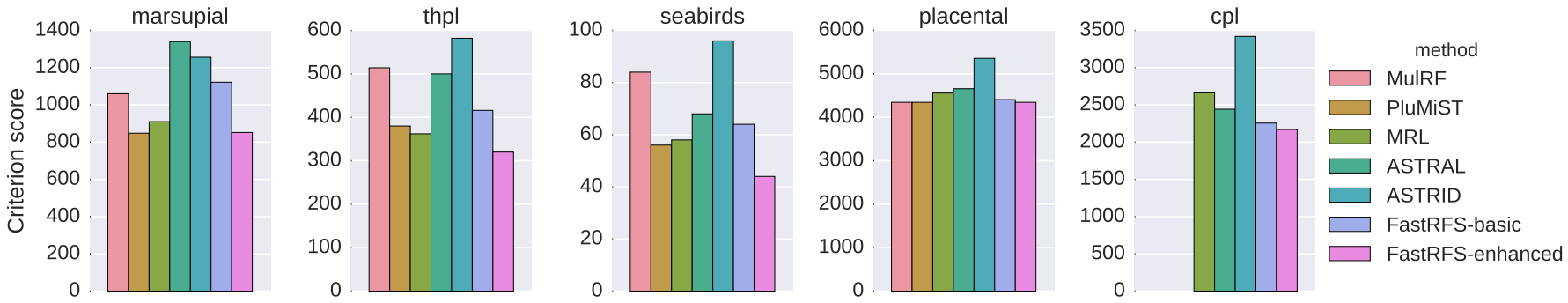
Constrained optimization

Bipartition-constrained Robinson-Foulds Supertree Problem:

- Input: Set $\mathcal{T}$ of source trees and set X of bipartitions on S
- Output: Tree T on S that draws its bipartition set $C(T)$ from X, and that minimizes the RF distance to $\mathcal{T}$ among all such supertrees

FastRFS (Vachaspati and Warnow, Bioinformatics 2016) solves this problem in polynomial time, using dynamic programming.

FastRFS criterion scores



The CPL dataset has 2228 species, and so represents the largest and most difficult dataset; MulRF and PluMiST could not complete on it.

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 about supertree methods

- Supertree methods are important techniques for large-scale phylogeny estimation
- New approaches provide good accuracy but are limited to relatively small datasets.
- SuperFine can help, but only when the constraint tree computed by the Strict Consensus Merger is not too unresolved.
- Potential directions: distance-based supertree methods, constrained approaches, others?
- Or perhaps just do something else to enable divide-and-conquer?

Disjoint Tree Merger (DTM) methods

- Input: set $\mathcal{T}$ of leaf-disjoint trees on subsets of S, and auxiliary information (e.g., distance matrix D or estimated tree T on S)
- Output: Tree $\hat{T}$ on S such that $\hat{T}|_{L(t)} = t$ for all $t \in \mathcal{T}$, using the auxiliary information (somehow)

Note: some DTM methods allow blending, and some do not.

DTM methods

Examples:

- NJmerge (Molloy and Warnow): uses distance matrix, and modifies Neighbor Joining (Saitou and Nei) to obey constraints
- TreeMerge (Molloy and Warnow): addresses weaknesses in NJMerge
- Guide Tree Merger (Smirnov and Warnow): uses estimated tree, does not allow blending, finds tree minimizing RF distance to estimated tree
- Incremental Tree Building (Zhang et al.): uses quartet trees computed from distances, absolute fast converging

All are outstanding for species tree estimation (improve runtime and/or accuracy). Results for maximum likelihood gene tree estimation also potentially strong!

Summary

- Supertree methods synthesize information from existing trees, but are based on NP-hard optimization problems
- Disjoint Tree Mergers: a new type of supertree method (but not really the same)
- Divide-and-conquer approaches can use supertree methods or disjoint tree methods to advantage
- Lots of interesting research still to do