Fast, accurate species tree estimation with ASTRID

Pranjal Vachaspati
ASTRID estimates species trees

**INPUT:** Gene trees
**OUTPUT:** Species tree computed from internode distance matrices

Similar approach to NJst (Liu & Yu, 2011)
Internode distance matrices

Gene Trees > Internode Distance Matrices > Average Matrix > Species Tree

A B C D E F
A - 3 4 4 - 2
B - - 3 3 - 3
C - - - 2 - 4
D - - - - 4
E - - - - - -
F - - - - - -
Distance methods for species tree estimation

Distance methods take a distance matrix as input, and output a species tree.

NJst (predecessor to ASTRID) used neighbor-joining.

ASTRID normally uses FastME-2 (Lefort, Desper & Gascuel 2015)
  - Faster than neighbor-joining
  - More accurate

Can also use other methods (BIONJ*, RapidNJ, UPGMA*) in certain cases.
Evaluating ASTRID with simulated data

Model Species Tree | True Gene Trees | Alignments | Estimated Gene Trees | Est. Species Tree

Compare
High accuracy on datasets with ILS

47-taxon simulated data based on avian biological dataset

High ILS (47% AD - average distance between true gene trees and true species tree)
High accuracy on datasets with ILS

Mammalian simulated dataset

- 37-taxon simulated data based on avian biological dataset
- Moderate ILS (29% AD - average distance between true gene trees and true species tree)
ASTRID is extremely fast

Simulated ILS dataset from ASTRAL-II paper (Mirarab & Warnow, 2015)

1000 gene trees, high ILS

Completes in just 30 seconds!

Competing methods much slower: ASTRAL 5.7.3 took 90 seconds for 100 taxa, 2 hours for 1000 taxa
ASTRID can scale to extremely large datasets

43,183 taxon supertree dataset based on RNAsim simulation

ASTRID completed in 5 hours 42 minutes (FastME+NNIs as distance method)

Running time dominated by distance method

- with RapidNJ, only 25 minutes! (but worse accuracy)

High RAM requirements - 132 GB for FastME analysis.

ASTRAL could not run (ran out of memory); MRL took over 24 hours
Getting ASTRID: github.com/pranjalv123/ASTRID

Available for Linux, Mac, and Windows

GPL-licensed
Using ASTRID

Easy-to-use command line interface, similar to ASTRAL:

ASTRID -i <input tree file> -o <output tree file>

Usually, this is sufficient!
Summary

ASTRID: accurate, extremely fast species tree estimation

Available at github.com/pranjalv123/ASTRID

Join the ASTRID user group at

https://groups.google.com/d/forum/astrid-users or astrid-users@googlegroups.com
Using ASTRID: multi-individual datasets

Create a file mapping individual names to species names:

<file:species_mapping.txt>
species1:indiv1,indiv2,indiv3
species2:indiv4,indiv5
...

ASTRID -i input.trees -o output.trees -a species_mapping.txt