ASTRID
Accurate Species TRrees from Internode Distances

Pranjal Vachaspati and Tandy Warnow

University of Illinois at Urbana-Champaign

Support: Roy J. Carver graduate fellowship, UIUC (PV), NSF DBI-1461364 (TW), Grainger Foundation (PV, TW).
The phylogenomic pipeline

Recover **species trees** from **whole genomes**

Sequences → Alignments → Gene Trees → Species Tree
The phylogenomic pipeline

Recover **species trees** from **whole genomes**

- **Sequences**
  - ATCG
  - TTTG
  - AGTA
  - GGTAA
  - AATC
  - AGTGA
  - GTCA
  - TTGG
  - CCATA

- **Alignments**
  - A--T-CG
  - CT--T-G
  - AC--T--A
  - GGT--A
  - AA--T--C
  - AGT--GA
  - G-T--C
  - TC--TGG
  - CCAT--A

- **Gene Trees**

- **Species Tree**
The phylogenomic pipeline

Recover **species trees** from **whole genomes**

Sequences → Alignments → Gene Trees → Species Tree

- **ATCG**
- **TTTG**
- **AGTA**

- **GGTTA**
- **AATC**
- **AGTGA**

- **GTG**
- **TTGG**
- **CCATA**

**MSA** (PASTA, MuSCLE, CLUSTAL, etc.) → **Gene Trees**

**ML estimator** (RAxML, FastTree, etc.)
Species tree estimation
Genome-scale data is getting really big

(Su 2013, sulab.org)
Incomplete lineage sorting (ILS) causes gene tree discordance
We want fast, accurate, statistically consistent species tree estimators

The probability of reconstructing the species tree goes to 1
How do we know how good our methods are in practice?
Simulations help assess method accuracy

Species Tree (random or biological)

SimPhy

Gene Trees

Indelible

Alignments
Simulations help assess method accuracy
Compare trees by counting different branches

0.5 Robinson-Foulds (RF) topological error
Compare trees by counting different branches

\[(\{A,B\},\{C,D,E\})\]
\[(\{A,B,C\},\{D,E\})\]
\[(\{A,B\},\{C,D,E\})\]
\[(\{A,B,D\},\{C,E\})\]

0.5 Robinson-Foulds (RF) topological error
Species trees can be estimated in several ways

<table>
<thead>
<tr>
<th>Class</th>
<th>Examples</th>
<th>Stat. Cons?</th>
<th>Scalable?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concatenated Maximum Likelihood (CA-ML)</td>
<td>RAxML, FastTree, IQTree</td>
<td>No (unpartitioned)</td>
<td>Moderate</td>
</tr>
<tr>
<td>Coestimation of gene trees and species trees</td>
<td>*BEAST, BEST</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Summary methods</td>
<td>ASTRAL, MP-EST, NJst</td>
<td>Yes</td>
<td>Varies</td>
</tr>
</tbody>
</table>
Species trees can be estimated in several ways

<table>
<thead>
<tr>
<th>Class</th>
<th>Examples</th>
<th>Stat. Cons?</th>
<th>Scalable?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concatenated Maximum Likelihood (CA-ML)</td>
<td>RAxML, FastTree, IQTree</td>
<td>No (unpartitioned) Open (partitioned)</td>
<td>Moderate</td>
</tr>
<tr>
<td>Coestimation of gene trees and species trees</td>
<td>*BEAST, BEST</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Summary methods</td>
<td>ASTRAL, MP-EST, NJst</td>
<td>Yes</td>
<td>Varies</td>
</tr>
</tbody>
</table>
Species trees can be estimated in several ways

<table>
<thead>
<tr>
<th>Class</th>
<th>Examples</th>
<th>Stat. Cons?</th>
<th>Scalable?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concatenated Maximum Likelihood (CA-ML)</td>
<td>RAxML</td>
<td>No (unpartitioned)</td>
<td>Moderate</td>
</tr>
<tr>
<td></td>
<td>FastTree</td>
<td>Open (partitioned)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>IQTree</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coestimation of gene trees and species</td>
<td>*BEAST</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>trees</td>
<td>BEST</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Summary methods</td>
<td>ASTRAL</td>
<td>Yes</td>
<td>Varies</td>
</tr>
<tr>
<td></td>
<td>MP-EST</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>NJst</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Species trees can be estimated in several ways

<table>
<thead>
<tr>
<th>Class</th>
<th>Examples</th>
<th>Stat. Cons?</th>
<th>Scalable?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concatenated Maximum Likelihood (CA-ML)</td>
<td>RAxML, FastTree, IQTree</td>
<td>No (unpartitioned) Open (partitioned)</td>
<td>Moderate</td>
</tr>
<tr>
<td>Coestimation of gene trees and species trees</td>
<td>*BEAST, BEST</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Summary methods</td>
<td>ASTRAL, MP-EST, NJst</td>
<td>Yes</td>
<td>Varies</td>
</tr>
</tbody>
</table>
Pros and cons of NJst (Liu & Yu, 2011)

- ONLY method other than ASTRAL-2 that runs on large datasets
- Better accuracy than MP-EST (previous leading method)
- Straightforward implementation
- Statistically consistent
- **Drawbacks:**
  - Slow
  - Works poorly in presence of incomplete gene trees
  - Difficult to use

- ASTRID fixes these problems, and more
ASTRAL, NJst have good accuracy on large datasets.

- MP-EST does not run on more than 100 taxa.
  (Mirarab et al., Bioinformatics 2015)
ASTRAL, NJst can handle large datasets

1000 genes, “medium” levels of recent ILS

(Mirarab et al., Bioinformatics 2015)
NJst uses neighbor joining to compute species trees.

Gene Trees → Internode Distance Matrices → Average Internode Distance Matrix → NJ Species Tree
ASTRID uses a variety of distance-based methods to compute species trees.
Better distance-based tree estimation methods

- Distance-based methods estimate trees from distance matrices
- Neighbor Joining (NJ) is statistically consistent, but improved methods have been developed:
  - WEIGHBOR, BIONJ, MVR (minimum variance reduction)
    - Estimates of short distances are more reliable than long distances
    - PhyD* (Criscolo and Gascuel, 2008) provides these methods even when distance matrix is missing elements
  - FastME (Desper and Gascuel, 2002) finds an approximate minimum-evolution tree in $O(n^2)$ time
    - Substantially faster than NJ, which is an $O(n^3)$ algorithm
Questions

- What is the best distance method for ASTRID?
- How does ASTRID’s performance compare to other phylogenetic methods?
- Is ASTRID robust to short sequences?
- Does ASTRID have good accuracy under high levels of ILS?
- Can ASTRID handle cases where each gene tree is on a subset of the taxa?
- How fast is ASTRID?
Questions

- What is the best distance method for ASTRID?
- How does ASTRID’s performance compare to other phylogenetic methods?
- Is ASTRID robust to short sequences?
- Does ASTRID have good accuracy under high levels of ILS?
- Can ASTRID handle cases where each gene tree is on a subset of the taxa?
- How fast is ASTRID?
Questions

- What is the best distance method for ASTRID?
- How does ASTRID’s performance compare to other phylogenetic methods?
- Is ASTRID robust to short sequences?
- Does ASTRID have good accuracy under high levels of ILS?
- Can ASTRID handle cases where each gene tree is on a subset of the taxa?
- How fast is ASTRID?
Questions

- What is the best distance method for ASTRID?
- How does ASTRID’s performance compare to other phylogenetic methods?
- Is ASTRID robust to short sequences?
- Does ASTRID have good accuracy under high levels of ILS?
- Can ASTRID handle cases where each gene tree is on a subset of the taxa?
- How fast is ASTRID?
Questions

- What is the best distance method for ASTRID?
- How does ASTRID’s performance compare to other phylogenetic methods?
- Is ASTRID robust to short sequences?
- Does ASTRID have good accuracy under high levels of ILS?
- Can ASTRID handle cases where each gene tree is on a subset of the taxa?
- How fast is ASTRID?
Questions

- What is the best distance method for ASTRID?
- How does ASTRID’s performance compare to other phylogenetic methods?
- Is ASTRID robust to short sequences?
- Does ASTRID have good accuracy under high levels of ILS?
- Can ASTRID handle cases where each gene tree is on a subset of the taxa?
- How fast is ASTRID?
Experimental Datasets (from prior publications)

<table>
<thead>
<tr>
<th>Dataset</th>
<th># taxa</th>
<th># genes</th>
<th>AD%</th>
<th># sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mammalian Sim¹²</td>
<td>37</td>
<td>200</td>
<td>21-50</td>
<td>250-1000</td>
</tr>
<tr>
<td>Avian Sim³ ⁴</td>
<td>48</td>
<td>1000</td>
<td>29-60</td>
<td>250-1500</td>
</tr>
<tr>
<td>ASTRAL-2 Sim</td>
<td>50-1000</td>
<td>1000</td>
<td>9-69</td>
<td>300-1500</td>
</tr>
</tbody>
</table>

¹Generated by Mirarab et al., Bioinformatics 2014
²Based on biological data from Song et al., PNAS 2012
³Generated by Mirarab et al., Science 2014
⁴Based on biological dataset from Jarvis, Mirarab et al., Science 2014
Use FastME for complete distance matrices
Use BioNJ* for incomplete matrices

FastME doesn’t work on incomplete matrices.

Avian Dataset
48 taxa, complete gene trees

ASTRID-bionj
ASTRID-fastme

Number of genes
RF Topological Error
Both ASTRAL and ASTRID are substantially more accurate than MP-EST
Both ASTRAL and ASTRID are substantially more accurate than MP-EST
ASTRID is slightly more accurate than ASTRAL on some datasets...

![Graph showing the comparison between ASTRAL-2 and ASTRID-fastme for RF Topological Error in Avian Simulated (48 taxa) datasets. The x-axis represents the number of genes, ranging from 10 to 800, while the y-axis shows the RF Topological Error, ranging from 0.45 to 0.00. The graph shows a downward trend as the number of genes increases.]
...about the same on others...
...and slightly less accurate on some.

[Graph showing RF topological error with ASTRAL-2 and ASTRID-fastme compared to different numbers of genes and speciation rates.]
ASTRID is very fast

48-taxon avian simulated dataset

Running time (s)

Number of genes

NJst

ASTRID-fastme
ASTRID is very fast

48-taxon avian simulated dataset

Running time (s)

NJst
ASTRAL-2
ASTRID-fastme

Number of genes

0 200 400 600 800

Running time (s)

0 50 100 150 200 250 300 350

Number of genes

0 200 400 600 800
ASTRID is a lot faster than ASTRAL on large datasets.

1000 genes and 1000 taxa:
ASTRAL takes 12 hours, ASTRID-FastME takes 0.5 hours.
Answers

- What is the best distance method for ASTRID? FastME
- How does ASTRID’s performance compare to other phylogenetic methods? Comparable or better
- Is ASTRID robust to short sequences? Not as much as ASTRAL
- Does ASTRID have good accuracy under high levels of ILS? Yes
- Can ASTRID handle cases where each gene tree is on a subset of the taxa? Yes (see paper)
- How fast is ASTRID? Extremely fast
What is the best distance method for ASTRID? **FastME**

How does ASTRID’s performance compare to other phylogenetic methods? **Comparable or better**

Is ASTRID robust to short sequences? **Not as much as ASTRAL**

Does ASTRID have good accuracy under high levels of ILS? **Yes**

Can ASTRID handle cases where each gene tree is on a subset of the taxa? **Yes (see paper)**

How fast is ASTRID? **Extremely fast**
Answers

- What is the best distance method for ASTRID? FastME
- How does ASTRID’s performance compare to other phylogenetic methods? Comparable or better
- Is ASTRID robust to short sequences? Not as much as ASTRAL
- Does ASTRID have good accuracy under high levels of ILS? Yes
- Can ASTRID handle cases where each gene tree is on a subset of the taxa? Yes (see paper)
- How fast is ASTRID? Extremely fast
Answers

- What is the best distance method for ASTRID? FastME
- How does ASTRID’s performance compare to other phylogenetic methods? Comparable or better
- Is ASTRID robust to short sequences? Not as much as ASTRAL
- Does ASTRID have good accuracy under high levels of ILS? Yes
- Can ASTRID handle cases where each gene tree is on a subset of the taxa? Yes (see paper)
- How fast is ASTRID? Extremely fast
Answers

- What is the best distance method for ASTRID? FastME
- How does ASTRID’s performance compare to other phylogenetic methods? Comparable or better
- Is ASTRID robust to short sequences? Not as much as ASTRAL
- Does ASTRID have good accuracy under high levels of ILS? Yes
- Can ASTRID handle cases where each gene tree is on a subset of the taxa? Yes (see paper)
- How fast is ASTRID? Extremely fast
Answers

- What is the best distance method for ASTRID? FastME
- How does ASTRID’s performance compare to other phylogenetic methods? Comparable or better
- Is ASTRID robust to short sequences? Not as much as ASTRAL
- Does ASTRID have good accuracy under high levels of ILS? Yes
- Can ASTRID handle cases where each gene tree is on a subset of the taxa? Yes (see paper)
- How fast is ASTRID? Extremely fast
ASTRID - fast and accurate

- Construct species trees from input gene trees
- Flexibility in choice of distance methods
- Statistically consistent
- ASTRID and ASTRAL are the fastest and most accurate coalescent-based methods
- Linear runtime scaling in number of gene trees
- Orders of magnitude faster than competing methods on biological datasets
- Easy to use

http://www.github.com/pranjalv123/ASTRID