Tree-of-life reconstruction using ASTRAL: complexity, support, and parallelism

Siavash Mirarab
University of California, San Diego (ECE)

in collaboration with
Warnow lab (UIUC) & Sebastian Roch (UW-Madison)
Phylogeny

Leaves are Species (a.k.a. **taxa**)

**Tree topology:** The branching structure, showing evolutionary relationships

**Branch length and width:** can be related to time between speciation events and the size of the populations, but we will draw them arbitrarily
Statistical inference of phylogenies
Statistical inference of phylogenies

Gorilla: ACTGCACACCG
Human: ACTGCCCCCG
Chimpanzee: AATGCCCCCG
Orangutan: CTGCACACCGG

\[ D \]
\[ P(D|T) \]
\[ T \]
Applications

source: Scaduto et al., PNAS, 2010


source: Gire et al., Science, 2014
Applications

Nothing in biology makes sense except in the light of **evolution** (Dobzhinsky)

Nothing in the evolution makes sense except in the light of the **phylogeny**

Sequence data growth

- Rapid growth in the number of sequences

Sequence data growth

- Rapid growth in the number of sequences
- Our focus has shifted to “whole genomes”

Phylogenomics

More data (#genes)

Inference error

More data (#genes)
Gene tree discordance
Gene tree discordance

The species tree

A gene tree
 Causes of gene tree discordance include:

  • **Incomplete Lineage Sorting (ILS)**
  • Duplication and loss
  • Horizontal Gene Transfer (HGT)
Incomplete Lineage Sorting (ILS)

- A random process related to the coalescence of alleles across various populations.
Incomplete Lineage Sorting (ILS)

- A random process related to the coalescence of alleles across various populations

Tracing alleles through generations
Incomplete Lineage Sorting (ILS)

- A random process related to the coalescence of alleles across various populations
- Omnipresent: possible for every tree
  - Likely for short branches or large population sizes
MSC and Identifiability

- A statistical model called multi-species coalescent (MSC) can generate ILS.
MSC and Identifiability

• A statistical model called multi-species coalescent (MSC) can generate ILS.

• Any species tree defines a unique distribution on the set of all possible gene trees
MSC and Identifiability

• A statistical model called multi-species coalescent (MSC) can generate ILS.

• Any species tree defines a unique distribution on the set of all possible gene trees.

• In principle, the species tree can be identified despite high discordance from the gene tree distribution.
Multi-gene tree estimation pipelines

Approach 1: Concatenation

Approach 2: Summary methods

There are also other approaches:
co-estimation (e.g., *BEAST), site-based (SVDQuartets)
Multi-gene tree estimation pipelines

**Approach 1: Concatenation**

Gene tree estimation

- Orangutan
- Gorilla
- Chimpanzee
- Human

Statistically **inconsistent** [Roch and Steel, 2014]

**Approach 2: Summary methods**

Gene tree estimation

- Orangutan
- Gorilla
- Chimpanzee
- Human

Can be statistically consistent given **true gene trees**
Multi-gene tree estimation pipelines

Approach 1: Concatenation

Statistically inconsistent [Roch and Steel, 2014]

Approach 2: Summary methods

STAR, STELLS, GLASS, BUCKy (population tree), MP-EST, NJst (ASTRID), …

ASTRAL, ASTRAL-II, ASTRAL-III

Can be statistically consistent given true gene trees
This talk: ASTRAL

- Optimization problem
- Dynamic programming solution
- Accuracy in simulation studies
  - With strong model violations
- Sample complexity
- Time complexity
- Handling of non-standard input
Unrooted quartets under MSC model

For a quartet (4 species), the unrooted species tree topology has at least 1/3 probability in gene trees (Allman, et al. 2010)
For a quartet (4 species), the unrooted species tree topology has at least 1/3 probability in gene trees (Allman, et al. 2010)

The most frequent gene tree = The most likely species tree
Unrooted quartets under MSC model

For a quartet (4 species), the unrooted species tree topology has at least 1/3 probability in gene trees (Allman, et al. 2010)

The most frequent gene tree = The most likely species tree

\[ \theta_1 = 1 - \frac{2}{3} e^{-d} \]
Unrooted quartets under MSC model

For a quartet (4 species), the unrooted species tree topology has at least 1/3 probability in gene trees (Allman, et al. 2010)

The most frequent gene tree = The most likely species tree

shorter branches ⇒ more discordance ⇒ a harder species tree reconstruction problem
More than 4 species

For >4 species, the species tree topology can be different from the most like gene tree (called anomaly zone) (Degnan, 2013)
For >4 species, the species tree topology can be different from the most like gene tree (called anomaly zone) (Degnan, 2013)

1. Break gene trees into \( \binom{n}{4} \) quartets of species
2. Find the dominant tree for all quartets of taxa
3. Combine quartet trees

Some tools (e.g., BUCKy-p [Larget, et al., 2010])
More than 4 species

For >4 species, the species tree topology can be different from the most like gene tree (called anomaly zone) (Degnan, 2013)

Alternative:
weight all $3(n^4)$ quartet topologies by their frequency and find the optimal tree

(probabilities are made-up just as an example)
Notations

- $n$ = the number of species
- $m$ = the number of gene trees
- $f$ = the length of the shortest branch
Maximum Quartet Support Species Tree

• Optimization problem:

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

\[ \text{Score}(T) = \sum_{1}^{m} |Q(T) \cap Q(t_i)| \]

the set of quartet trees induced by T

a gene tree
Maximum Quartet Support Species Tree

- Optimization problem:

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

\[ \text{Score}(T) = \sum_{1}^{m} |Q(T) \cap Q(t_i)| \]

- **Theorem**: Statistically consistent under the multi-species coalescent model when solved exactly
Maximum Quartet Support Species Tree

- Optimization problem: NP-Hard [Lafond & Scornavaccaori, 2016]

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

\[ \text{Score}(T) = \sum_{1}^{m} |Q(T) \cap Q(t_i)| \]

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

[Mirarab, et al., Bioinformatics, 2014]

• ASTRAL solves the problem exactly using **dynamic programming**
  
  • Exponential running time (feasible for $n < 18$)
ASTRAL-I
[Mirarab, et al., Bioinformatics, 2014]

- ASTRAL solves the problem exactly using **dynamic programming**
  - Exponential running time (feasible for \( n < 18 \))

- Introduced a **constrained version** of the problem
  - Draws the set of branches in the species tree from a given set \( \mathcal{X} = \{ \text{all bipartitions in all gene trees} \} \)
  - Species tree branches tend to be in at least one gene tree
  - **Theorem**: the constrained version is **statistically consistent**
  - Running time: \( O(n^2 m|\mathcal{X}|^2) \)
1. Faster calculation of the score function inside DP

- $O(nm|X|^2)$ instead of $O(n^2m|X|^2)$
ASTRAL-II
[Mirarab and Warnow, Bioinformatics, 2015]

1. Faster calculation of the score function inside DP
   - $O(nm|X|^2)$ instead of $O(n^2m|X|^2)$

2. Add extra bipartitions to the set $X$ using heuristics
   - Consensus + support + subsampling species
   - Using quartet-based distances to find likely branches
   - Complete incomplete gene trees
1. Faster calculation of the score function inside DP
   - $O(nm|\mathcal{X}|^2)$ instead of $O(n^2m|\mathcal{X}|^2)$

2. Add extra bipartitions to the set $\mathcal{X}$ using heuristics
   - Consensus + support + subsampling species
   - Using quartet-based distances to find likely branches
   - Complete incomplete gene trees

3. Ability to take as input gene trees with polytomies
Simulation study

Vary parameters: the number of species, the number of genes, and the amount of ILS.
Simulation study

- Vary parameters: the number of species, the number of genes, and the amount of ILS.
- Compare to other strong methods (NJst, MP-EST, concatenation)
Simulation study

- Vary parameters: the number of species, the number of genes, and the amount of ILS.

- Compare to other strong methods (NJst, MP-EST, concatenation)

- Evaluate using the FN rate: the percentage of branches (bipartitions) in the true tree that are missing from the estimated tree
Number of species impacts estimation error in the species tree

1000 genes, “medium” levels of ILS, simulated species trees
[S. Mirarab, T. Warnow, 2015]
ASTRAL: accurate and scalable

1000 genes, "medium" levels of ILS, simulated species trees
[S. Mirarab, T. Warnow, 2015]
ASTRAL: accurate and scalable

1000 genes, “medium” levels of ILS, simulated species trees
[S. Mirarab, T. Warnow, 2015]
Tree error as a function of # species

1000 genes, “medium” levels of ILS, simulated species trees
[S. Mirarab, T. Warnow, 2015]
Running time as function of # species

A new implementation called ASTRID (Warnow lab) is much faster (minutes)

1000 genes, “medium” levels of ILS, simulated species trees
[S. Mirarab, T. Warnow, 2015]
Horizontal Gene Transfer (HGT)

[R. Davidson et al., BMC Genomics. 16 (2015)]

Model violation: the simulated discordance is due to both ILS and randomly distributed HGT.

50 species, 10 genes

~30% discordance
~35%
~55%
~70%
Horizontal Gene Transfer (HGT)

[Davidson et al., BMC Genomics. 16 (2015)]

Randomly distributed HGT is tolerated with enough genes

50 species, varying # genes

~30% discordance
~55%
~35%
~70%
Used by the biologists

- Plants: Wickett, et al., 2014, PNAS
- Birds: Prum, et al., 2015, Nature
- Xenoturbella, Cannon et al., 2016, Nature
- Xenoturbella, Rouse et al., 2016, Nature
- Flatworms: Laumer, et al., 2015, eLife
- Angiosperms: Huang et al., 2016, MBE
- Worms: Andrade, et al., 2015, MBE
UCSD Work

1. Statistical support
2. Sample complexity
3. Running time complexity
4. Parallelism and new data types
Going beyond the topology

[Sayyari and Mirarab, MBE, 2016]

- **Branch length:** simply a function of the level of discordance

\[
\theta_1 = 1 - \frac{2}{3}e^{-d}
\]

\[
\theta_1 = 70\%, \quad \theta_2 = 15\%, \quad \theta_3 = 15\%
\]
Going beyond the topology

[Sayyari and Mirarab, MBE, 2016]

- **Branch length**: simply a function of the level of discordance

- A single quartet ($n=4$)
  - Easy: just reverse the discordance formula
  - gives the ML estimate

\[
\theta_1 = 1 - \frac{2}{3} e^{-d}
\]

\[
\theta_1 = 70\% \quad \theta_2 = 15\% \quad \theta_3 = 15\%
\]
Branch length for $n > 4$

- Simply average all quartet frequencies “around” that branch
- Justified given some assumptions

\[
\theta_1 = 1 - \frac{2}{3} e^{-d}
\]
Branch length for \( n > 4 \)

- Simply **average** all quartet frequencies “around” that branch
- Justified given some assumptions
- Can be done **efficiently** in \( \Theta(n^2 m) \) for all branches

\[
\theta_1 = 1 - \frac{2}{3}e^{-d}
\]
Branch support

- **Traditional** approach: Multi-locus bootstrapping (MLBS)
  - **Slow**: requires bootstrapping all genes (e.g., $100 \times m$ ML tree inferences)
  - **Inaccurate** and hard to interpret
    [Mirarab et al., Sys bio, 2014; Bayzid et al., PLoS One, 2015]

- We can do better!

[Image of graph showing bootstrap support vs. percent correct]
Branch support idea: $n=4$

- Recall quartet frequencies follow a multinomial distribution

  $$m_1 = 80 \quad m_2 = 63 \quad m_3 = 57$$

- $P(\text{topology seen in } m_1 / m \text{ gene trees is the species tree } ) = P(\theta_1 > 1/3 ) = P(\text{a 3-sided coin tossed } m \text{ times is biased towards the side that shows up } m_1 \text{ times})$
Branch support idea: \( n = 4 \)

- Recall quartet frequencies follow a multinomial distribution

\[
m_1 = 80 \quad m_2 = 63 \quad m_3 = 57
\]

- \( P(\theta_1 > 1/3) = P(\text{a 3-sided coin tossed } m \text{ times is biased towards the side that shows up } m_1 \text{ times}) \)

- Can be analytically solved
Posterior

\[ \Gamma t^{z_1} \left( \frac{1-t}{2} \right)^{n-z_1} \]

\[ P\left( \theta_1 > \frac{1}{3} \mid \bar{Z} = \bar{z} \right) = \frac{\int_{\frac{1}{3}}^{1} P(\bar{Z} = \bar{z} \mid \theta_1 = t) f_{\theta_1}(t) dt}{P(\bar{Z} = \bar{z})} \]

\[ \sum_{j=1}^{3} \int_{\frac{1}{3}}^{1} P(\bar{Z} = \bar{z} \mid \theta_j = t) f_{\theta_j}(t) dt \]

- Fast to calculate
- Depends on the frequency of not just the first topology, but also the frequency of second and third topologies

Prior: Yule process become conjugate
Prior

• All three topologies have equally prior

\[ Pr(\theta_1 > \frac{1}{3}) = Pr(\theta_2 > \frac{1}{3}) = Pr(\theta_3 > \frac{1}{3}) = \frac{1}{3} \]

• The species tree generated through a Birth-only (Yule) process with rate \( \lambda \)

  • Turns out to be the conjugate prior

  • (default) \( \lambda = 0.5 \rightarrow \) uniformly distributed branch lengths
Background: How many targets are enough?

The number of loci required for confidently resolving a phylogenetic relationship depends on its "hardness". Discordant gene trees due to incomplete lineage sorting (ILS) can make branches of a species tree very difficult to resolve. ILS is a function of the branch length and population size. This means that shorter branches or larger population sizes increase ILS, and can make species tree reconstruction more difficult. The co-PI has recently developed a new approach for estimating the support of a branch in a coalescent-based framework based on the proportion of gene trees that support quartet topologies in gene trees [31]. This new approach is analogous to finding the probability that a three-sided die is loaded towards each facet by observing outcomes of many tosses. We can also ask the reverse question: if we have an estimate of the degree of bias of a die, how many tosses are needed to confidently find its loaded side. Similarly, this approach [31] sheds light on the relationship between the number of genes required to achieve a level of support for a branch of a certain length in coalescent units (Fig. 3). While the co-PI's method of calculating support, which is implemented in ASTRAL [32, 33], gives the basic mathematical framework, more method development is needed (see below).

Research Approach

Suitable specimens of more than 80 species of Sabellidae and >120 Terebelliformia species have been collected from localities around the world in the last 15 years by the PI. Further collecting will be undertaken for some key taxa for a few more transcriptomes and for targeted DNA capture. We also have agreements from other experts in Sabellidae and Terebelliformia to provide other needed ethanol-preserved specimens for targeted capture. We will obtain ~ 50% of the known sabellid diversity and ~25% of terebelliforms for targeted capture sequencing (~500 species total). Our plan is to sample all genera, particularly focusing on their type species. This should allow for the generation of robust phylogenies, from which we will revise the taxonomy. Specimens will be vouchered at the SIO Benthic Invertebrate Collection and biodiversity information curated on the Encyclopedia of Life.

Sequencing (transcriptomes and targeted capture of DNA)

A targeted capture approach will be used with an appropriate number of loci (see below) to generate robust phylogenies of Sabellidae and Terebelliformia. Two different sets of targets will be designed from transcriptomes, across Sabellidae and Terebelliformia respectively. We have already generated new transcriptomes for nine sabellids, a serpulid and a fabriciid (bold terminals in Fig. 4), and nine Terebelliformia (bold terminals in Fig. 5) and combined this data with the few publicly available transcriptomes. Based on direct sequencing data already obtained for several genes for Sabellidae and Terebelliformia [Rouse in prep.; Stiller et al. in prep.], our sampling arguably spans the extant diversity of each clade. Several more transcriptomes will be needed to ensure the targeted capture methods will be robust. Methods for generating these transcriptomes will follow those previously used by us [6, 34].

How many targets are needed?

A targeted capture pipeline typically starts from a large number of loci sequenced from a small number of species using genome-wide approaches (e.g., transcriptomics) [7, 8, 35]. This initial dataset is then used to select a smaller subset of loci for the targeted capture phase, which involves a larger set of species. To reduce the cost and effort, one would want to minimize the number of loci in the second phase, as long as sufficient loci are selected to confidently resolve relationships of interest. To date the number of loci has been determined in an ad hoc manner, either by the ultra conserved elements discovered [7, 8] or limitations of the targeted capture technology [13]. Initial analyses of our two datasets demonstrates how the number of loci affects support (Fig. 3).

- Increased number of genes ($m$) $\Rightarrow$ increased support
- Decreased discordance $\Rightarrow$ increased support
How about $n > 4$?

- **Locality Assumption**: All four clusters around a branch are correct
  - Treat branches independently

\[ k = n_1 \times n_2 \times n_3 \times n_4 \]
How about \( n > 4 \)?

- **Locality Assumption**: All four clusters around a branch are correct
  - Treat branches independently

- \( k \) quartets around a branch?
  - Independence assumption would be too liberal (\( m \times k \) tosses of the coin)
  - Fully dependent assumption:
    - \( k \) quartets give noisy estimates of a single hidden true frequency.
    - Simply average their frequencies

- \( m \) tosses, \( m \times k \) readings
Simulation studies

- Assumption **violations**:
  - Estimated gene trees instead of true gene trees
  - Estimated species trees: the locality assumption can be violated

- Measuring the support **accuracy**: ROC curves — based on the number of false positive and false negatives above various thresholds of support
Results (Avian, ROC)

Avian simulated dataset (48 taxa, 1000 genes)

[Sayyari and Mirarab, MBE, 2016]
Branch length accuracy

With true gene trees, ASTRAL correctly estimates BL.
Branch length accuracy

With error-prone estimated gene trees, ASTRAL underestimates BL
Sample complexity

Shubhanshu Shekhar  Sebastien Roch
Sample complexity?

- How many genes are needed to guarantee an arbitrarily high probability of finding the true species tree?

![Species tree diagram](image)
Sample complexity?

• **How many genes** are needed to guarantee an arbitrarily high probability of finding the true species tree?

• … asymptotically, as the problem gets more difficult.

• \( f \): the length of the shortest branch (difficulty)

• Find \( m \), as a function of \( f \) and \( n \) for probability of error \( \epsilon \)
Theorem 1

Consider a model species tree with minimum branch length \( f < \log(\sqrt{2}) \). Then, for any \( \epsilon > 0 \), ASTRAL (exact) returns the true species tree with probability at least \( 1 - \epsilon \) if the number of input error-free gene trees satisfies

\[
m > \frac{9}{2} \log \left( \frac{4}{\epsilon} \binom{n}{4} \right) \frac{1}{(1 - e^{-f})^2}
\]  

(1)

Theorem 2

For any \( \rho \in (0, 1) \) and \( a \in (0, 1) \), there exist constants \( f_0 \) and \( n_0 \) such that the following holds. For all \( n \geq n_0 \) and \( f \leq f_0 \), there exists a species tree with \( n \) leaves and shortest branch length \( f \) such that when ASTRAL (exact) is used with \( m \leq \frac{a \log n}{5f^2} \) gene trees, the event \( E \) that ASTRAL (exact) reconstructs the wrong tree has probability

\[
P(E) \geq 1 - \rho.
\]  

(3)
The sample complexity of the ASTRAL optimization problem (exact solution) is

\[ O(\log(n)f^{-2}) \]

such that when ASTRAL (exact) is used with \( m \leq \frac{a \log n}{5f^2} \) gene trees, the event \( E \) that ASTRAL (exact) reconstructs the wrong tree has probability

\[ P(E) \geq 1 - \rho. \]
Simulations match theory
Simulations match theory
ASTRAL v.s. ASTRID: depends
Running time complexity & ASTRAL-III (unpublished work)
ASTRAL-III new features

• Handling new dataset types (multiple individuals)

• Bounded polynomial running time: $O((nm)^{2.73})$

• GPU and CPU parallelism

• Using a polytree to reduce speed to $O(|\mathcal{U}|(nm)^{1.73})$

• $\mathcal{U} = \{\text{unique nodes in gene trees}\}$; $|\mathcal{U}| = O(nm)$
Multiple individuals

• What if we sample multiple individuals from each species?

• In recently diverged species individuals *may* have different trees for each gene

• Sampling multiple individuals may provide extra signal
Extending ASTRAL to multiple individuals

• Optimization problem: quartet-based score easily extends (simply ignore multi-individual quartets)

• Dynamic programming: simply adjust boundary conditions

• **Challenge**: forming constrained search space (set $\mathcal{X}$)

  • New heuristics: repeated subsampling of individuals and taking a consensus among subsamples
Multiple individuals helpful?

Yes, it marginally helps accuracy
Multiple individuals helpful?

Yes, it marginally helps accuracy

But not if sequencing effort is kept fixed
Multiple individuals helpful?

- Variable effort: 500 genes; 200 species
- Fixed effort: genes x inds = 1000; 200 species
- Fixed effort: 500 genes; species x inds = 200

Species tree error (RF distance)

1 ind. 2 ind. 5 ind.

Yes, it marginally helps accuracy

But not if sequencing effort is kept fixed
Asymptotic running time?

• Simple discrete math question:

  • \( \mathcal{X} = \) a set of subsets of some set \( \mathcal{L} \).

  \[ \mathcal{Y} = \{ (a, b) \in \mathcal{X} \mid a \cap b = \emptyset, \ a \cup b \in \mathcal{X} \} \]

  • Clearly, \( |\mathcal{Y}| < |\mathcal{X}|^2 \)

  • What’s the maximum \( |\mathcal{Y}| \) with respect to \( |\mathcal{X}| \)?
Asymptotic running time?

• Simple discrete math question:

  • $X$ = a set of subsets of some set $L$.
  $Y = \{ (a, b) \in X \mid a \cap b = \emptyset, a \cup b \in X \}$

  • Clearly, $|Y| < |X|^2$

  • What’s the maximum $|Y|$ with respect to $|X|$?

• Turns out to be rather challenging

  • Daniel Kane and Terence Tao proved: $|Y| = O(|X|^{1.73})$
Bounding ASTRAL running time

• ASTRAL running time is $O(n m |\mathcal{X}|^{1.73})$. 
Bounding ASTRAL running time

• ASTRAL running time is $O(nm|\mathcal{X}|^{1.73})$.

• What is $|\mathcal{X}|$?
  • ASTRAL-I: edges in gene trees $\Rightarrow |\mathcal{X}|=O(nm)$
  • ASTRAL-II: ASTRAL-I + uncontrolled heuristics
  • ASTRAL-III: control heuristics to $|\mathcal{X}|=O(nm)$
Bounding ASTRAL running time

• ASTRAL running time is $O(nm|\mathcal{X}|^{1.73})$.

• What is $|\mathcal{X}|$?
  • ASTRAL-I: edges in gene trees $\Rightarrow |\mathcal{X}| = O(nm)$
  • ASTRAL-II: ASTRAL-I + uncontrolled heuristics
  • ASTRAL-III: control heuristics to $|\mathcal{X}| = O(nm)$

• ASTRAL-III is bounded at $O((nm)^{2.73})$
  • Bounding the running time does not hurt accuracy (simulations)
Further improvements

• Use a polytree to overly all the input gene trees into one data structure

• Allows us to spend time for each unique node in gene trees once

• $O(|U|(nm)^{1.73})$ unique nodes in gene trees = $O(nm)$

• 3X running time improvement
Parallelism

• We need $O((nm)^{1.73})$ weights
• “almost” independent
• can send each task to a different CPU or GPU core
• Can now infer trees with 10,000 species & 400 genes in less than a day
Moving forward …

- ASTRAL, like all other two-step approaches, is sensitive to errors in the input gene trees

- Can it be changed to use characters directly? possible but slow for binary characters …

- More broadly, can alternative scalable methods be developed for better gene tree estimation?
• ASTRAL, like all other two-step approaches, is sensitive to errors in the input gene trees

• Can it be changed to use characters directly? possible but slow for binary characters …

• More broadly, can alternative scalable methods be developed for better gene tree estimation?

• ASTRAL scales to 10K leaves. We have 90K bacterial genomes.

• Can we scale further? … divide-and-conquer …
Moving forward …

• ASTRAL, like all other two-step approaches, is sensitive to errors in the input gene trees

  • Can it be changed to use characters directly? possible but slow for binary characters …

  • More broadly, can alternative scalable methods be developed for better gene tree estimation?

• ASTRAL scales to 10K leaves. We have 90K bacterial genomes.

  • Can we scale further? … divide-and-conquer …

• Theory: can the running time be further improved? Can the sample complexity be established for heuristic ASTRAL?
Summary

• ASTRAL is one of the leading methods for species tree reconstruction from gene trees
  • Can handle all types of inputs used in practice (missing data, polytomies, multiple individuals, …)
  • Has high accuracy given good gene trees
  • Seems robust to some model violations (but not high gene tree error)
  • Is scalable to very large datasets (10K leaves)
• Combines CS theory+statistics+heuristic techniques +efficient implementation+parallelism+software support
Theoretical sample complexity results

How many genes are enough to reconstruct the tree?

\[ m \geq \frac{9}{2} \log \left( \frac{4\binom{n}{4}}{\varepsilon} \right) \frac{c}{\alpha^2 f^2} \]