ASTRAL-Pro: Quartet-Based Species-Tree Inference despite Paralogy
Introduction

- **The Problem:** Given a set of gene trees (multi-copy) which evolved under the Gene Duplication and Loss model (GDL), infer the corresponding species tree (single-copy).

- **Existing methods:** DupTree, MulRF, FastMulRFS, ASTRAL-multi, and more...

- **So why do we need ASTRAL-Pro?**
Gene Duplication and Loss

ASTRAL solves the Maximum Quartet Support Species Tree [MQSST] problem.

\[ T^* = \arg \max_{T: C(T) \in \mathcal{X}} \sum_{G \in \mathcal{G}} QS(G, T) \]

- \( C(T) \) = Set of bipartitions in tree \( T \)
- \( \mathcal{X} \) = Set of given bipartitions
- \( \mathcal{G} \) = Set of input gene trees
- \( QS(T_1, T_2) \) = Quartet similarity between \( T_1 \) and \( T_2 \), i.e. number of common quartets between \( T_1 \) and \( T_2 \)
ASTRAL

Anchors and Tripartition

Anchors u and v

Tripartition T(u) = A|B|C
For tripartitions $M = M_1|M_2|M_3$, $P = P_1|P_2|P_3$ and $I_{ij} = |M_i \cap P_j|$, 

$$QI(M, P) = \text{Number of shared quartets between trees containing } M \text{ and } P$$
where $M$ and $P$ are anchors.

$$= \sum_{i,j,k \in \{1,2,3\}} I_{i1}I_{j2}I_{k3}(I_{i1} + I_{j2} + I_{k3} - 3)$$

$W(P) =$ Number of quartets in gene trees where $P$ is an anchor

$$= \frac{1}{2} \sum_{G \in \mathcal{G}} \sum_{M \in \mathcal{P}(G)} QI(P, M)$$

$$T^* = \arg \max_{T:C(T) \in \mathcal{X}} \sum_{p \in I(T)} W(p)$$

ASTRAL solves this problem in polynomial time with dynamic programming.
Definitions

$\alpha_G$ maps the leaves of gene trees to the species the gene is sampled from.

\[\alpha_G(b_1) = \alpha_G(b_2) = B\]
\[\alpha_G(c_1) = \alpha_G(c_2) = C\]
**Definitions**

**Tagged Tree:** Rooted tree where each internal node is tagged as duplication or speciation node.
**Definitions**

\( SQ = \) Quartet \( Q \) where (1) \(|\alpha_G(Q)| = 4\), (2) LCA of at least 3 elements is a speciation.

**Anchor LCA**, \( \psi(Q) = \) LCA of anchors of quartet \( Q \)
Definitions

Equivalent SQs: \( Q_1 \sim Q_2 \iff \alpha_G(Q_1) = \alpha_G(Q_2) \land \psi_G(Q_1) = \psi_G(Q_2) \)
MLQST Problem

Per-locus Quartet score, \( q(S, G) = \) Number of equivalent classes in \( G \) induced in \( S \).

MLQST Problem: Find \( S^* \) such that,

\[
S^* = \arg \max_S \sum_{G \in G} p(S, G)
\]

- For singly-labeled trees, MLQST problem becomes MQSST problem.
- ASTRAL-Pro, an extension of ASTRAL can solve MLQST problem.
ASTRAL-Pro

Three main changes to ASTRAL-

1. When computing tripartitions, uses species labels.
2. Changes the calculation of $W(P)$
3. Sums over only speciation nodes when computing $W$
ASTRAL-Pro

For $M_w = M_1 | M_2 | M_3$, and $I_{ij} = |M_i \cap P_j|

\[ QI_{pro}(P, M_w) = \sum_{i,j,k \in \{1,2,3\}, j < k} \left( \frac{I_{1i}}{2} \right) I_{2j} I_{2k} + \sum_{i,j,k \in \{1,2,3\}} \frac{I_{1i} I_{2j} I_{3k} (I_{1i} + I_{2j} - 2)}{2} \]

\[ W_{pro}(P) = \sum_{G \in G} \sum_{w \in I(G)} QI_{pro}(P, M_w) \times 1_{speciation} \]

Changing the calculation of $W$ in ASTRAL solves the MLQST problem.
Tagging trees - For any node \( u \) with children \( u_1 \) and \( u_2 \), if \( \alpha_G(u_1) \) and \( \alpha_G(u_2) \) are disjoint sets, then \( u \) is tagged as speciation node. Otherwise, label \( u \) as duplication node.

Rooting trees - Set the internal node as root which has the least number of duplication and loss events in its corresponding tagged tree.

Default bipartition set - Singly labeled trees are sampled from multi-label trees and bipartitions are generated from sampled trees.

ASTRAL-Pro is statistically consistent under the GDL model (Arvestad et. al. 2009) assuming an accurate tagging on the input trees.
Performance Study
Overview of Experiment

● Methods Compared
  ○ DupTree: Solves Gene Tree Parsimony problem (GTP) with heuristic search.
  ○ MulRF: Solves the RFS-multree problem (RFS adapted for multrees) with heuristic search.
  ○ ASTRAL-multi: ASTRAL modified to handle multiple alleles.

● Datasets
  ○ Simulated Datasets (S25, S100):
    ■ Varying GDL rates
    ■ Varying the amounts of ILS
    ■ Varying both the number of species and gene trees
  ○ Biological Datasets:
    ■ 103 species plant dataset
    ■ 16 species fungal dataset
Dataset simulation

● Step 1: Generate true gene trees
  ○ Generated under DLCoal model, which combines ILS and GDL
● Step 2: Generate alignments from those true trees
● Step 3: Infer gene trees with error from alignments
Advantages of Utilizing Multi-copy gene trees

Dataset: S25

Gene Trees = 10,000

(200-900 single-copy across 50 replicas)

- Makes sense. More gene trees tends to mean more accuracy
Varying GDL

Dataset: S25
Species = 25
Gene Trees = 1000
Varying ILS

Dataset: S25

Species = 25

Gene Trees = 1000
Varying Gene & Species Tree Amounts

\( k = \text{Gene Trees} \)

\( n = \text{Species Trees} \)
S100 Summary

- Rankings are calculated by mean error over 10 replicates
- A-Pro clearly the best
- MulRF ends up being the second best

<table>
<thead>
<tr>
<th></th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>4th</th>
</tr>
</thead>
<tbody>
<tr>
<td>MulRF</td>
<td>42</td>
<td>67</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>DupTree</td>
<td>28</td>
<td>8</td>
<td>15</td>
<td>69</td>
</tr>
<tr>
<td>A-Pro</td>
<td>105</td>
<td>14</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ASTRAL-multi</td>
<td>12</td>
<td>14</td>
<td>71</td>
<td>23</td>
</tr>
</tbody>
</table>
Plant Dataset

- Previous study had inferred 424 single-copy genes; 9,683 multi-copy genes.
  - Had to throw out all the multi-copy genes as they didn’t have a method!
- All 9,683 multi-copy gene trees were analyzed in this experiment.
- DupTree returned clearly inaccurate tree
- A-pro’s tree for the most part agreed with the previous ASTRAL tree.
  - A-pro however, has higher localPP support.
Summary

- We saw the advantage of using multi-copy trees over merely single-copy trees.
- We saw that ASTRAL-Pro performed extremely well under all simulated conditions.
  - Under the vast majority of conditions it outperformed all other methods put up against it.
- We saw that ASTRAL-Pro performed well on biological data.
  - Comparable (and possibly better than) the ASTRAL tree inferred from single-copy gene trees.
Backup Slides
<table>
<thead>
<tr>
<th>Condition</th>
<th>Parameter Ranges</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default model</td>
<td>( n = 25; k = 1,000; \tau \sim LN(21.25; 0.2) )</td>
</tr>
<tr>
<td></td>
<td>( \hat{\lambda}<em>+ = 4.9 \times 10^{-10}; \hat{\lambda}</em>- = \hat{\lambda}_+; N_e = 4.7 \times 10^8 )</td>
</tr>
<tr>
<td></td>
<td>( C \approx 5; \text{ILS} \approx 70% )</td>
</tr>
<tr>
<td></td>
<td>MGTE = 15% (500 bp) or 36% (100 bp)</td>
</tr>
<tr>
<td>Varying ( \hat{\lambda}<em>+, \hat{\lambda}</em>- ) (DupLoss rate)</td>
<td>( \hat{\lambda}<em>+ \in {4.9, 2.7, 1.9, 0.52, 0} \times 10^{-10} ) ( \hat{\lambda}</em>- \in {1, 0.5, 0.1, 0} \times \hat{\lambda}_+ ); ( C \approx {5, 2, 1, 0.2, 0} )</td>
</tr>
<tr>
<td>Varying ( \hat{\lambda}_+, N_e ) (dup rate, ILS)</td>
<td>( \hat{\lambda}_+ \in {4.9, 1.9, 0} \times 10^{-10} ); ( N_e \in {4.7, 1.9, 0.48, 0.0001} \times 10^8 )</td>
</tr>
<tr>
<td></td>
<td>( \text{ILS} \approx {70, 52, 20, 0} % ); ( C \approx {5, 1, 0} )</td>
</tr>
<tr>
<td></td>
<td>MGTE ( \approx {15, 15, 15, 16} % ) (500 bp) or ( {36, 36, 36, 35} % ) (100 bp) as ( N_e ) changes</td>
</tr>
<tr>
<td>Varying ( n )</td>
<td>( n \in {10, 25, 100, 250, 500} )</td>
</tr>
<tr>
<td></td>
<td>MGTE ( \approx {15, 15, 17, 18, 18} % ) (500 bp)</td>
</tr>
<tr>
<td></td>
<td>or ( {34, 36, 40, 43, 43} % ) (100 bp)</td>
</tr>
<tr>
<td>Varying ( k )</td>
<td>( k \in {25, 100, 250, 1,000, 2, 500, 10,000} )</td>
</tr>
</tbody>
</table>