ASTRAL Tutorial

• Instructor: Siavash Mirarab, smirarab@ucsd.edu

• Github site: https://github.com/smirarab/ASTRAL

• Email: astral-users@googlegroups.com
Gene tree discordance

The species tree

A gene tree
Gene tree discordance

Causes of gene tree discordance include:

• Duplication and loss
• Horizontal Gene Transfer (HGT) and Hybridization
• Incomplete Lineage Sorting (ILS)
Gene evolution model (MSC)

Sequence evolution model (GTR)

ACTGCACACCG
ACTGC--CCCCG
AATGC--CCCCG
--CTGCACACCGG

CTGAGCATCG
CTGAGC--TCG
ATGAGC--TC-
CTGA--CAC--G

AGCAGCATCGTG
AGCAGC--TCGTG
AGCAGC--TC--TG
C--TA--CACGGTG

CAGGCACGCACGAA
AGC--CAGGC--CAT
ATGGCAGGC--C--TA
AGCTAC--CAGGGAT
Two-step approach

Step 1: infer gene trees (e.g., ML)
Two-step approach

Step 1: infer gene trees (e.g., ML)

Step 2: gene tree summarization

Gene tree

Chimp  Human  Orang.  Gorilla

Gene tree

Chimp  Gorilla  Human  Orang.

Gene tree

Chimp  Orang.  Human  Gorilla

Gene tree

Chimp  Human  Orang.  Gorilla

ACTGCACACCCG  CTGAGCATCG  AGCAGCATCGTG  CAGGCACGCACGAA
ACTGC–CCCCCG  CTGAGC–TCG  AGCAGC–TGCTG  AGC–CACGC–CATA
ASTRAL

- **Input:** Unrooted gene trees
  - Can have missing data
  - Can have polytomies
  - Can have multiple individuals per species

- **Output:** The estimated unrooted species tree
  - Will have branch lengths in coalescent units on internal branches (not super accurate)
  - Will have a measure of support called localPP
A bit on the input
Typical pipeline to produce input

1. Find **orthologous** parts of the genomes for species of interest

2. **Align** sequences per ortholog using your favorite MSA method (e.g., PASTA, UPP, MAFFT, etc.)

3. Infer **“gene trees”** per ortholog using Maximum Likelihood methods (e.g., RAxML or FastTree)

Optionally: perform various **filtering** steps to remove errors
Should you filter?

• Filtering genes based on missing data?

  • Generally not beneficial [Molloy and Warnow, 2018]
Should you filter?

- Filtering genes based on missing data?
  - Generally not beneficial [Molloy and Warnow, 2018]

- Filtering genes based on gene tree estimation error?
  - Depends on conditions [Molloy and Warnow, 2018]
Should you filter?

- Filtering genes based on missing data?
  - Generally not beneficial [Molloy and Warnow, 2018]

- Filtering genes based on gene tree estimation error?
  - Depends on conditions [Molloy and Warnow, 2018]

- Filtering fragmentary sequences while keeping the gene?
  - Often beneficial [Sayyari, Whitfield, and Mirarab, 2018]
Should you filter?

• Filtering genes based on missing data?
  • Generally not beneficial [Molloy and Warnow, 2018]

• Filtering genes based on gene tree estimation error?
  • Depends on conditions [Molloy and Warnow, 2018]

• Filtering fragmentary sequences while keeping the gene?
  • Often beneficial [Sayyari, Whitfield, and Mirarab, 2018]

• Filtering super long branches helps (talk today by Uyen Mai)
Multi-locus Bootstrapping (MLBS) is possible.


- You need a file with bootstrap replicates for each gene tree.

- MLBS not suggested as seem to degrade accuracy compared to simply using ML gene trees.

Bioinformatics, 2014, Mirarab et al.
Low support branches

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

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 support branches**
  
  • 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 support branches**
  - 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

BMC Bioinformatics, 2018, Zhang et al.
A bit on the output
For a quartet (4 species), the most probable unrooted quartet tree (among the gene trees) is the unrooted species tree topology (Allman, et al. 2010)
Unrooted quartets under MSC model

For a quartet (4 species), the most probable unrooted quartet tree (among the gene trees) is the unrooted species tree topology (Allman, et al. 2010)

The most frequent gene tree = The most likely species tree
For 5 or more species, the unrooted species tree topology can be different from the most probable gene tree (called “anomaly zone”) (Degnan, 2013)
For 5 or more species, the unrooted species tree topology can be different from the most probable 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])
For 5 or more species, the unrooted species tree topology can be different from the most probable gene tree (called “anomaly zone”) (Degnan, 2013).

Alternative:
weight all $3\binom{n}{4}$ quartet topologies by their frequency and find the optimal tree.
Maximum Quartet Support Species Tree

\[
Score(T) = \sum_{1}^{k} |Q(T) \cup Q(t_i)|
\]

- Optimization problem (NP-Hard):

Find the species tree with the maximum number of induced quartet trees shared with the collection of input gene trees.
Maximum Quartet Support Species Tree

\[ \text{Score}(T) = \sum_{1}^{k} | Q(T) \cup Q(t_i) | \]

- Optimization problem (NP-Hard):
  
  Find the species tree with the maximum number of induced quartet trees shared with the collection of input gene trees

- Statistically consistent under the multi-species coalescent model when solved exactly
Maximum Quartet Support Species Tree

\[ \text{Score}(T) = \sum_{1}^{k} | Q(T) \cup Q(t_i) | \]

- Optimization problem (NP-Hard):
  
  Find the species tree with the maximum number of induced quartet trees shared with the collection of input gene trees

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

- ASTRAL: an exact solution using dynamic programming a constrained version that can run on large datasets
Beyond topology, ASTRAL can …

• estimate length of internal branches in coalescent units: \( \frac{\text{# generations}}{\text{population size}} \)
Beyond topology, ASTRAL can ... 

- estimate length of internal branches in coalescent units: \# generations / population size

- estimate a measure of branch support called local posterior probability

- The probability of the branch being correct assuming gene trees are generated by the MSC model
Beyond topology, ASTRAL can …

• estimate **length** of internal branches in **coalescent units**: \# generations / population size

• estimate a measure of branch **support** called **local posterior probability**

  • The probability of the branch being correct assuming gene trees are generated by the MSC model

• test for **polytomies**
Beyond topology, ASTRAL can …

• estimate **length** of internal branches in **coalescent units**: # generations / population size

• estimate a measure of branch **support** called **local posterior probability**

  • The probability of the branch being correct assuming gene trees are generated by the MSC model

• test for **polytomies**

• annotate species tree branches with quartet-based measures of discordance
A bit on the search space
ASTRAL versions

• ASTRAL-I (<v. 4.7.3) restricts the search space to combinations of bipartitions seen in gene trees.
  
• This make it fast but it remains statistically consistent
ASTRAL versions

• ASTRAL-I (<v. 4.7.3) restricts the search space to combinations of bipartitions seen in gene trees.
  • This make it fast but it remains statistically consistent
• ASTRAL-II (<v. 5.1.0) increased the search space heuristically
  • Improved the accuracy at the expense of running time
  • Can handle polytomies in input gene trees
ASTRAL versions

- ASTRAL-I (≤v. 4.7.3) restricts the search space to combinations of bipartitions seen in gene trees.
  - This makes it fast but it remains statistically consistent.
- ASTRAL-II (≤v. 5.1.0) increased the search space heuristically.
  - Improved the accuracy at the expense of running time.
  - Can handle polytomies in input gene trees.
- ASTRAL-III (>v. 5.1.1) changed the search space again for a better running time versus accuracy trade-off.
  - Improved running time for unresolved trees; makes it feasible to remove very low support branches from gene trees.
Modifying the Search Space

- For small enough datasets, you can run ASTRAL in *exact* mode, which will find the globally optimal tree! (\(-x\) option)

- ASTRAL’s default usage will not run in exact mode, and will constrain the search space largely using the gene trees.

- It may be beneficial to *expand* the search space, using the “\(-e\)” or “\(-f\)” options (see tutorial at GitHub site). In particular, you can add species trees you’ve estimated using other methods, such as concatenation, SVDquartets, or species trees that other studies have suggested.
A bit on the running time
Empirical running time as a function of the number of genes

Avian simulations:
- \( n = 48 \) species
- \( k = 256 \) to \( 16,384 \) gene trees
- Relatively high ILS
- Low/moderate gene tree error
- 4 replicates per dot

Empirical running time seems to increase close to \( O(k^2) \)
Empirical running time as a function of the number of species

Simphy simulations:
- $n = 200$ to $10,000$ species
- $k = 1000$ gene trees
- Varying ILS
- Varying gene tree error
- 20 replicates everywhere, except for 10000 that has only 2 replicates

Empirical running time seems to increase close to $O(n^{1.8})$
Parallelization and scaling limits

- We have now developed ASTRAL-MP to use parallelism.
- Can analyze datasets with 10,000 species and 1000 genes in less than a day given 24 cores and a GPU
- [https://github.com/smirarab/ASTRAL/tree/MP-similarity](https://github.com/smirarab/ASTRAL/tree/MP-similarity)
Examples
Simple run

java -jar ~/workspace/ASTRAL/astral.5.6.1.jar -i 1KP-genetrees.tre -o 1KP-speciestrees.tre 2> 1KP-log.txt

- Runs ASTRAL version 5.6.1 on 1KP-genetrees.tre and save results in 1KP-speciestrees.tre with logs saved to 1KP-log.txt.

Open the tree in FigTree

- ASTRAL trees are unrooted.
- You will have to root them.
- You may have to open 2-3 times in FigTree (not sure why)

Length in coalescent units
• Often better to look at cladogram
By default, nodes labelled by localPP [Sayyari and Mirarab, 2016]
The log *(1KP-log.txt)* includes ...

This is ASTRAL version 5.6.1
424 trees read from 1KP-genetrees.tre
...
Number of taxa: 103 (103 species)
...
Taxon occupancy: {Pseudolycopodiella_caroliniana=282, Kadsura_heteroclita=234, Entransia_fimbriata=294, ...}
...
Number of Clusters after addition by greedy: 11043
...
Final quartet score is: 339023690
Final normalized quartet score is: 0.8946722590876793
...
ASTRAL finished in 36.372 secs
Annotate (score) a given tree and compute quartet support

```
java -jar ~/workspace/ASTRAL/astral.5.6.1.jar
  -i 1KP-genetrees.tre -q 1KP-speciestrees.tre
  -o 1KP-speciestrees-qs.tre -t 1
```

- Scores `1KP-speciestrees.tre` based on the `1KP-genetrees.tre` and annotates branches with the quartet score of the branch (`-t 1`), saving results into `1KP-speciestrees-qs.tre`.

- Check out other annotation options at [https://github.com/smirarab/ASTRAL/blob/master/astral-tutorial.md#extensive-branch-annotations](https://github.com/smirarab/ASTRAL/blob/master/astral-tutorial.md#extensive-branch-annotations)
Quartet scores as branch-specific estimate of gene tree discordance
Annotate (score) a given tree and compute quartet support

```
java -jar ~/workspace/ASTRAL/astral.5.6.1.jar
   -i 1KP-genetrees.tre -q 1KP-speciestrees.tre
   -o 1KP-speciestrees-qsall.tre -t 8
```

- Scores `1KP-speciestrees.tre` based on the `1KP-genetrees.tre` and annotates branches with the quartet score of the branch and the two alternative resolutions (`-t 8`), saving results into `1KP-speciestrees-qsall.tre`. 
Quartet scores of alternative resolutions

Quartet scores for all three resolutions around the branch. Example: 52%, 21%, 27%
Quartet scores of alternative resolutions

Example: 52%, 21%, 27%

Hard to read
Discovista: visualizing discordance

- [https://github.com/esayyari/DiscoVista](https://github.com/esayyari/DiscoVista)
Discovista: visualizing discordance

- https://github.com/esayyari/DiscoVista

Handling multiple individuals

• Handling multiple individuals is supported (the relevant paper has been in review for a long time)

• The mapping between names in gene trees and names in the species tree should be provided using the \(-a\) option. Two formats are supported (either can be used):

  Format 1:
  
  cat: siamesecat12,persiancat17,wildcat
dog: dog1,labrador,bulldog-1,bulldog-2
horse: pony,shire-2,mustang110

  Format 2:
  
  cat 3 siamesecat12 persiancat17 wildcat
dog 4 dog1 labrador bulldog-1 bulldog-2
horse 3 pony shire-2 mustang110
Main publications


Some other ASTRAL-related papers

- Testing for polytomies in species trees using quartet frequencies (Sayyari and Mirarab). Genes (2018) (Note: Implemented in option `-t 10`)
- Filtering loci is not beneficial! (Molloy and Warnow), Systematic Biology (2018)
- ASTRAL consistent under models of missing data (Nute, Molloy, Chou, and Warnow). BMC Genomics (2018)
- Fragmentary sequences can negatively impact ASTRAL trees (Sayyari, Whitfield, and Mirarab). Molecular Biology and Evolution (2017)
- How many genes does ASTRAL need? (Shekhar, Roch, and Mirarab). Transactions on Computational Biology and Bioinformatics (2017)
- Using ASTRAL as a supertree method (Vachaspati and Warnow). Bioinformatics (2017)
For more info

• Contact us:
  Tandy Warnow, warnow@illinois.edu,
  Siavash Mirarab, smirarabbaygi@eng.ucsd.edu

• Software available at Github site:
  https://github.com/smirarab/ASTRAL

• See tutorial and README at GitHub site:
  https://github.com/smirarab/ASTRAL/blob/master/astral-tutorial.md

• Email: astral-users@googlegroups.com

• More related papers at
  http://tandy.cs.illinois.edu/papers-all.html and
  http://eceweb.ucsd.edu/~smirarab/publications.html