[some]
Methodological advances since the first avian phylogenomics project
[mostly our work]

Siavash Mirarab
University of California, San Diego

TENT tree from Jarvis et al, Science, 2014
Challenges
Challenges

• Errors and incompleteness in data due to annotation, assembly, or other unknown origins
Challenges

• Errors and incompleteness in data due to annotation, assembly, or other unknown origins

• Models of sequence evolution
Challenges

• Errors and incompleteness in data due to annotation, assembly, or other unknown origins

• Models of sequence evolution

• Gene tree discordance
  • True discordance
  • Spurious discordance
Challenges

• Errors and incompleteness in data due to annotation, assembly, or other unknown origins

• Models of sequence evolution

• Gene tree discordance
  • True discordance
  • Spurious discordance

• Scalability
Challenges

• Errors and incompleteness in data due to annotation, assembly, or other unknown origins

• Models of sequence evolution

  • Gene tree discordance

    • True discordance

    • Spurious discordance

• Scalability
Gene tree discordance
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)**
Incomplete Lineage Sorting (ILS)

- Can occur when multiple alleles of a gene persist (fail to coalesce) during the lifetime of an ancestral population

“gene” here simply refers to a recombination-free part of the genome
Incomplete Lineage Sorting (ILS)

• Can occur when multiple alleles of a gene persist (fail to coalesce) during the lifetime of an ancestral population

“gene” here simply refers to a recombination-free part of the genome
Incomplete Lineage Sorting (ILS)

- Can occur when multiple alleles of a gene persist (fail to coalesce) during the lifetime of an ancestral population

- Always possible. Likely for:
  - Short branches (# generations)
  - Large populations

“gene” here simply refers to a recombination-free part of the genome
Incomplete Lineage Sorting (ILS)

• Can occur when multiple alleles of a gene persist (fail to coalesce) during the lifetime of an ancestral population

• Always possible. Likely for:
  • Short branches (# generations)
  • Large populations
  • Both characterize rapid radiations

“gene” here simply refers to a recombination-free part of the genome
Gene evolution model

Sequence evolution model

Species tree

Gene tree

Sequence data (Alignments)

ACTGACACCCG
ACTGC–CCCCCG
AATGC–CCCCCG
–CTGACACCGG

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

AGCAGCATCGTG
AGCAGC–TCGTG
AGCAGC–TC–TG
CTA–CACGGTG

CAGGCACGCACGAA
AGC–CACGC–CATA
ATGGCACGC–C–TA
AGCTAC–CACGGAT

Gene tree

Gene tree

Gene tree

Gene tree

Gene tree
Gene evolution model

Sequence evolution model

Species tree

- Gorilla
- Human
- Chimp
- Orangutan

Gene tree

- Chimp
- Human
- Orang.

Gene tree

- Gorilla
- Human
- Chimp

Gene tree

- Chimp
- Human
- Orang.

Gene tree

- Chimp
- Human
- Orang.

Sequence data (Alignments)

- ACTGCACACCG
- ACTGC–CCCCCG
- AATGC–CCCCCG
- –CTGCAcacacgg

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

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

- CAGGCACGCACGAA
- AGC–CACGC–CATA
- ATGGCACGC–C–TA
- AGCTAC–CACGGAT
Multi-species coalescent (MSC) model

- A statistical gene tree evolution model for ILS
  [Pamilo and Nei, 1988] [Rannala and Yang, 2003]
- Does not model recombination within a gene
Multi-species coalescent (MSC) model

- A statistical gene tree evolution model for ILS
  [Pamilo and Nei, 1988] [Rannala and Yang, 2003]
- Does not model recombination within a gene
- In theory, we can infer the species tree given a large randomly distributed sample of recombination-free, reticulation-free, orthologous, error-free gene trees
Gene evolution model

Step 1: infer gene trees (traditional methods)

ACTGCACACCG
ACTGC–CCCCCC
AATGC–CCCCCC
–CTGCACACCG

CTGAGCATCG
CTGAGC–TCG
ATGAGC–TC–
CTGACACGG

AGCAGCATCGTG
AGCAGC–TCGTG
AGCAGC–TC–TG
CTA–CACGGTG

CAGGCACGCACGAA
AGC–CACGC–CATA
ATGGCACGC–C–TA
AGCTAC–CACGGAT
Step 1: infer gene trees (traditional methods)

ACTGCACACCG
ACTGC−CCCCCG
AATGC−CCCCCG
−CTGCACACCGG

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

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

CAGGCACGCACGAA
AGC−CACGC−CATA
ATGGCACGC−C−TA
AGCTAC−CACGGAT
Challenge 1:
Inferring the species tree from a set of gene trees is difficult for large datasets

Step 1: infer gene trees (traditional methods)
Number of species impacts estimation error in the species tree

1000 genes, “medium” levels of ILS, simulated species trees
[S. Mirarab, T. Warnow, 2015]
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)

\[ \theta_1 = 70\% \quad \theta_2 = 15\% \quad \theta_3 = 15\% \]

\[ d = 0.8 \]
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
More than 4 species

For 5 or more species, the unrooted species tree topology can be different from the most probable gene tree (called “anomaly zone”) (Degnan, 2013)
More than 4 species

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])
More than 4 species

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])

Alternative:
weight all \( \binom{3}{4} \) quartet topologies
by their frequency
and find the optimal tree

(probabilities are made-up just as an example)
Maximum Quartet Support Species Tree

$$Score(T) = \sum_{1}^{k} \left| Q(T) \cup Q(t_i) \right|$$

- Optimization problem:

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:
  
  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 [Mirarab, et al, Bioinformatics 2014]
Maximum Quartet Support Species Tree

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

- Optimization problem:
  
  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 [Mirarab, et al, Bioinformatics 2014]

- ASTRAL: an exact solution using dynamic programming
Number of species impacts estimation error in the species tree

1000 genes, “medium” levels of ILS, simulated species trees
[S. Mirarab, T. Warnow, 2015]
ASTRAL: accurate and scalable

1000 genes, “medium” levels of ILS, simulated species trees
[S. Mirarab, T. Warnow, 2015]
ASTRAL: accurate and scalable

1000 genes, “medium” levels of ILS, simulated species trees
[S. Mirarab, T. Warnow, 2015]
ASTRAL versions

• ASTRAL-I (<v. 4.7.3): 2014 - 2015
ASTRAL versions

• ASTRAL-I (<v. 4.7.3): 2014 - 2015

• ASTRAL-II (<v. 5.1.0): 2015 - 2017
  
  • Increased the accuracy by expanding the search space and improved the scalability
  
  • Can handle polytomies in input gene trees
ASTRAL versions

- ASTRAL-I (<v. 4.7.3): 2014 - 2015
- ASTRAL-II (<v. 5.1.0): 2015 - 2017
  - Increased the accuracy by expanding the search space and improved the scalability
  - Can handle polytomies in input gene trees
- ASTRAL-III (>v. 5.1.1): since 2017
  - Better running time, and better search space
  - Especially improved for unresolved trees, making it feasible to remove very low support branches
ASTRAL used widely

Early use:
- Plants: Wickett, et al., 2014, PNAS
- Birds: Prum, et al., 2015, Nature
- Xenoturbella, Cannon et al., 2016, Nature
- Xenoturbella, Rouse et al., 2016, Nature
- Flatworms: Laumer, et al., 2015, eLife
- Angiosperms: Huang et al., 2016, MBE
- Worms: Andrade, et al., 2015, MBE
Comparison to concatenation: depends on the level of discordance

Simulations, 200 species, deep ILS
[Mirarab and Warnow, 2016]
Comparison to concatenation: depends on the level of discordance

Simulations, 200 species, deep ILS
[Mirarab and Warnow, 2016]
Step 1: infer gene trees (traditional methods)

ACTGCACACCCG  ACTGC–CCCCCG  AATGC–CCCCCG  –CTGCACACCGG

ACTGACATCG  CTGAGC–TCG  ATGAGC–TC–  CTGA–CAG–G

Step 2: infer species trees

Gene tree

Gene tree

Gene tree

Gene tree

Gorilla  Human  Chimp  Orangutan
Step 2: infer species trees

Challenge 2:
Gene trees will have errors that will look like true discordance

Gene tree

Gene tree

Gene tree

Gene tree

Gorilla   Human   Chimp   Orangutan

ACTGCACACCG
ACTGCACCCCCG
AATGACCCCG
-CTGCACACGG
CTGAGCATCG
CTGAGCCTCG
ATGAGCCTCG
CTGACACCG
AGCAGCATCGTG
AGCAGCCTCGTG
AGCAGCCTCG
C-TA-CACGGTG
CAGGCACGCACGAA
AGCCACGCCTATA
ATGGCACGCCTA
AGCTACCATCGAT

Step 1: infer gene trees (traditional methods)

Challenge 2:
Gene trees will have errors that will look like true discordance
Comparison to concatenation: depends on the level of gene tree error

Simulations, 200 species, deep medium level ILS
[Mirarab and Warnow, 2016]
Comparison to concatenation: depends on the level of gene tree error

Simulations, 200 species, deep medium level ILS
[Mirarab and Warnow, 2016]
How to deal with gene tree error?

- **Statistical binning** (used in Jarvis *et al.* [Mirarab et al, Science, 2014] and other forms of binning have emerged since [Bayzid, et al, PLOS One, 2015])
How to deal with gene tree error?

• **Statistical binning** (used in Jarvis *et al.*) [Mirarab *et al.*, Science, 2014] and other forms of binning have emerged since [Bayzid, *et al.*, PLOS One, 2015]

• There are new **site-based methods**, which avoid gene trees altogether
  
  • **SVDQuartets** [Chou, BMC Genomics, 2015]
How to deal with gene tree error?

- **Statistical binning** (used in Jarvis *et al.*) [Mirarab *et al*, Science, 2014] and other forms of binning have emerged since [Bayzid, *et al*, PLOS One, 2015]

- There are new **site-based methods**, which avoid gene trees altogether
  - **SVDQuartets** [Chou, BMC Genomics, 2015]

- *BEAST*: **co-estimate** gene trees and species trees.
  - Its recent second version is more scalable
How to deal with gene tree error?

• **Statistical binning** (used in Jarvis *et al.*) [Mirarab et al, Science, 2014] and other forms of binning have emerged since [Bayzid, et al, PLOS One, 2015]

• There are new **site-based methods**, which avoid gene trees altogether
  
  • **SVDQuartets** [Chou, BMC Genomics, 2015]

• **BEAST**: co-estimate gene trees and species trees.
  
  • Its recent second version is more scalable

• **revPoMo**: concatenation with ILS-aware sequence evolution models [Schrempf et al, J. Theor. Bio., 2016]
How to deal with gene tree error?

- **Statistical binning** (used in Jarvis *et al.*) [Mirarab et al, Science, 2014] and other forms of binning have emerged since [Bayzid, et al, PLOS One, 2015]

- There are new **site-based methods**, which avoid gene trees altogether
  - **SVDQuartets** [Chou, BMC Genomics, 2015]

- **BEAST**: co-estimate gene trees and species trees.
  - Its recent second version is more scalable


- ASTRAL-III allows a different solution …
Contract low support branches

- It helps to contract very low support branches
- Mostly helps in the presence of low support gene trees
- Helps most for large numbers of gene trees

Simulations: 100 taxa, simpiphy, ILS: around 46% true discordance
FastTree, support from bootstrapping

BMC Bioinformatics, 2018, Zhang et al.
ASTRAL-III on all 14,446 unbinned gene trees

Beyond topology, ASTRAL estimates …

• **length** of internal branches in **coalescent units**: 
  \# generations / population size
Beyond topology, ASTRAL estimates ...

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

- a measure of branch **support** called **local posterior probability** [Sayyari and Mirarab, MBE, 2016]
Beyond topology, ASTRAL estimates …

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

- A measure of branch **support** called **local posterior probability** [Sayyari and Mirarab, MBE, 2016]

- **P-values** for a **polytomy test** [Sayyari and Mirarab, Genes, 2018]
Beyond topology, ASTRAL estimates …

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

- a measure of branch **support** called **local posterior probability** [Sayyari and Mirarab, MBE, 2016]

- P-values for a **polytomy** test 
  [Sayyari and Mirarab, Genes, 2018]

- quartet-based measures of gene tree discordance
Discovista: visualizing discordance

- https://github.com/esayyari/DiscoVista
Discovista: visualizing discordance

- https://github.com/esayyari/DiscoVista
How about hybridization?

• **PhyloNet** suit of tools make an effort to distinguish ILS and hybridization [e.g., Yu et al, 2014, PNAS]

  • Scalability remains to be tested

• **PhyloNetworks** takes a pseudo-likelihood approach [e.g., Solís-Lemus et al, 2016]
Other source of discordance?

- There has been progress on duplication and loss, but perhaps less relevant to avian phylogenomics
Challenges

• Errors and incompleteness in data due to annotation, assembly, or other unknown origins

• Models of sequence evolution

• Gene tree discordance
  • True discordance
  • Spurious discordance

• Scalability
Challenges

- Sequences that look like they may be incorrect
- Perhaps wrong annotation
Challenges

- Sequences that look like they may be incorrect
- Perhaps wrong annotation
TreeShrink
[Mai and Mirarab, BMC Genomics, 2018]

• Automatically detect long branches in gene trees

• It learns a distribution of branch length per species and looks for outliers
  • Avoids removing species that have long branches in all genes
  • Reduces discordance of gene trees
Fragmentary sequences

• Sequence of some species is present for some gene, but just a small portion of it
Filtering fragments

[Sayyari et al, MBE, 2017]

- Added fragmentation to simulated data with patterns similar to the Misof. et. al. insect (transcriptome data)
- Filtering simply removes fragmentary data from genes but keeps the gene
Should you remove whole genes?

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

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

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

• Errors and incompleteness in data due to annotation, assembly, or other unknown origins

• Models of sequence evolution

• Gene tree discordance
  • True discordance
  • Spurious discordance

• Scalability
Many tools have improved speed since 2014!

• ASTRAL-MP: super scalable ASTRAL using GPU and CPU multi-threading [under review]

• RAxML-ng+ParGenes: scalable gene tree estimation

• ASTRID, which is similar to NJst, and is quite good, but is super-fast
Challenges

• Errors and incompleteness in data due to annotation, assembly, or other unknown origins

• Models of sequence evolution

• Gene tree discordance
  • True discordance
  • Spurious discordance

• Scalability
Some new models …

- IQ-TREE:
  - PMFS: Wang et al., Systematic Biology, 2018
  - Heterotachy (GHOST): Crotty et al.
  - Partition models: Chernomor et al., Systematic Biology, 2016
- AA-biochemical model: Braun, ISMB, 2018
Summary

• There are better methods of species tree estimation and data correction available

• Sequence evolution models have not changed dramatically

• Scalability has improved and is not an issue for some but not all analyses

• Many challenges remain!
Acknowledgments

Jim Leebens-mack (UGA)
Norman Wickett (U Chicago)
Gane Wong (U of Alberta)
S.M. Bayzid
Nam Nguyen (now at UIUC)
Tandy Warnow
Bastien Boussau (Université Lyon)
Erich Jarvis (Duke, HMMI)
Tom Gilbert (U Copenhagen)
Guojie Zhang (BGI, China)
Erfan Sayyari
Chao Zhang
Maryam Hashemi
John Yin
Uyen Mai
Ed Braun (U Florida)
Siavash Mirarab
Maryam Hashemi
John Yin
Uyen Mai

HHMI
NSERC
ALFRED P. SLOAN FOUNDATION
NSF
华大基因