ASTRAL-II

Discussion and Criticism

Michael Nute

3/8/16

Summary of ASTRAL (original)

- ASTRAL-II Follows up on ASTRAL [1], which finds the species tree that maximizes the quartet-score using a constrained search.
  - **Quartet Score**: for every quartet in the species tree, for every gene tree in the data, add 1 if the topology from the species tree is the same as the one in the gene tree.
  - **Constraint**: searches *only* trees whose bipartitions appear in *at least* one gene tree.
- Pros:
  - Statistically consistent under the multi-species coalescent
  - Fast for large numbers of genes, taxa
- Cons:
  - Search space might not capture the actual species tree

ASTRAL-II Improvements

1. Improves speed with some shortcuts for counting quartet frequencies
2. Adds method for handling unresolved gene trees (polytomies)
3. Expands search space by adding bipartitions according to a heuristic using a similarity matrix:

$$[S]_{ij} = \sum_{g \in \mathcal{G}} \sum_{q \in \binom{n}{4}} \mathbb{I}\{g\big|_q = x_i x_j | **\}$$

   - I.e. the similarity of x_i and x_j is the count of the number of quartets with them together among all gene trees.
   - Adds all the bipartitions from the UPGMA tree on S.
   - Adds the closure of bipartitions from the original search space (may imply some that are not already in the set).

Issues with ASTRAL-II

- Doesn't consider how likely a bipartition is to have been missing from the search space
  - Not necessarily a drawback because the method is sensible and fast
  - Can still omit bipartitions that were likely to have been missed
- Could design a heuristic to choose bipartitions based on low-frequency quartets to find those that were probably in the species tree but never showed up in a gene tree.

Questions?

References:

[1] Mirarab, S., et al. (2014). ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics (Oxford, England)*, *30*(17), i541–8.

[2] Mirarab, S., & Warnow, T. (2015). ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics (Oxford, England)*, *31*(12), i44–52.