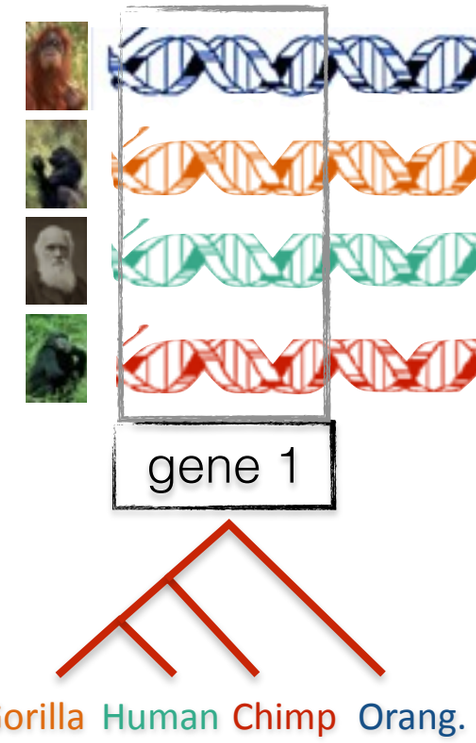
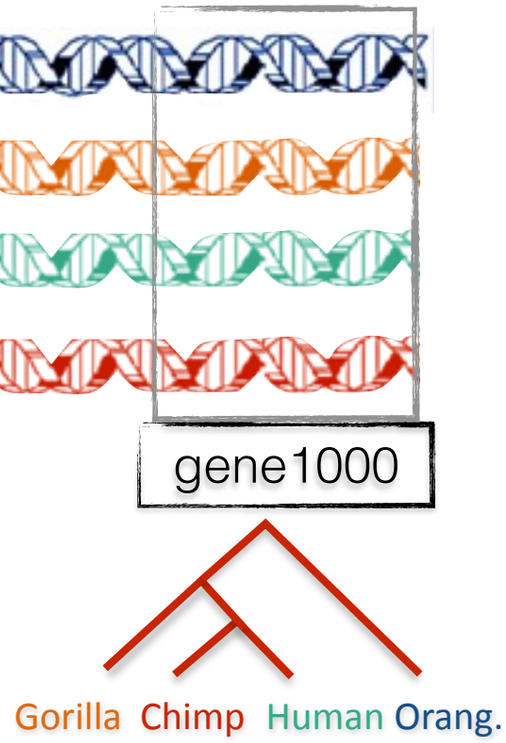
ASTRAL-III: increased scalability and impacts of contracting low support branches

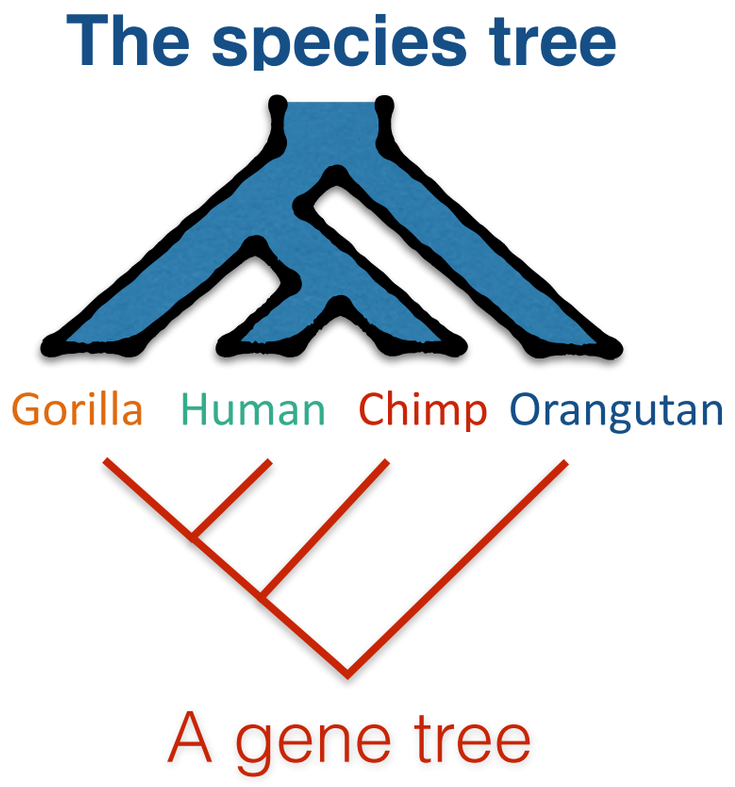
Chao Zhang
Maryam Rabiee
Erfan Sayyari
Siavash Mirarab

University of California, San Diego

Gene tree discordance





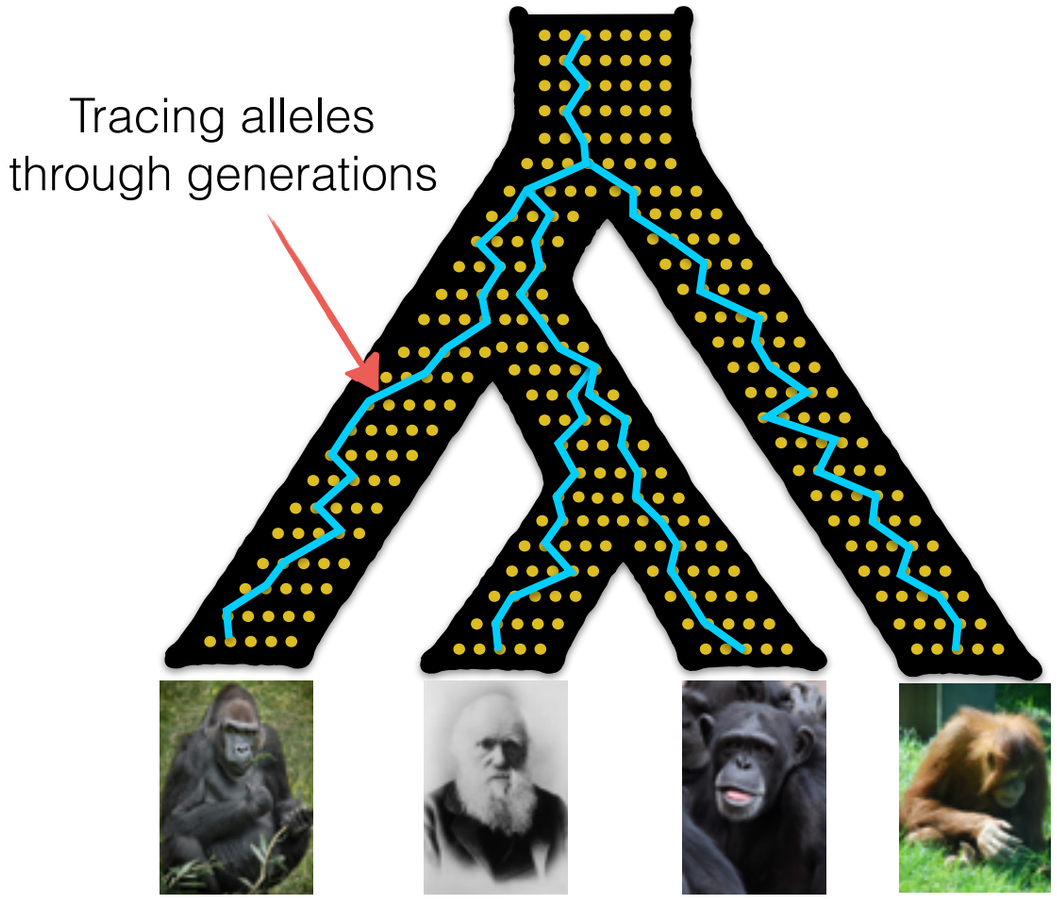


Causes of gene tree discordance include:

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

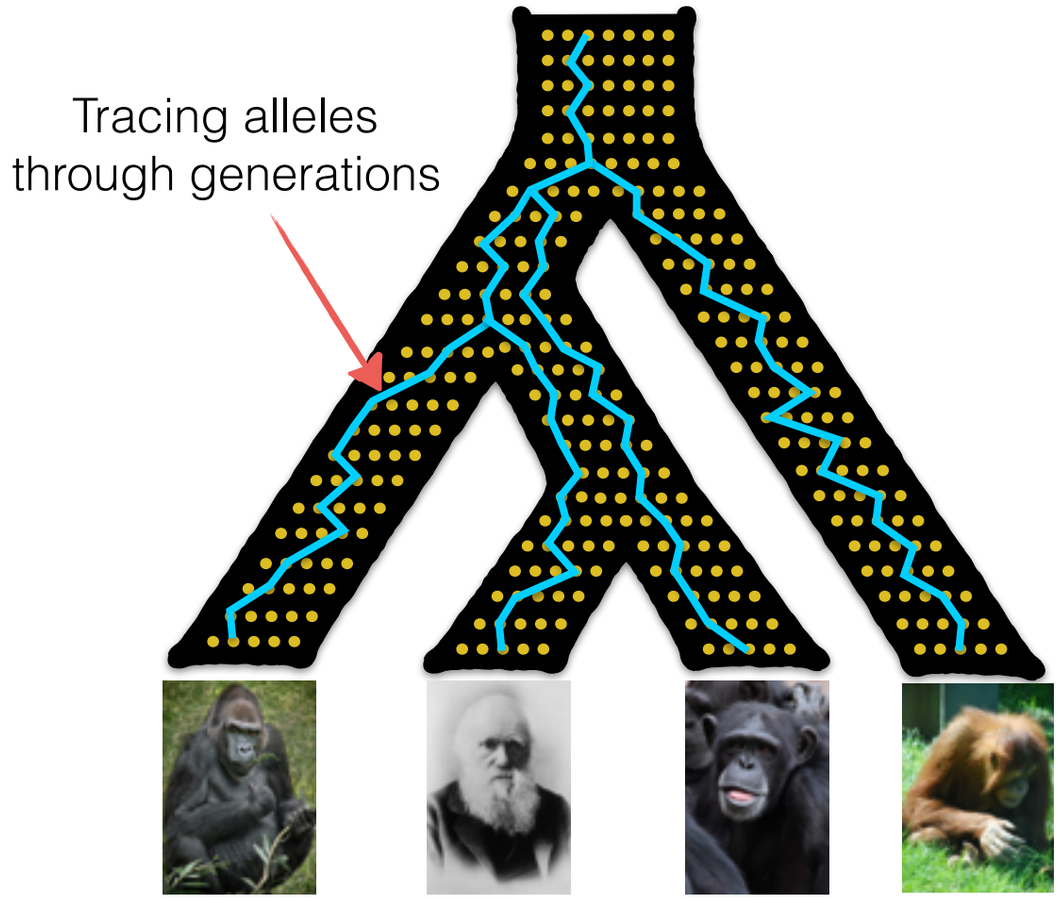
Incomplete Lineage Sorting (ILS)

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



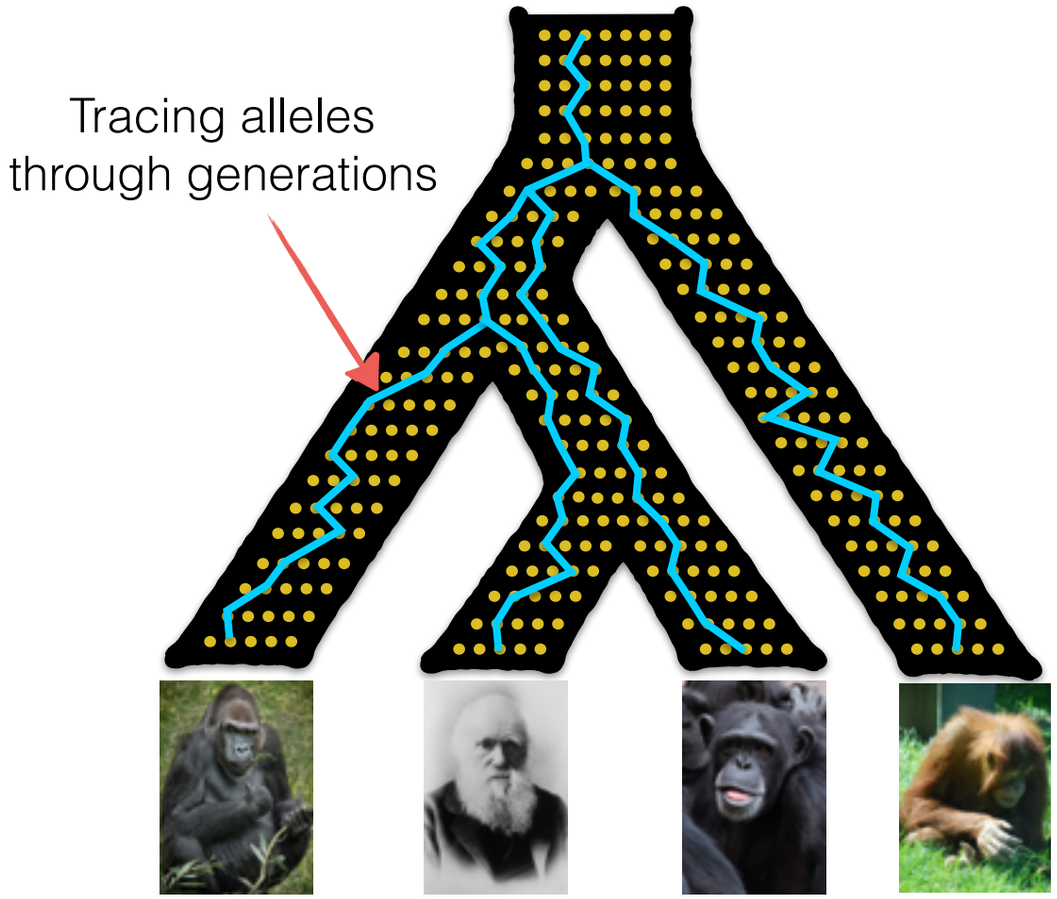
Incomplete Lineage Sorting (ILS)

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



Incomplete Lineage Sorting (ILS)

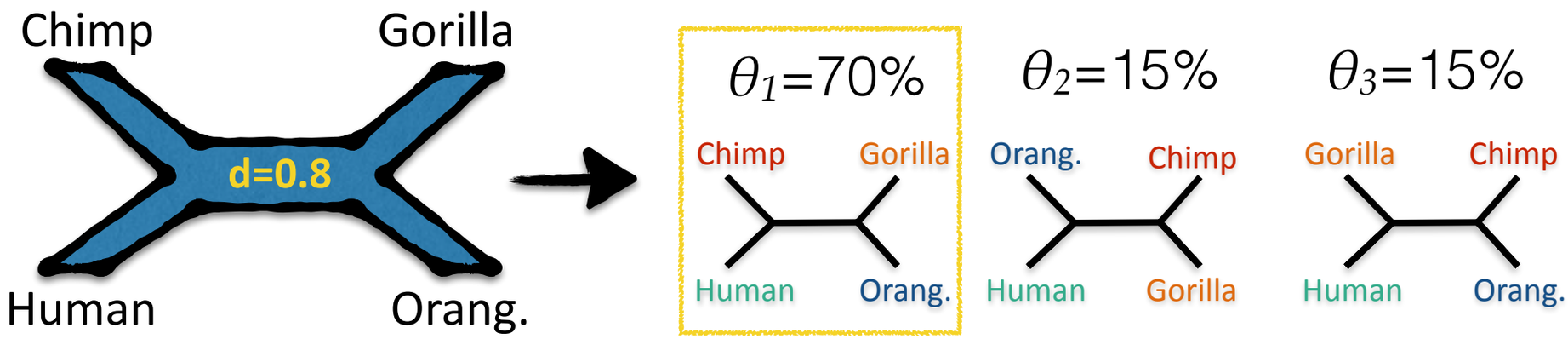
A random process related to the coalescence of alleles across various populations



Multi-species coalescent (MSC) model captures ILS

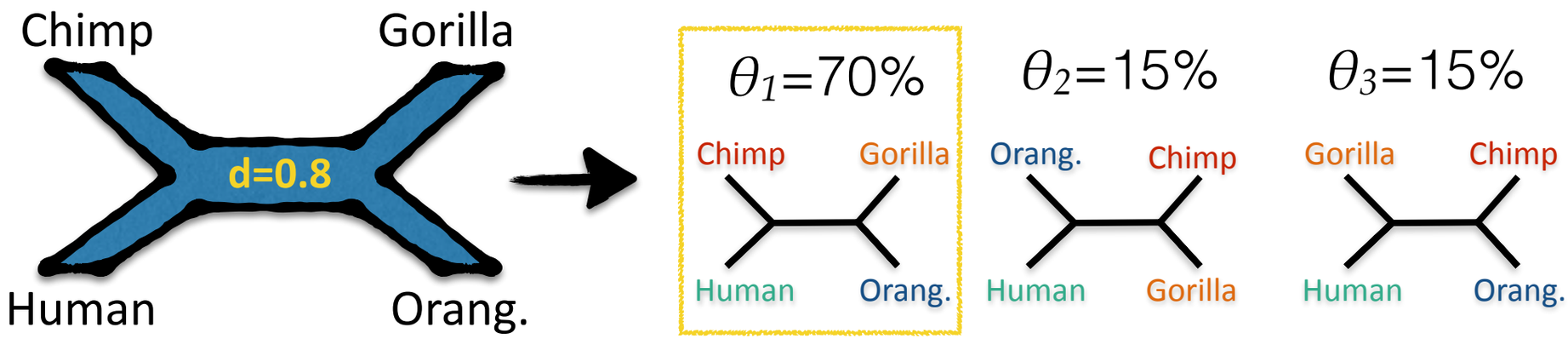
Unrooted quartets under MSC model

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



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)



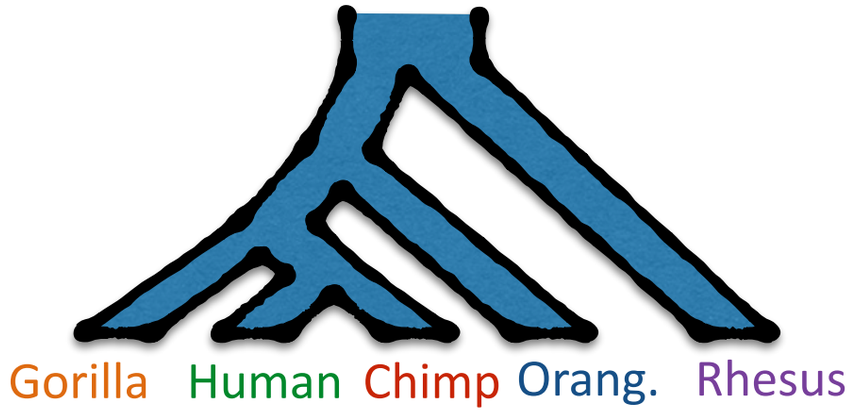
The most frequent gene tree

=

The most likely species tree

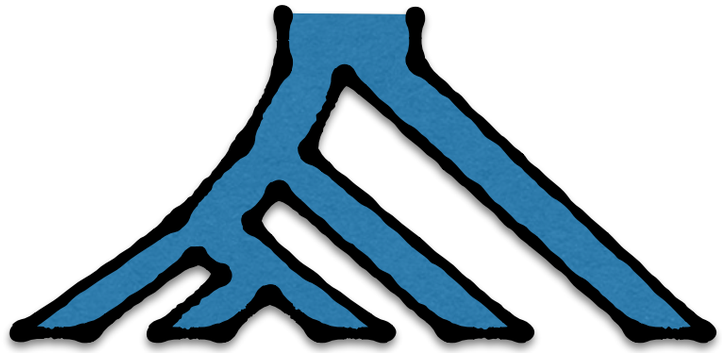
More than 4 species

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



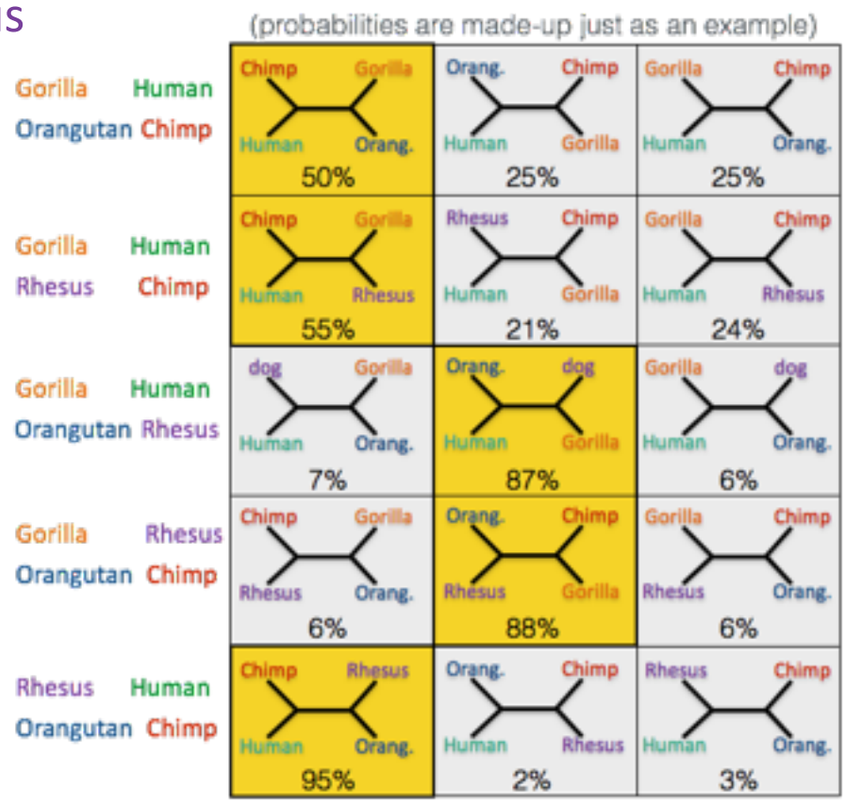
More than 4 species

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



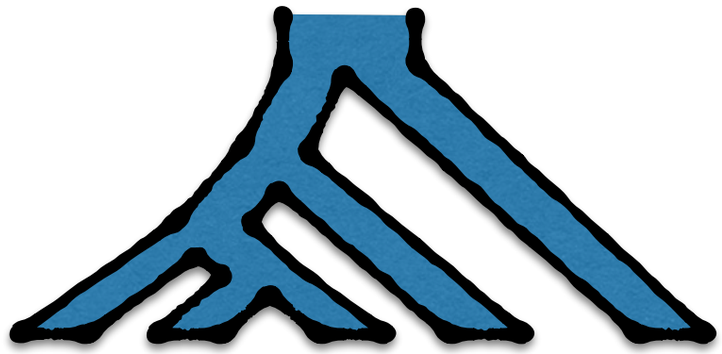
1. Break gene trees into $\binom{n}{4}$ quartets of species
2. Find the dominant tree for all quartets of taxa
3. Combine quartet trees

   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)



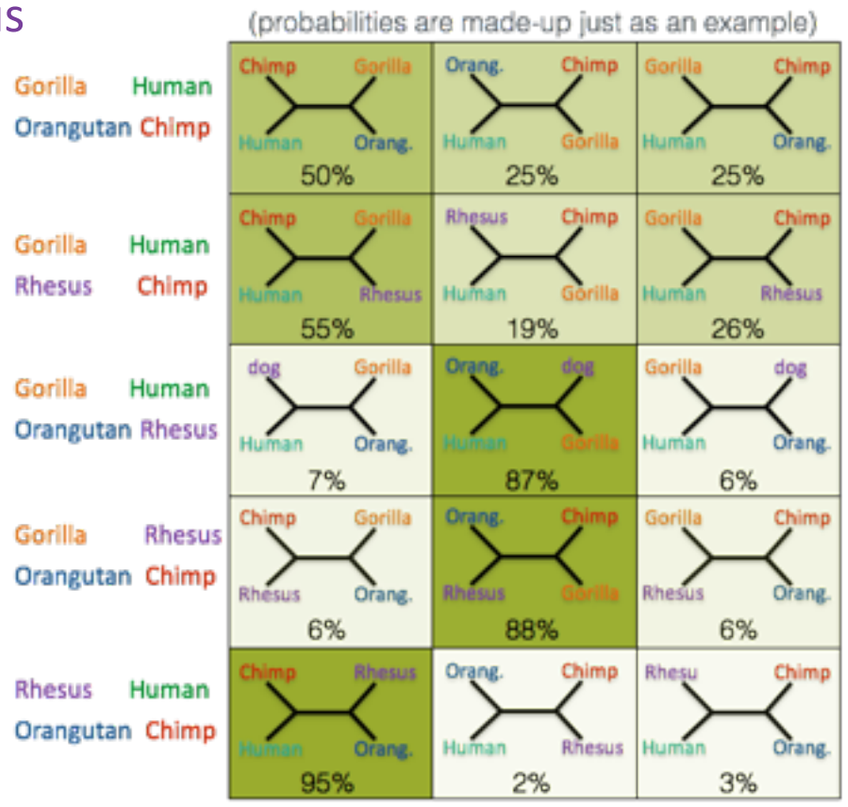
Gorilla Human Chimp Orang. Rhesus

1 s

2

3

Alternative:
weight all $3\binom{n}{4}$ quartet topologies
by their frequency
and find the optimal 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 k input gene trees

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

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 k input gene trees

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

the set of quartet trees induced by T

a gene tree

- **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 k input gene trees

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

the set of quartet trees induced by T

a gene tree

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

ASTRAL-I

[Mirarab, et al., Bioinformatics, 2014]

- ASTRAL solves the problem using **dynamic programming**

$$S(\mathcal{A}) = \max_{\mathcal{A}' \subset \mathcal{A}} \{S(\mathcal{A}') + S(\mathcal{A} - \mathcal{A}') + w(\mathcal{A}' | \mathcal{A} - \mathcal{A}' | \mathcal{L} - \mathcal{A})\}$$

ASTRAL-I

[Mirarab, et al., Bioinformatics, 2014]

- ASTRAL solves the problem using **dynamic programming**

$$S(\mathcal{A}) = \max_{\mathcal{A}' \subset \mathcal{A}} \{S(\mathcal{A}') + S(\mathcal{A} - \mathcal{A}') + w(\mathcal{A}' | \mathcal{A} - \mathcal{A}' | \mathcal{L} - \mathcal{A})\}$$

- Introduced a **constrained version** of the problem

$$S(\mathcal{A}) = \max_{\{\mathcal{A}', \mathcal{A}-\mathcal{A}'\} \subset \mathcal{X}} \{S(\mathcal{A}') + S(\mathcal{A} - \mathcal{A}') + w(\mathcal{A}' | \mathcal{A} - \mathcal{A}' | \mathcal{L} - \mathcal{A})\}$$

  - Draws branches in the species tree from a given set
    $\mathcal{X}$ = {all bipartitions in all gene trees}
  - Th constrained version is statistically consistent
  - Running time: $O(n^2 k |\mathcal{X}|^{1.73})$

ASTRAL-II

[Mirarab and Warnow, Bioinformatics, 2015]

1. Faster calculation of the score function inside DP

   - $O(nk|\mathcal{X}|^{1.73})$ instead of $O(n^2k|\mathcal{X}|^{1.73})$

ASTRAL-II

[Mirarab and Warnow, Bioinformatics, 2015]

1. Faster calculation of the score function inside DP

   - $O(nk|\mathcal{X}|^{1.73})$ instead of $O(n^2k|\mathcal{X}|^{1.73})$

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

ASTRAL-II

[Mirarab and Warnow, Bioinformatics, 2015]

1. Faster calculation of the score function inside DP

   - $O(nk|\mathcal{X}|^{1.73})$ instead of $O(n^2k|\mathcal{X}|^{1.73})$

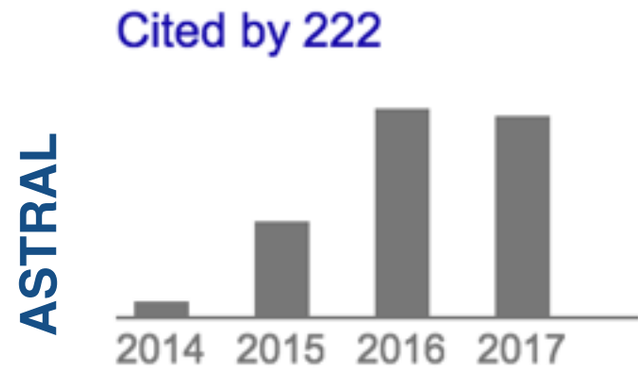
2. Add extra bipartitions to the set $\mathcal{X}$ using heuristics

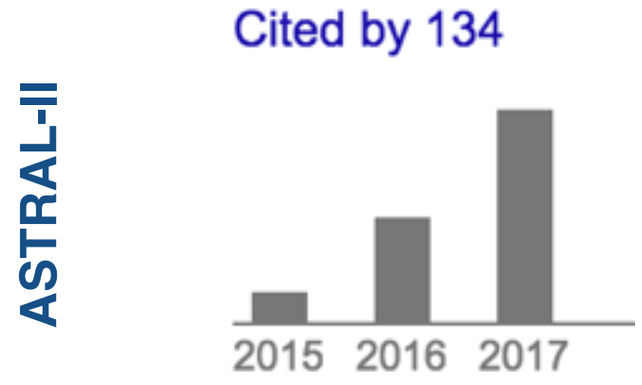
   - Consensus + support + subsampling species
   - Using quartet-based distances to find likely branches
   - Complete incomplete gene trees

3. Can handle input gene trees with polytomies in $O(n^3k|\mathcal{X}|^{1.73})$

ASTRAL 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
- Shrews: Giarla, et al., 2015, Syst. Bio.
- Frogs: Yuan et al., 2016, Syst. Bio.
- Tomatoes: Pease, et al., 2016, PLoS Bio.
- Angiosperms: Huang et al., 2016, MBE
- Worms: Andrade, et al., 2015, MBE



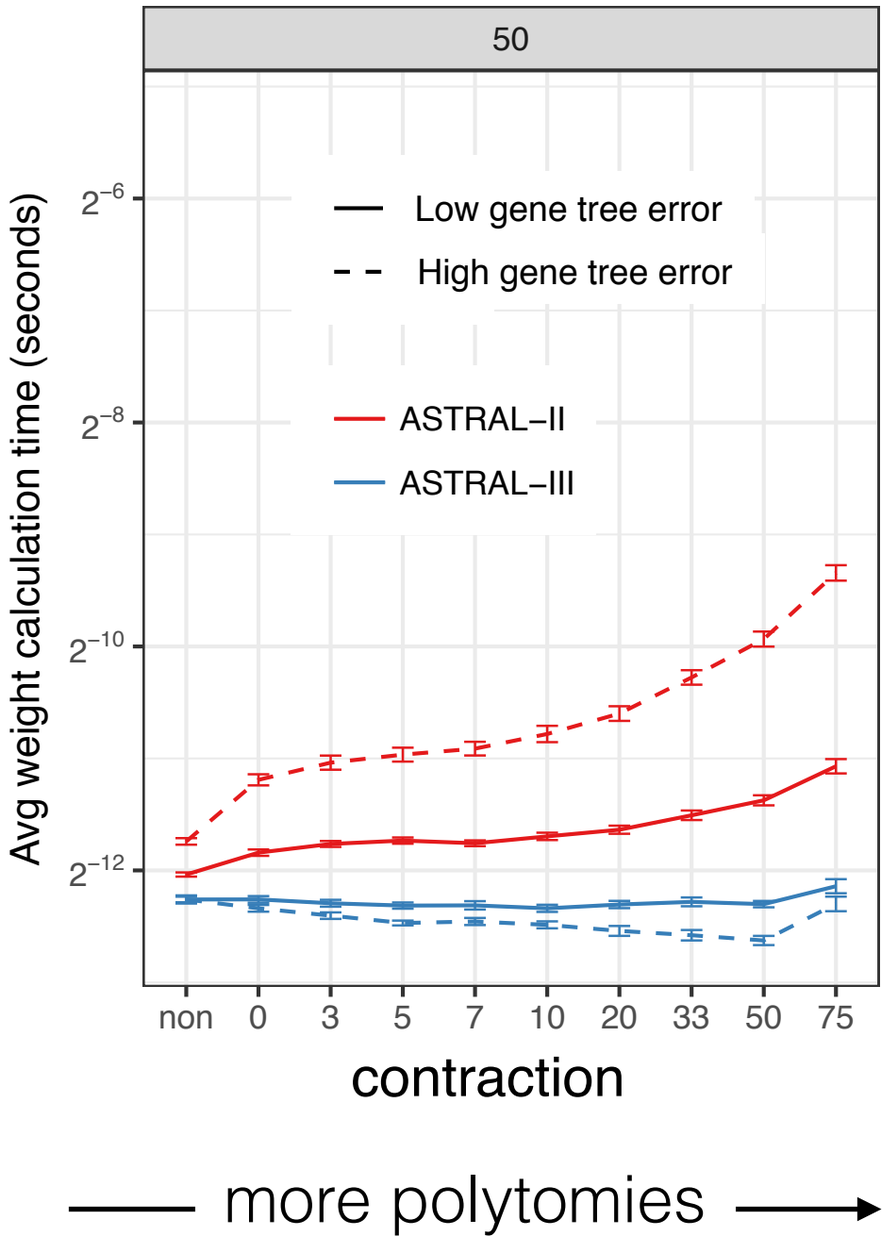


ASTRAL-III: improved running time

Feature 1:

ASTRAL-III can now analyze multifurcating gene trees with the same running time as binary gene trees:

$$O(nk|\mathcal{X}|^{1.73})$$



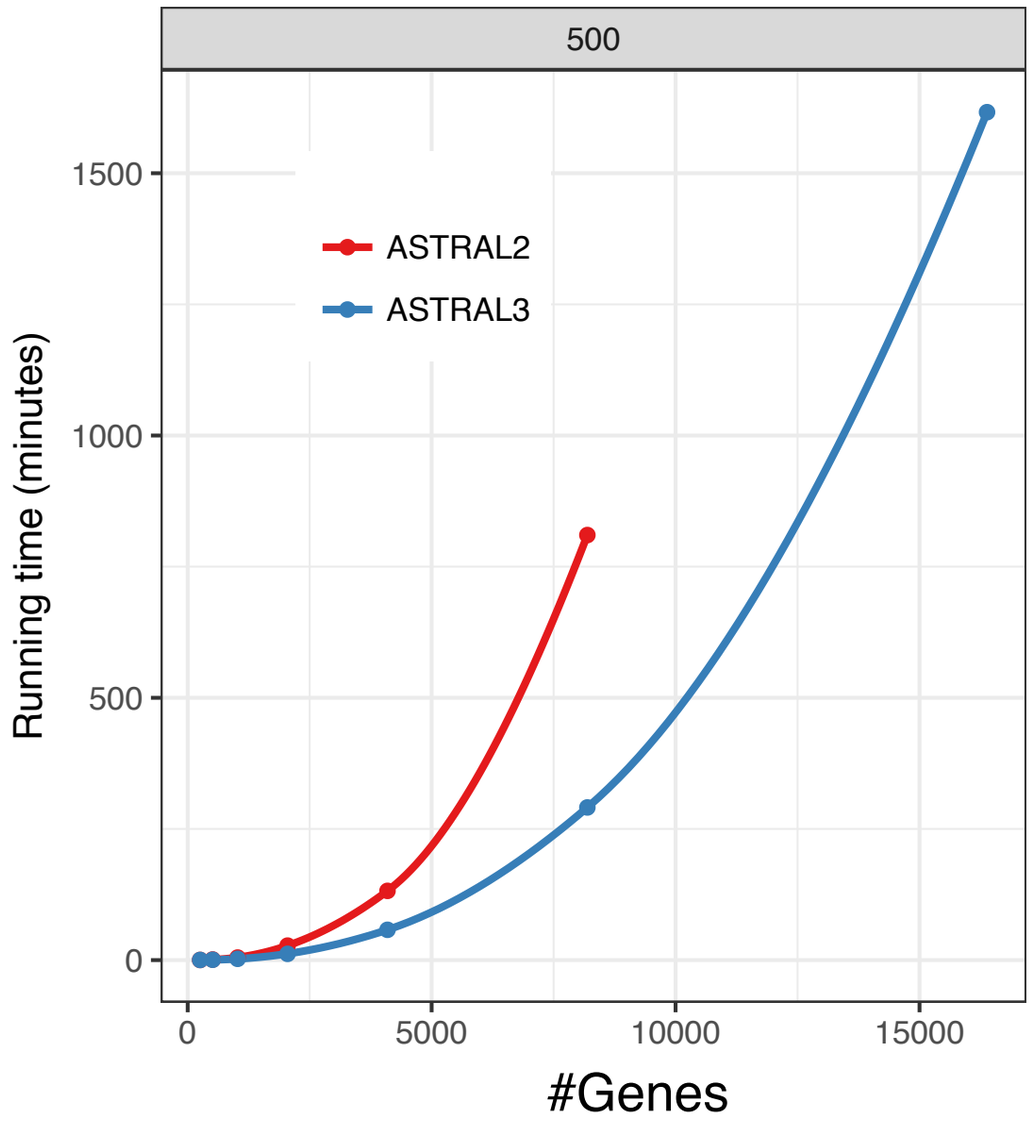
ASTRAL-III: improved running time

Feature 2:

Can exploit similarities between gene trees

- $O(\mathcal{D}|\mathcal{X}|^{1.73})$ where $\mathcal{D}$ is the sum of degrees of all unique gene tree nodes

- Overlay all gene trees onto a single polytree; use a bottom-up traversal to score a cluster



ASTRAL-III: improved running time

Feature 3.

An A*-like algorithm is used to trim the dynamic programming

- Compute an upper-bound on the score of a cluster without “expanding” it
- If the upper bound is below the best existing score, don’t expand

New features (journal version)

4. Restrict the search space to guarantee $|\mathcal{X}|=O(nk)$

- Overall running time is $O(\mathcal{D}(nk)^{1.73})= O((nk)^{2.73})$
- Does not impact accuracy
- Further improves speed

New features (journal version)

4. Restrict the search space to guarantee $|\mathcal{X}|=O(nk)$

   - Overall running time is $O(\mathcal{D}(nk)^{1.73})= O((nk)^{2.73})$

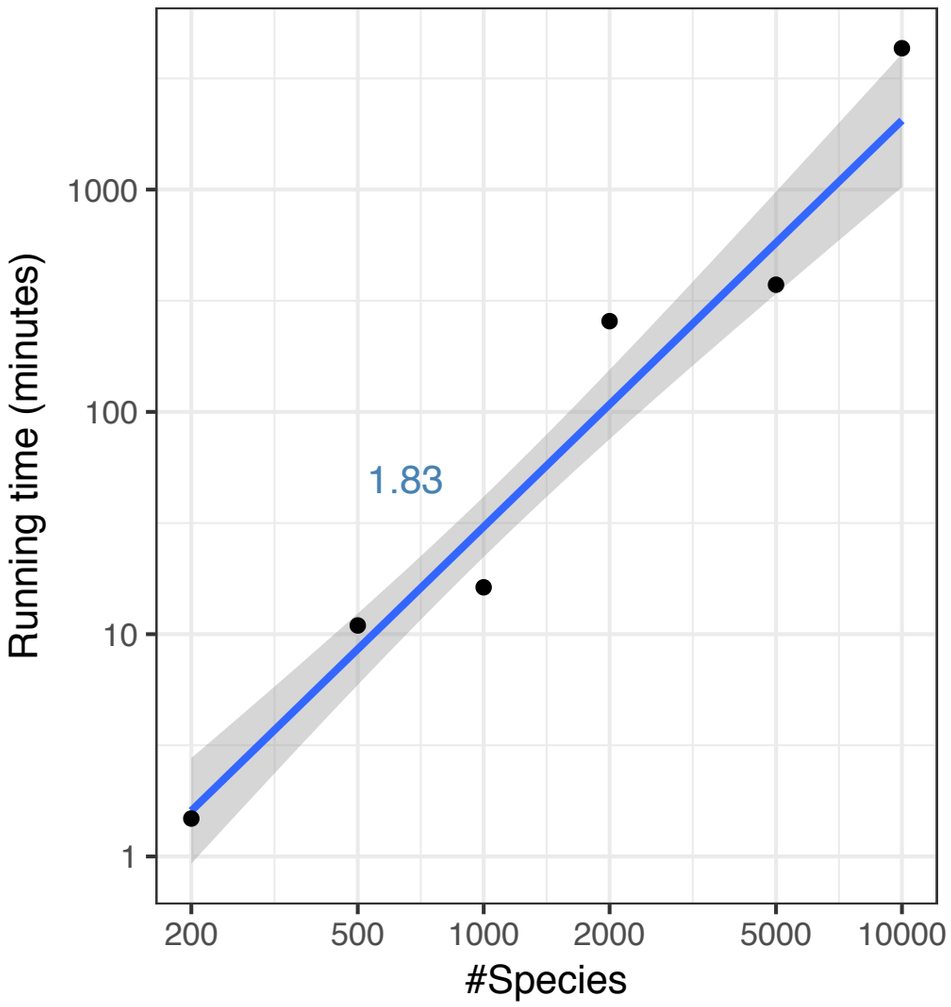
   - Does not impact accuracy

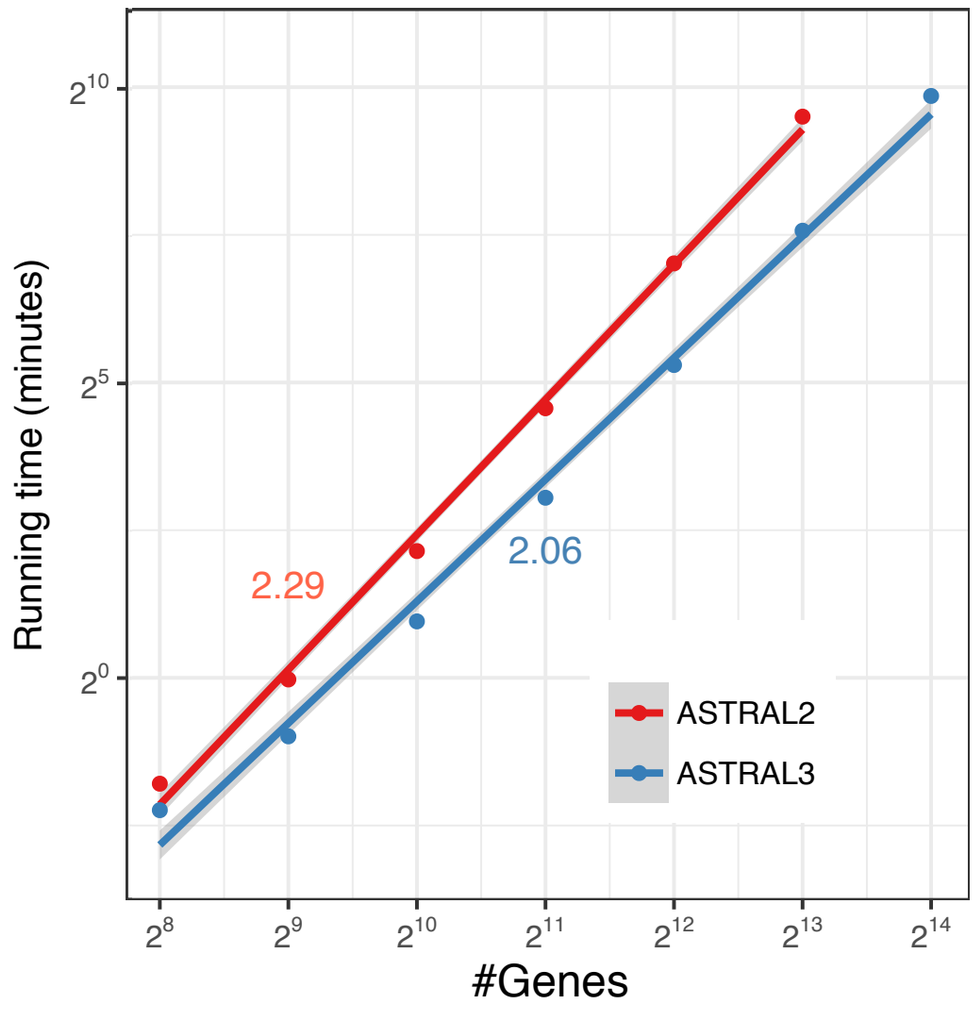
   - Further improves speed

5. The A*-like algorithm is further developed to compute even tighter upper bounds (complicated)

   - Improves speed only empirically and slightly

Empirical running time

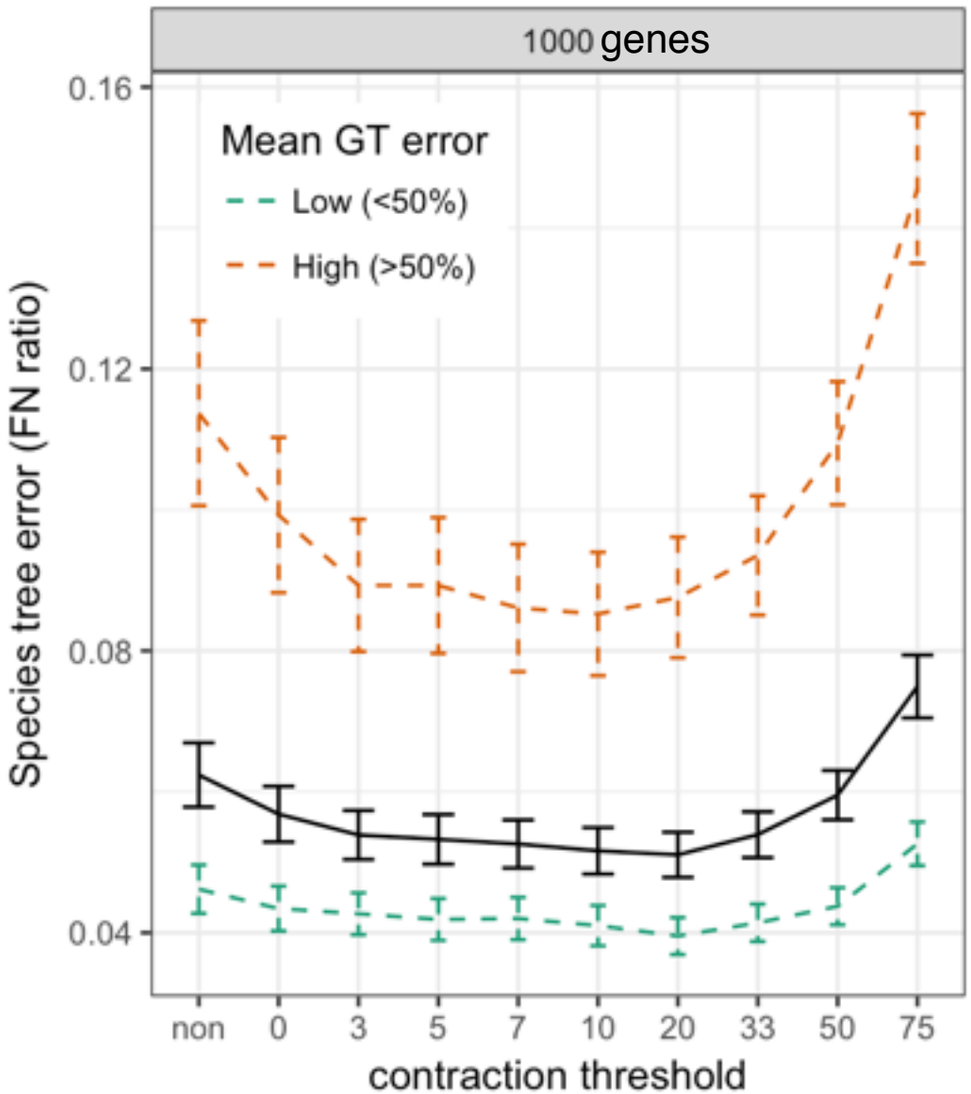




Empirical running time seems close to $O((nk)^2)$

Low support branches

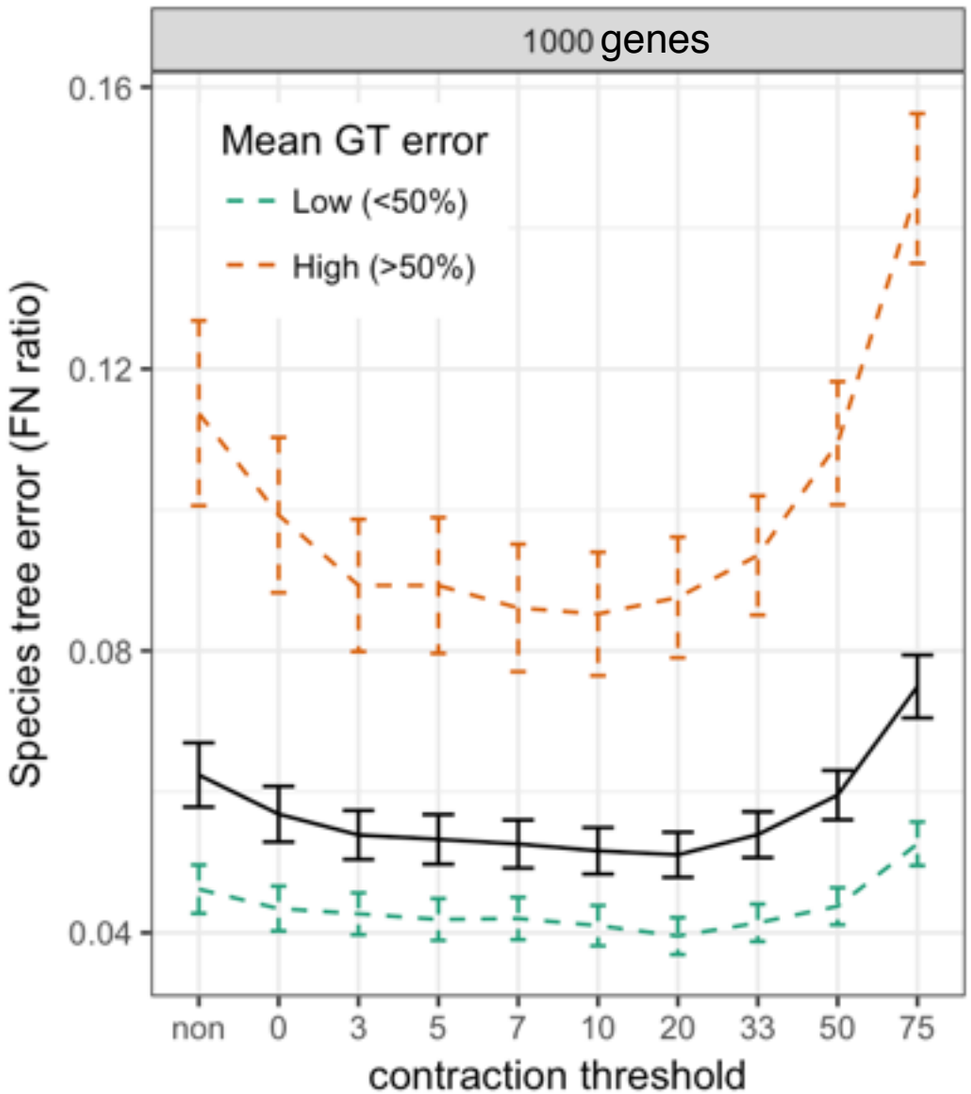
- Does it help to contract branches with low support?
- Yes, but only for **very low** supports



Simulations: 100 taxa, simphy,
ILS: around 46% true discordance
FastTree, support from bootstrapping

Low support branches

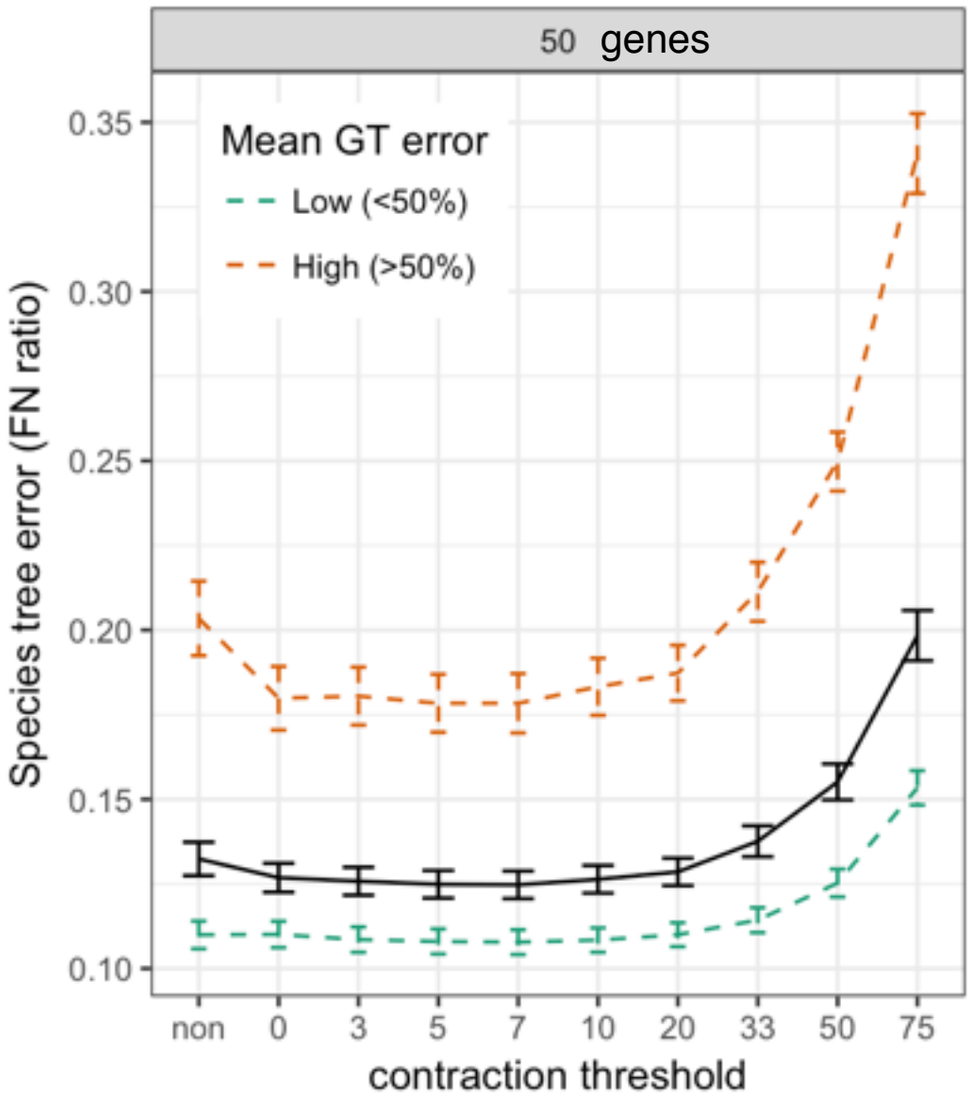
- Does it help to contract branches with low support?
- Yes, but only for **very low** supports
- Mostly helps in the presence of low support gene trees



Simulations: 100 taxa, simphy,
ILS: around 46% true discordance
FastTree, support from bootstrapping

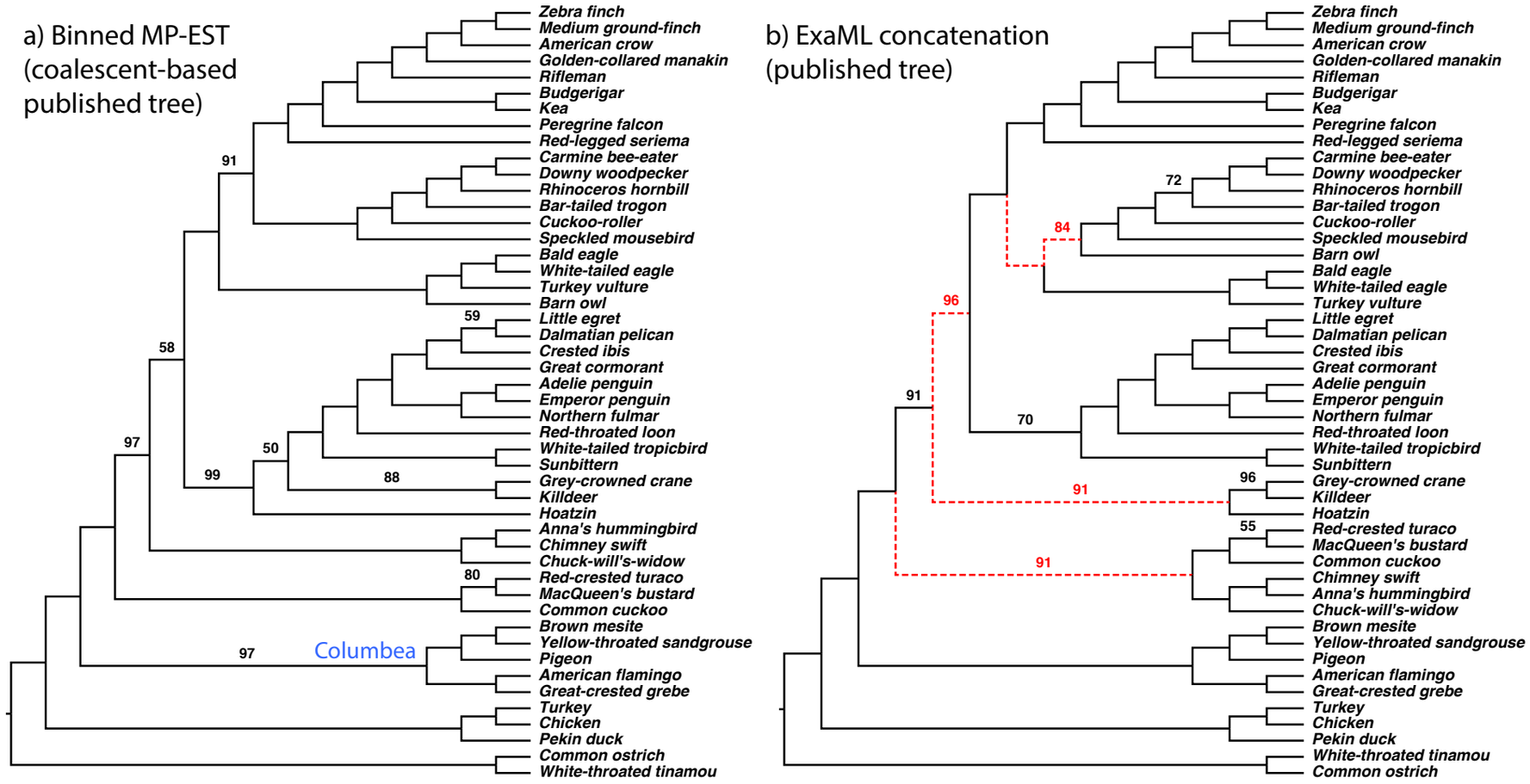
Low support branches

- Does it help to contract branches with low support?
- Yes, but only for **very low** supports
- Mostly helps in the presence of low support gene trees
- More genes allows for more filtering



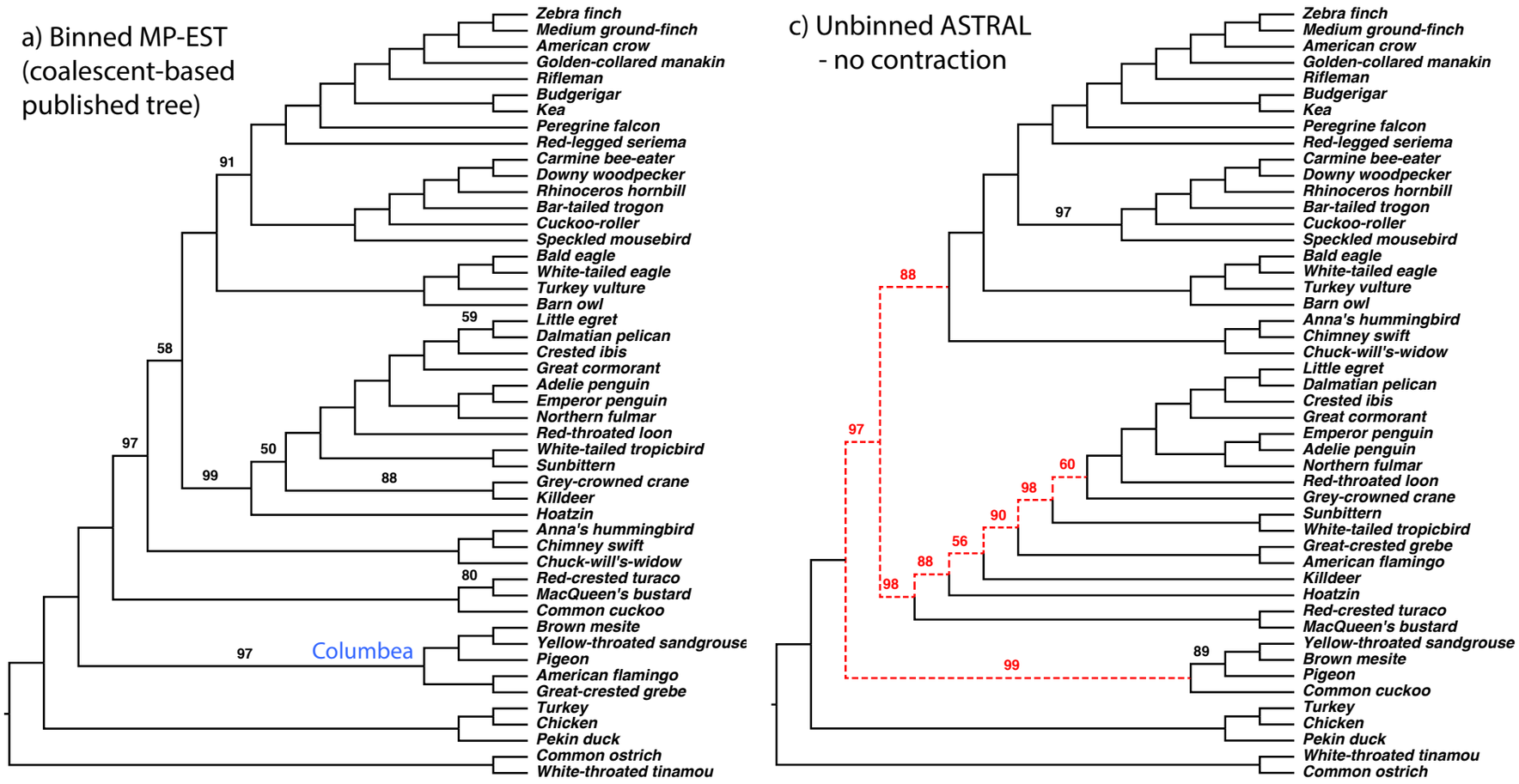
Simulations: 100 taxa, simphy,
ILS: around 46% true discordance
FastTree, support from bootstrapping

Avian biological dataset



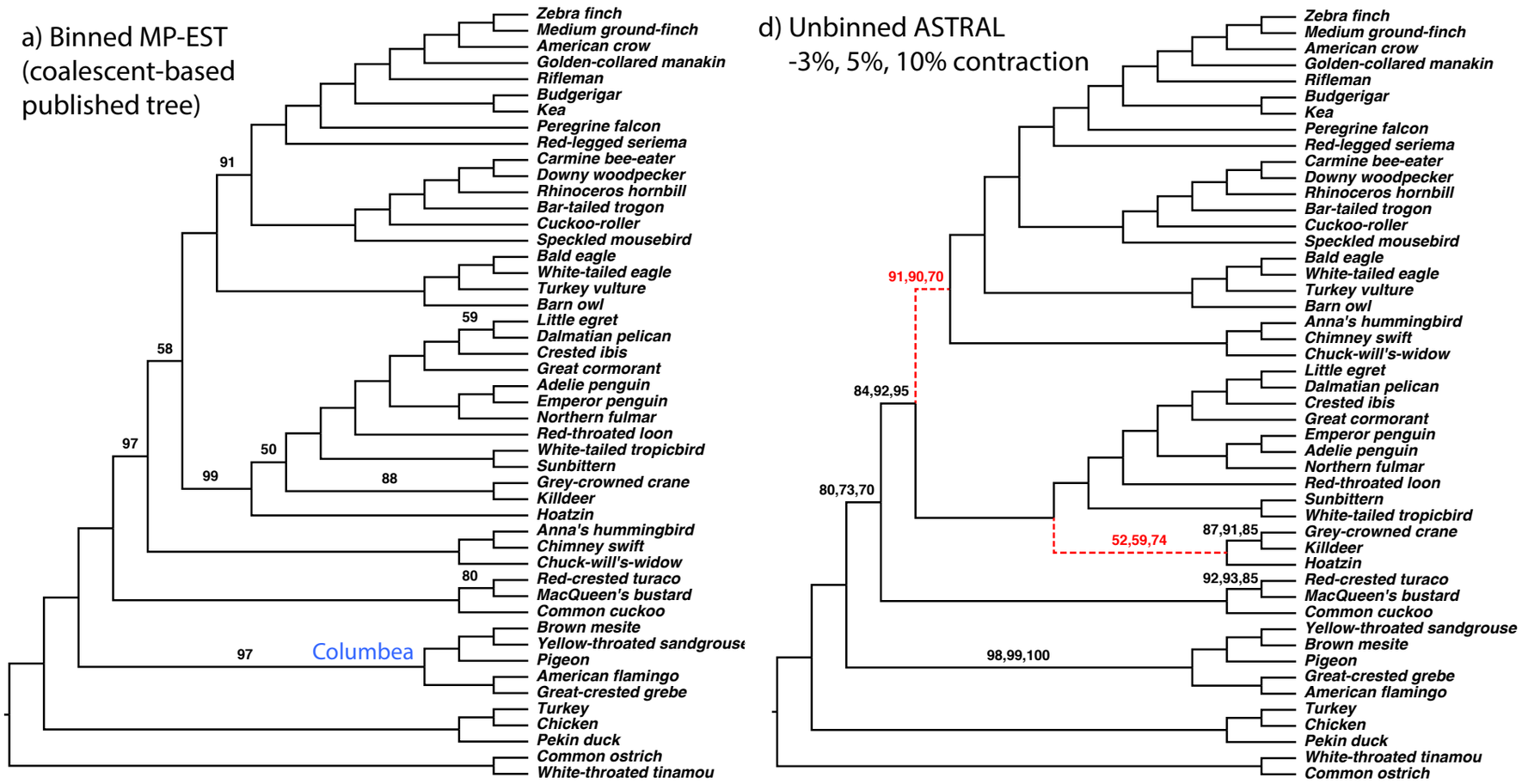
14K gene trees with very low gene tree resolution (25% mean BS)

Avian biological dataset



14K gene trees with very low gene tree resolution (25% mean BS)

Avian biological dataset



14K gene trees with very low gene tree resolution (25% mean BS)

Moving forward …

- ASTRAL, like all other two-step approaches, is sensitive to errors in the input gene trees
- What else can be done to reduce impacts of gene tree error?

Moving forward …

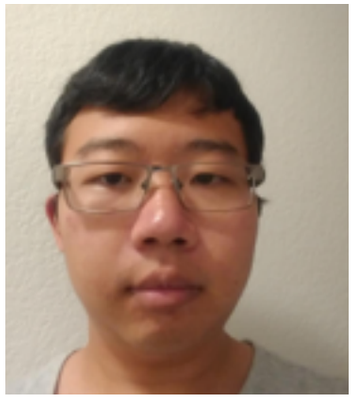
- ASTRAL, like all other two-step approaches, is sensitive to errors in the input gene trees
  - What else can be done to reduce impacts of gene tree error?

- ASTRAL scales to 10K species.
  We have 90K bacterial genomes.
  - Can we scale further? … divide-and-conquer …

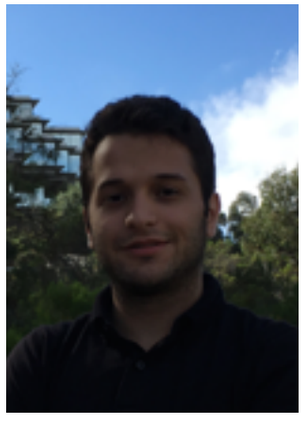
ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Tandy Warnow

UC San Diego

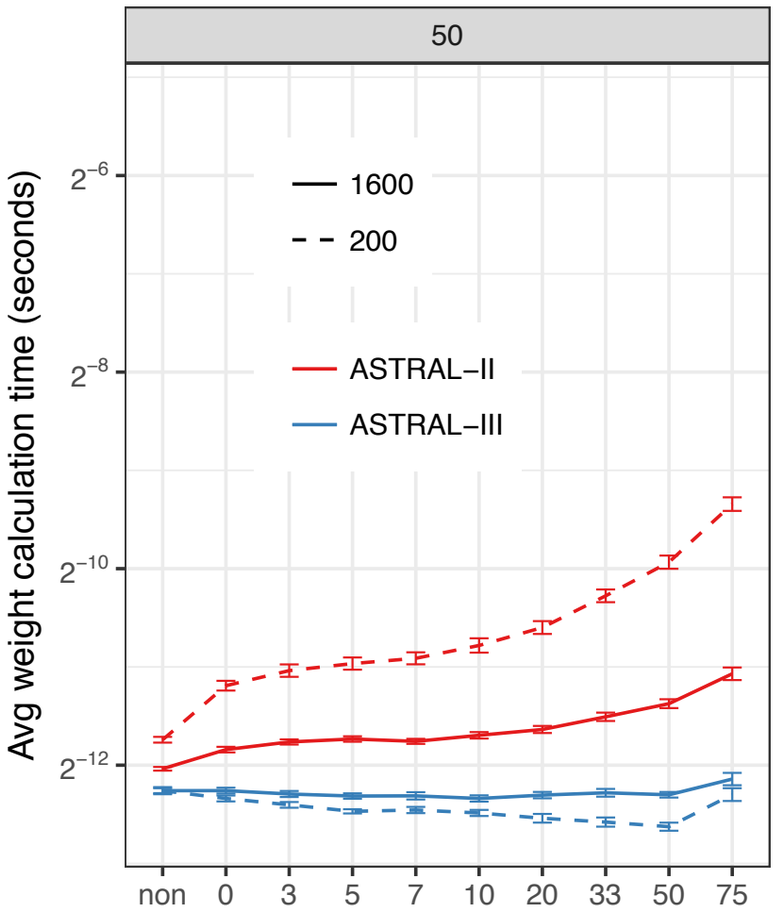


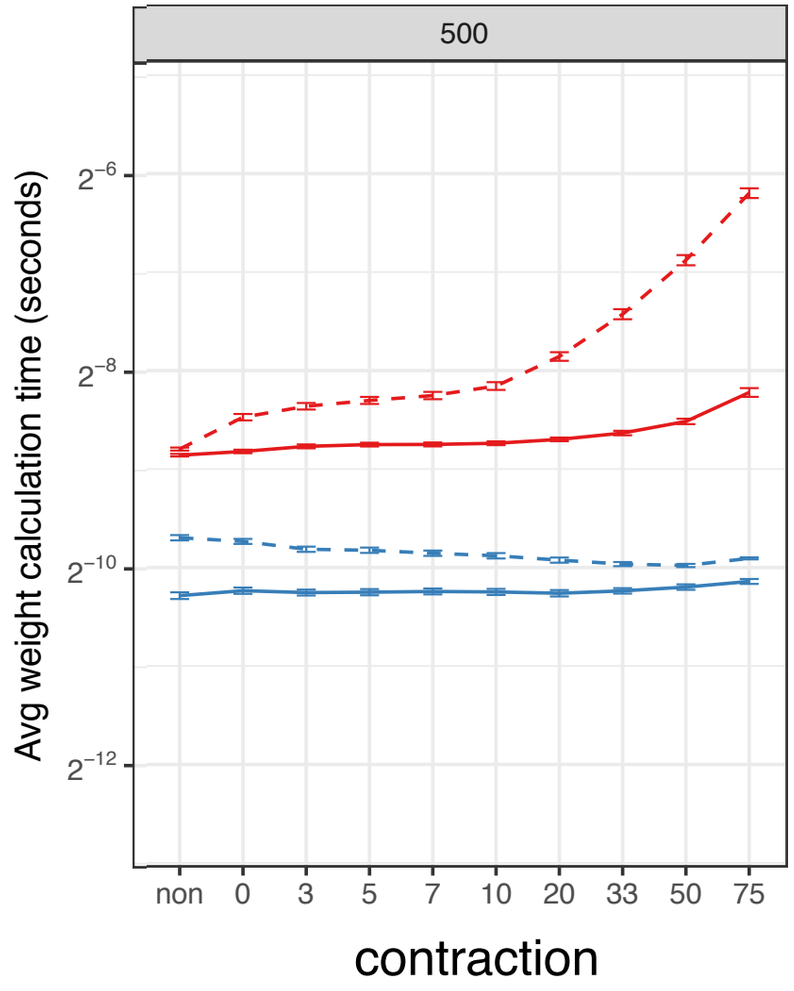
Chao Zhang



Erfan Sayyari

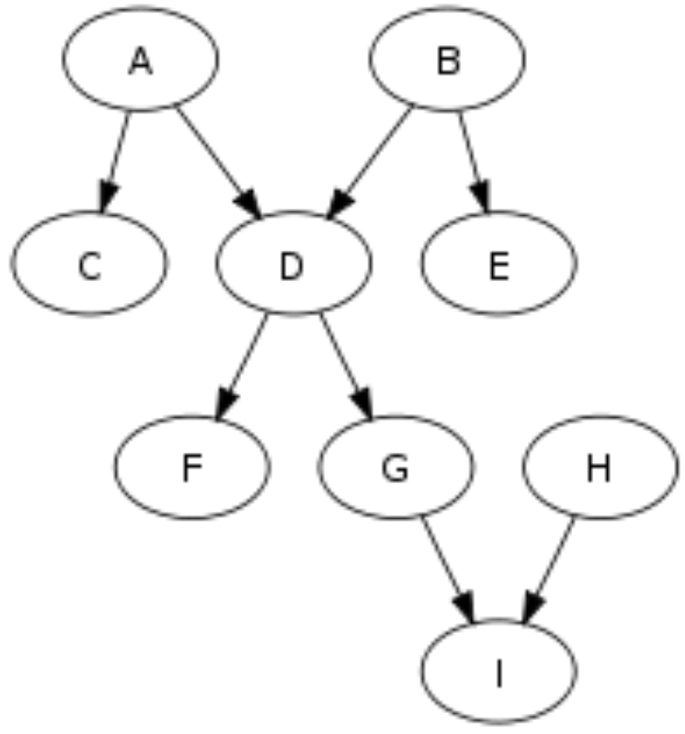
Maryam Rabiee


50
Avg weight calculation time (seconds)
2^{-6}
2^{-8}
2^{-10}
2^{-12}
1600
200
ASTRAL−II
ASTRAL−III
non
0
3
5
7
10
20
33
50
75


500
Avg weight calculation time (seconds)
2^{-6}
2^{-8}
2^{-10}
2^{-12}
non
0
3
5
7
10
20
33
50
75
contraction

Further improvements

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



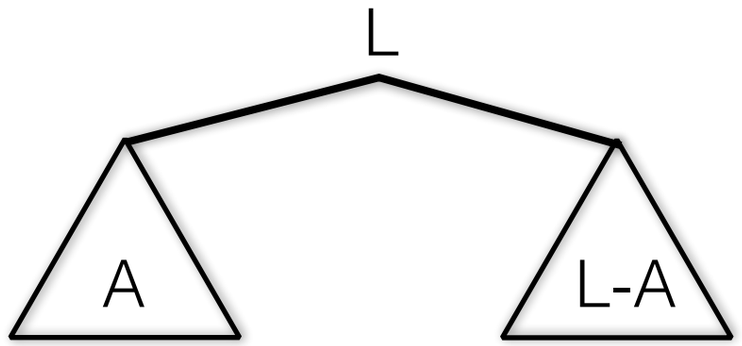
- Allows us to spend time for each unique node in gene trees once
  - $O(|\mathcal{D}|(nm)^{1.73})$

unique nodes in gene trees = $O(nm)$

- 3X running time improvement

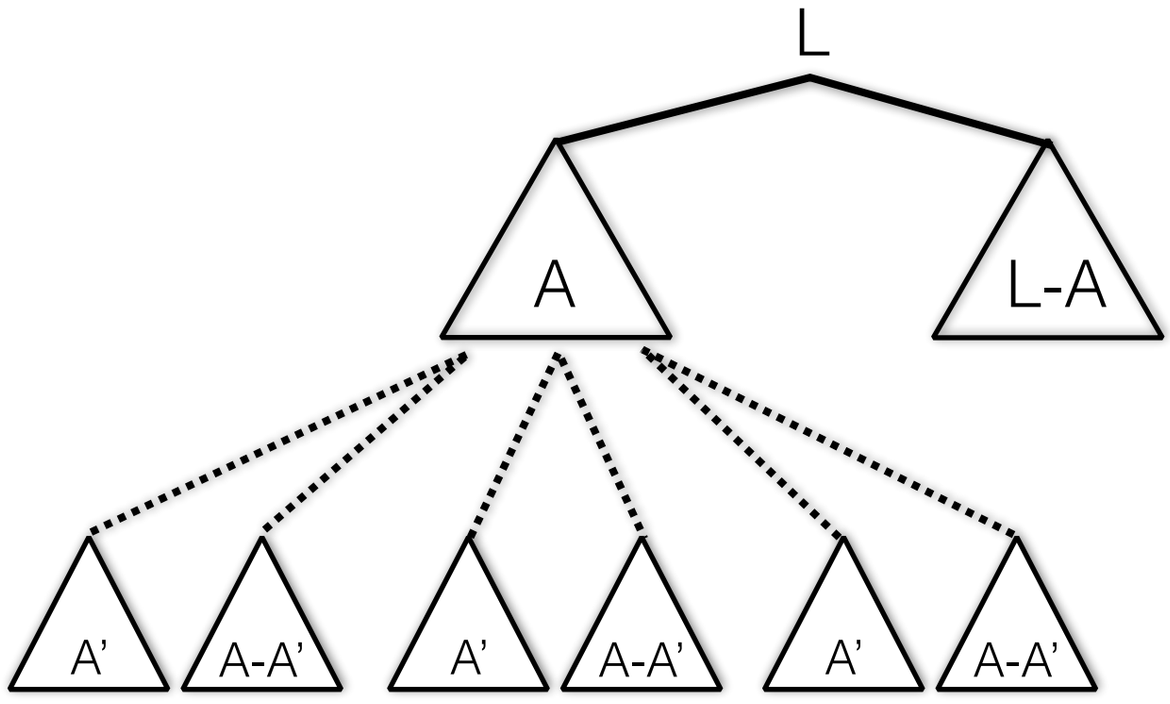
Dynamic programming

$$S(\mathcal{A}) = \max_{\mathcal{A}' \subset \mathcal{A}} \{S(\mathcal{A}') + S(\mathcal{A} - \mathcal{A}') + w(\mathcal{A}'|\mathcal{A} - \mathcal{A}'|\mathcal{L} - \mathcal{A})\}$$



Dynamic programming

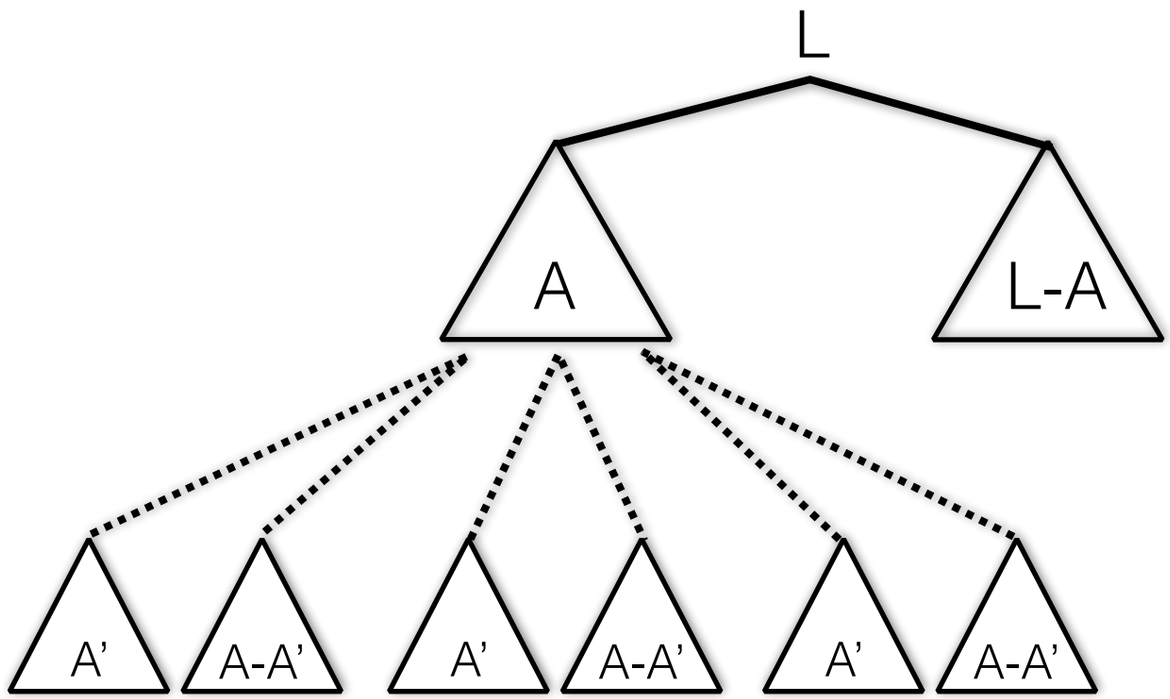
$$S(\mathcal{A}) = \max_{\mathcal{A}' \subset \mathcal{A}} \{S(\mathcal{A}') + S(\mathcal{A} - \mathcal{A}') + w(\mathcal{A}'|\mathcal{A} - \mathcal{A}'|\mathcal{L} - \mathcal{A})\}$$

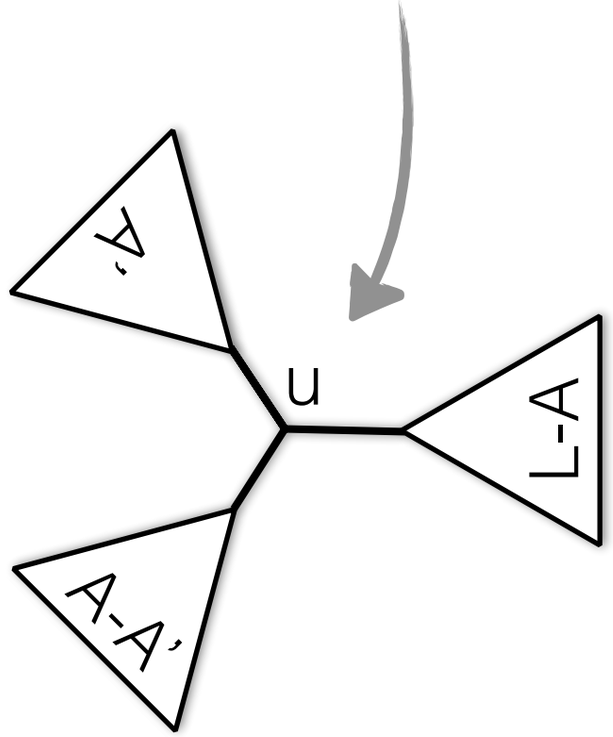


- Recursively break subsets of species into smaller subsets

Dynamic programming

$$S(\mathcal{A}) = \max_{\mathcal{A}' \subset \mathcal{A}} \{S(\mathcal{A}') + S(\mathcal{A} - \mathcal{A}') + w(\mathcal{A}' | \mathcal{A} - \mathcal{A}' | \mathcal{L} - \mathcal{A})\}$$



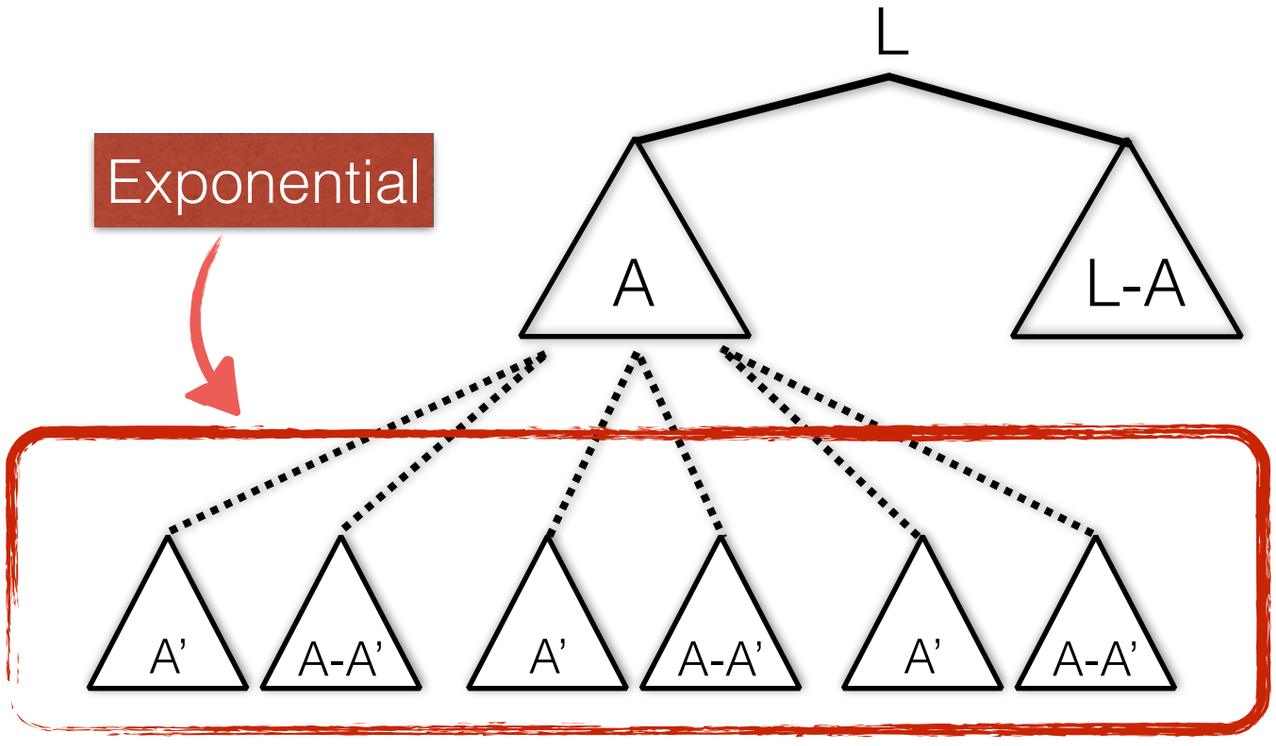


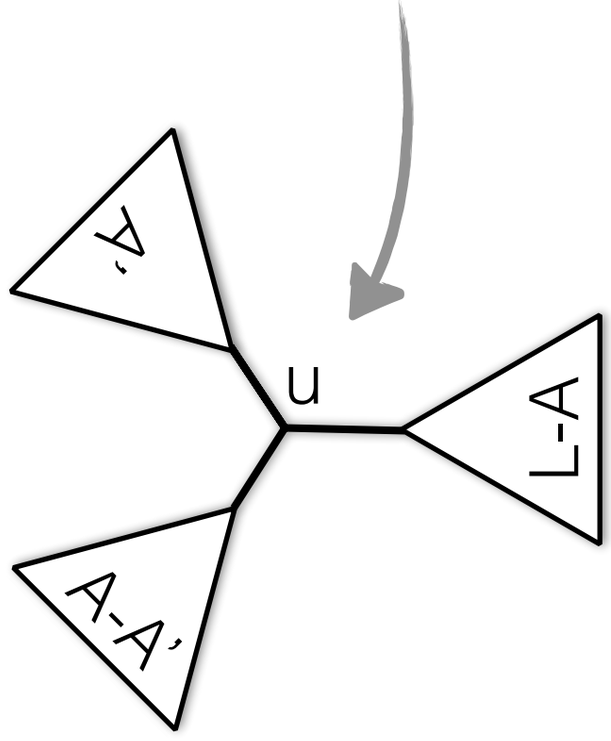
- Recursively break subsets of species into smaller subsets

- $w(u)$: Compare u against input gene trees and compute quartets from gene trees satisfied by u

Dynamic programming

$$S(\mathcal{A}) = \max_{\mathcal{A}' \subset \mathcal{A}} \{S(\mathcal{A}') + S(\mathcal{A} - \mathcal{A}') + w(\mathcal{A}' | \mathcal{A} - \mathcal{A}' | \mathcal{L} - \mathcal{A})\}$$



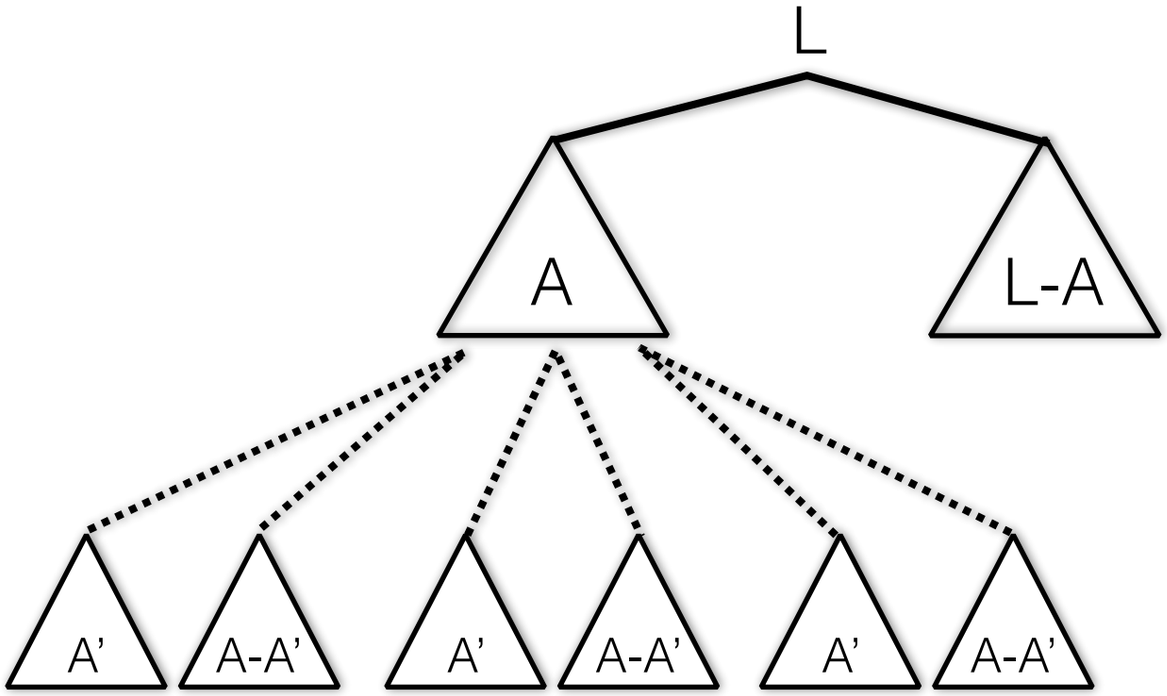


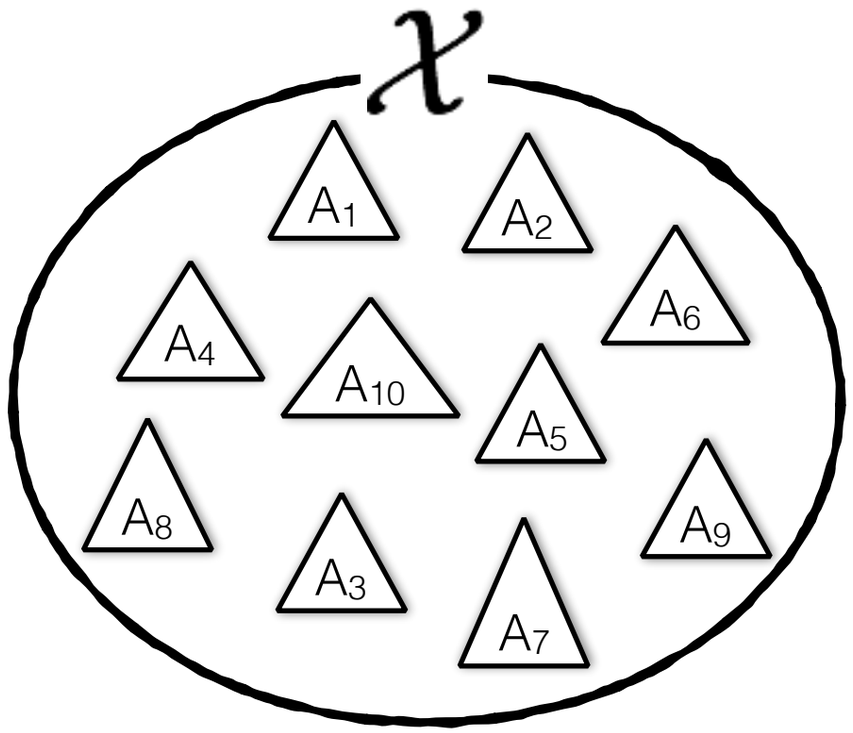
- Recursively break subsets of species into smaller subsets

- $w(u)$: Compare u against input gene trees and compute quartets from gene trees satisfied by u

Constrained version

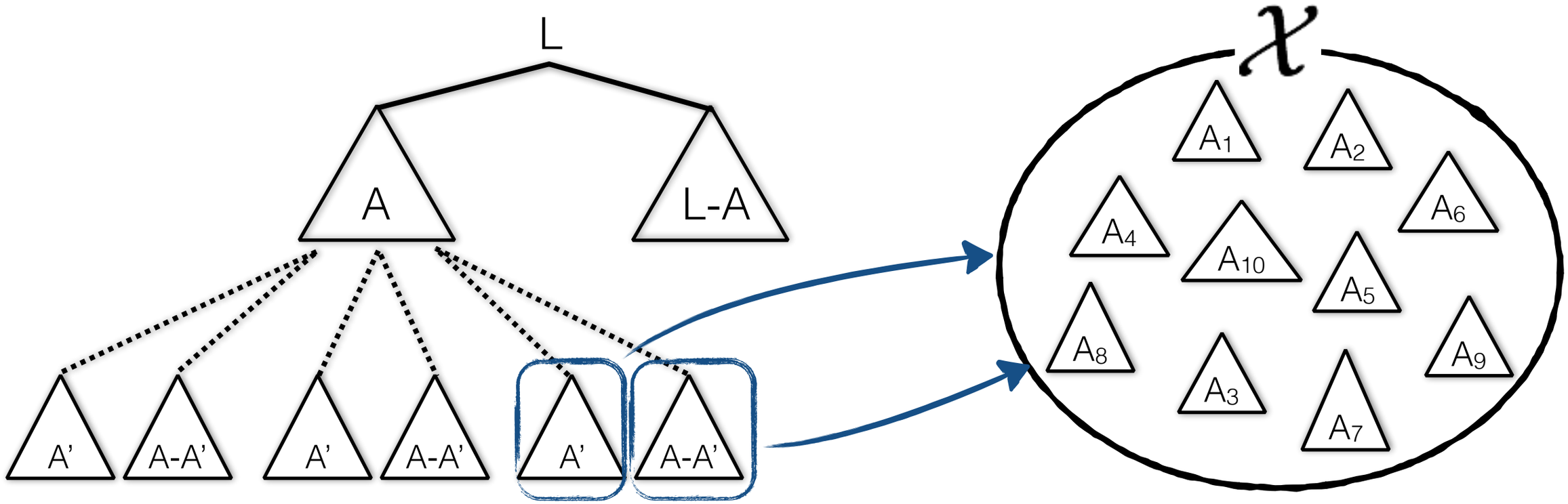
$$S(\mathcal{A}) = \max_{\{\mathcal{A}', \mathcal{A}-\mathcal{A}'\} \subset \mathcal{X}} \{S(\mathcal{A}') + S(\mathcal{A} - \mathcal{A}') + w(\mathcal{A}'|\mathcal{A} - \mathcal{A}'|\mathcal{L} - \mathcal{A})\}$$





Constrained version

$$S(\mathcal{A}) = \max_{\{\mathcal{A}', \mathcal{A}-\mathcal{A}'\} \subset \mathcal{X}} \{S(\mathcal{A}') + S(\mathcal{A} - \mathcal{A}') + w(\mathcal{A}'|\mathcal{A} - \mathcal{A}'|\mathcal{L} - \mathcal{A})\}$$



- Restrict “branches” in the species tree to a given constraint set $\mathcal{X}$.

Asymptotic running time?

- Simple discrete math question:

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

  - Clearly, $|\mathcal{Y}| < |\mathcal{X}|^2$

  - What's the maximum $|\mathcal{Y}|$ with respect to $|\mathcal{X}|$?

Asymptotic running time?

- Simple discrete math question:
  - $\mathcal{X}$ = a set of subsets of some set $\mathcal{L}$.
    $\mathcal{Y} = \{\ (a,b) \in \mathcal{X} \mid\ a \cap b = \varnothing\ ,\ a \cup b \in \mathcal{X}\ \}$
    - Clearly, $|\mathcal{Y}| < |\mathcal{X}|^2$
    - What's the maximum $|\mathcal{Y}|$ with respect to $|\mathcal{X}|$?
- Turns out to be rather challenging
  - Daniel Kane and Terence Tao proved: $|\mathcal{Y}| = O(|\mathcal{X}|^{1.73})$