Mathematical and Computational Challenges in Reconstructing Evolution

Tandy Warnow

The University of Illinois
Phylogeny (evolutionary tree)

From the Tree of the Life Website,
University of Arizona
• “Nothing in biology makes sense except in the light of evolution”

• “...... nothing in evolution makes sense except in the light of phylogeny ...”
This Talk

- Models of evolution, identifiability, statistical consistency
- Genome-scale phylogeny:
  - Incomplete lineage sorting and species tree estimation under the Multi-Species Coalescent model (MSC)
  - ASTRAL: non-parametric accurate and statistically consistent species tree estimation under the MSC
  - NJMerge: scaling species tree methods to large datasets
  - Statistical consistency when number of sites per locus is bounded
- Open questions
- Future directions
DNA Sequence Evolution (Idealized)

-3 mil yrs
-2 mil yrs
-1 mil yrs
today
Gene Tree Estimation

AGGGGCAT
TAGCCCC
TAGACTT
TGCACAA
TGCGCCCT

U
V
W
X
Y

U
V
W
X
Y
Markov Models of Sequence Evolution

The different sites are assumed to evolve \textit{i.i.d.} down the model tree (with rates that are drawn from a gamma distribution).
Markov Models of Sequence Evolution

The different sites are assumed to evolve \textit{i.i.d.} down the model tree (with rates that are drawn from a gamma distribution).

Simplest site evolution model (Jukes-Cantor, 1969):

- The model tree $T$ is binary and has substitution probabilities $p(e)$ on each edge $e$, with $0 < p(e) < 3/4$.
- The state at the root is randomly drawn from \{A,C,T,G\} (nucleotides)
- If a site (position) changes on an edge, it changes with equal probability to each of the remaining states.
- The evolutionary process is Markovian.

More complex models (such as the General Markov model) are also considered, often with little change to the theory.
Quantifying Error

FN: false negative (missing edge)
FP: false positive (incorrect edge)

50% error rate
Statistical Consistency
Questions

• Is the model tree identifiable?
• Which estimation methods are statistically consistent under this model?
• How much data does the method need to estimate the model tree correctly (with high probability)?
Answers?

• We know a lot about which site evolution models are identifiable, and which methods are statistically consistent.
• We know a little bit about the sequence length requirements for standard methods.

Take home message: need to limit (or not allow) heterogeneity to get identifiability!
Genome-scale data?

error

Data
Phylogeny + genomics = genome-scale phylogeny estimation.
Gene tree discordance

Incomplete Lineage Sorting (ILS) is a dominant cause of gene tree heterogeneity
Gene trees inside the species tree (Coalescent Process)

Gorilla and Orangutan are not siblings in the species tree, but they are in the gene tree.
Gene trees inside the species tree (Coalescent Process)

Gorilla and Orangutan are not siblings in the species tree, but they are in the gene tree.

Deep coalescence = INCOMPLETE LINEAGE SORTING (ILS): gene tree can be different from the species tree.

Courtesy James Degnan
1KP: Thousand Transcriptome Project

- 103 plant transcriptomes, 400-800 single copy “genes”
- Next phase will be much bigger
- Wickett, Mirarab et al., *PNAS* 2014

**Major Challenge:**
- Massive gene tree heterogeneity consistent with ILS
Avian Phylogenomics Project

- Approx. 50 species, whole genomes
- 14,000 loci
- Multi-national team (100+ investigators)
- 8 papers published in special issue of Science 2014

Major challenge:
- Massive gene tree heterogeneity consistent with ILS.
Traditional approach: concatenation

- Statistically **inconsistent** and can even be positively misleading (proved for unpartitioned maximum likelihood) [Roch and Steel, Theo. Pop. Gen., 2014]

Statistically consistent methods

- **Coalescent-based summary methods**: Estimate gene trees, and then combine together (*ASTRAL*, *ASTRID*, MP-EST, NJst, and others)

- **Co-estimation methods**: Co-estimate gene trees and species trees (*TOO EXPENSIVE*)

- **Site-based methods**: estimate the species tree from the concatenated alignment, and do not estimate gene trees (*NOT WELL STUDIED*)
Main competing approaches

Species

<table>
<thead>
<tr>
<th>gene 1</th>
<th>gene 2</th>
<th>\ldots</th>
<th>gene k</th>
</tr>
</thead>
</table>

Concatenation

Analyze separately

Summary Method
What about summary methods?
What about summary methods?

Techniques:
- Most frequent gene tree?
- Consensus of gene trees?
- Other?
Species tree estimation from unrooted gene trees

Theorem (Allman et al.): Under the multi-species coalescent model, for any four taxa A, B, C, and D, the most probable unrooted gene tree on \{A,B,C,D\} is identical to the unrooted species tree induced on \{A,B,C,D\}.
Species tree estimation from unrooted gene trees

Theorem (Allman et al.): Under the multi-species coalescent model, for any four taxa A, B, C, and D, the most probable unrooted gene tree on \{A,B,C,D\} is identical to the unrooted species tree induced on \{A,B,C,D\}.

Proof: For every four species, select most frequently observed tree as the species tree. Then combine quartet trees!
Species tree estimation from unrooted gene trees

Theorem (Allman et al.): Under the multi-species coalescent model, for any four taxa A, B, C, and D, the most probable unrooted gene tree on \{A,B,C,D\} is identical to the unrooted species tree induced on \{A,B,C,D\}.

Proof: For every four species, select most frequently observed tree as the species tree. Then combine quartet trees!
ASTRAL

[Mirarab, et al., ECCB/Bioinformatics, 2014]

• Optimization Problem (NP-Hard):

Find the species tree with the maximum number of induced quartet trees shared with the collection of input gene trees

\[ \text{Score}(T) = \sum_{t \in T} |Q(T) \cap Q(t)| \]

• **Theorem:** Statistically consistent under the multi-species coalescent model when solved exactly
Constrained Maximum Quartet Support Tree

• Input: Set $\mathcal{T} = \{t_1, t_2, \ldots, t_k\}$ of unrooted gene trees, with each tree on set $S$ with $n$ species, and set $X$ of allowed bipartitions

• Output: Unrooted tree $T$ on leafset $S$, maximizing the total quartet tree similarity to $\mathcal{T}$, subject to $T$ drawing its bipartitions from $X$. 

Theorems (Mirarab et al., 2014):

• If $X$ contains the bipartitions from the input gene trees (and perhaps others), then an exact solution to this problem is statistically consistent under the MSC.

• The constrained MQST problem can be solved in $O(|X|^{2nk})$ time. (We use dynamic programming, and build the unrooted tree from the bottom-up, based on “allowed clades” – halves of the allowed bipartitions.)

Conjecture: MQST is NP-hard
Constrained Maximum Quartet Support Tree

- **Input:** Set $\mathcal{T} = \{t_1,t_2,...,t_k\}$ of unrooted gene trees, with each tree on set $S$ with $n$ species, and set $X$ of allowed bipartitions.
- **Output:** Unrooted tree $T$ on leafset $S$, maximizing the total quartet tree similarity to $\mathcal{T}$, subject to $T$ drawing its bipartitions from $X$.

Theorems (Mirarab et al., 2014):

- If $X$ contains the bipartitions from the input gene trees (and perhaps others), then an exact solution to this problem is statistically consistent under the MSC.
Constrained Maximum Quartet Support Tree

• Input: Set $\mathcal{T} = \{t_1, t_2, \ldots, t_k\}$ of unrooted gene trees, with each tree on set $S$ with $n$ species, and set $X$ of allowed bipartitions.
• Output: Unrooted tree $T$ on leafset $S$, maximizing the total quartet tree similarity to $\mathcal{T}$, subject to $T$ drawing its bipartitions from $X$.

Theorems (Mirarab et al., 2014):
• If $X$ contains the bipartitions from the input gene trees (and perhaps others), then an exact solution to this problem is statistically consistent under the MSC.
• The constrained MQST problem can be solved in $O(|X|^2nk)$ time. (We use dynamic programming, and build the unrooted tree from the bottom-up, based on “allowed clades” – halves of the allowed bipartitions.)
Simulation study

- Variable parameters:
  - Number of species: 10 – 1000
  - Number of genes: 50 – 1000
  - Amount of ILS: low, medium, high
  - Deep versus recent speciation
  - 11 model conditions (50 replicas each) with heterogenous gene tree error
  - Compare to NJst, MP-EST, concatenation (CA-ML)
  - Evaluate accuracy using FN rate: the percentage of branches in the true tree that are missing from the estimated tree

Used SimPhy, Mallo and Posada, 2015
Tree accuracy when varying the number of species

1000 genes, “medium” levels of recent ILS
Tree accuracy when varying the number of species

1000 genes, “medium” levels of recent ILS
Accuracy in the presence of HGT + ILS

Data: Fixed, moderate ILS rate, 50 replicates per HGT rates (1)-(6), 1 model species tree per replicate on 51 taxa, 1000 true gene trees, simulated 1000 bp gene sequences using INDELible, 1000 gene trees estimated from GTR simulated sequences using FastTree-2.

---

Price, Dehal, Arkin 2015
Fletcher, Yang 2009

Davidson et al., RECOMB-CG, BMC Genomics 2015
Impact of Gene Tree Estimation Error
(from Molloy and Warnow 2017)

Error is fraction of bipartitions that are not recovered

Summary Methods
- Site-based Method
  - Error is fraction of bipartitions that are not recovered

High ILS (41% AD)

Mean GTEE

Species Tree Error

20-50% (N=33)
50-80% (N=35)
80-85% (N=4)

ASTRAL  ASTRID  MP-EST  SVDquartets  CA-ML

Summary Methods  Site-based Method
Impact of Gene Tree Estimation Error
(from Molloy and Warnow 2017)

Summary Methods

Error is fraction of bipartitions that are not recovered

Note: Summary methods better than CA-ML for low GTEE, then worse!
Summary (so far)

- ASTRAL has high accuracy, is relatively robust to gene tree estimation error, missing data, and HGT.
- ASTRID is also good, but not generally quite as accurate as ASTRAL.
- Concatenation using ML: unpartitioned CA-ML not statistically consistent under the MSC but can be more accurate than even the best summary methods when ILS is low enough or gene tree estimation error is high enough.
- Site-based methods (e.g., SVDquartets): promising but not as good on simulated data as CA-ML, and not as good as summary methods except under very high gene tree estimation error.
Running time as function of # species

1000 genes, “medium” levels of ILS, simulated species trees
[Mirarab and Warnow, ISMB, 2015]
But ASTRAL can fail to return a tree within 24 hrs on some very large datasets with high ILS.
Scalability to large datasets

• All the methods described here (except distance-based methods such as ASTRID, NJst) have computational challenges on datasets with large numbers of species:
  – SVDquartets: best accuracy obtained if computing all quartet trees
  – ASTRAL: can fail on some datasets with many species and genes (constraint space too big)
  – CA-ML: no current method scales to large numbers of species and genes
NJMerge

• General technique for scaling phylogeny estimation methods to large numbers of species

• Approach:
  – Divide species set into disjoint subsets
  – Construct species tree (constraint tree) on each subset
  – Merge together using a modification of Neighbor Joining, obeying constraint trees, using pairwise distances between species from the “internode distance matrix” (used in ASTRID and NJst)
  – Note: can sometimes fail to return a tree (0.2% of cases in our experiments)

• Molloy and Warnow, RECOMB-CG 2018 (to appear)
NJMerge + ASTRAL vs. ASTRAL:
Comparable accuracy and can analyze larger datasets
NJMerge + SVDquartets vs. SVDquartets:
Better accuracy and can analyze larger datasets!

Species Tree Error

Level of ILS

Running Time (m)

Level of ILS

SVDquartets  NJMerge+SVDquartets
NJMerge + RAxML vs. RAxML: Better accuracy and faster!

Species Tree Error

100 taxa, 25 introns

100 taxa, 100 introns

100 taxa, 1000 introns

1000 taxa, 1000 introns

Level of ILS

Running Time (m)

100 taxa, 25 introns

100 taxa, 100 introns

100 taxa, 1000 introns

1000 taxa, 1000 introns

Level of ILS

RAxML  NJMerge+RAxML  NJMerge+RAxML (in SERIAL)  NJMerge+RAxML (in PARALLEL)
NJMerge summary

- **NJMerge+ASTRAL**: generally as accurate and faster on large datasets than ASTRAL
- **NJMerge+SVDquartets**: more accurate and much faster, greater scalability than SVDquartets
- **NJMerge+CA-ML**: more accurate and much faster, greater scalability than CA-ML

Only limitation: NJMerge can sometimes fail to return a tree, but this only occurred in 0.2% of the datasets we examined in our experiments. (Other methods also fail to return a tree due to time constraints.)
Which type of method is best?

• What is the meaning of “best”?
  – Statistically consistent under the MSC?
  – Good accuracy on simulated data?
  – Good accuracy on biological data?
Genome-scale data?

error

Data = sites? genes?
Statistical consistency for summary methods

Data = true gene trees!
Impact of Gene Tree Estimation Error
(from Molloy and Warnow 2017)

Summary Methods
- Site-based Method
  - Error is fraction of bipartitions that are not recovered

Note:
Summary methods better than CA-ML for low GTEE, then worse!

Species Tree Error

Mean GTEE
- 20-50% (N=33)
- 50-80% (N=35)
- 80-85% (N=4)

Summary Methods
- ASTRAL
- ASTRID
- MP-EST

Site-based Method
- SVDquartets
- CA-ML
• Multiple studies show that *summary methods can be less accurate than concatenation* in the presence of high gene tree estimation error.

• Genome-scale data includes a range of markers, not all of which have substantial signal. Furthermore, removing sites due to model violations reduces signal.

• Some researchers also argue that “gene trees” should be based on **very short alignments**, to avoid intra-locus recombination.
What about performance on bounded number of sites?

• Question #1: Do any summary methods converge to the species tree as the number of loci increase, but where each locus has only a constant number of sites?
• Answer #1: Roch & Warnow, Syst Biol, March 2015:
  – Strict molecular clock: Yes for some new methods, even for a single site per locus
  – No clock: Unknown for all methods, including MP-EST, ASTRAL, etc.

What about performance on bounded number of sites?

• Question #1: Do any summary methods converge to the species tree as the number of loci increase, but where each locus has only a constant number of sites?
• Answer #2: Roch, Nute, & Warnow, accepted to Syst Biol.
  • No! Summary methods are not only not consistent, they can be positively misleading! (Felsenstein Zone)

• Question #2: What about concatenation using maximum likelihood?
• Answer: Roch, Nute, & Warnow, accepted to Syst Biol.
  • Not if fully partitioned! Concatenation using maximum likelihood, if fully partitioned is also not consistent and can be positively misleading (even if there is NO ILS)! (Felsenstein Zone)

Statistically consistent methods

• **Coalescent-based summary methods**: Estimate gene trees, and then combine together *(ASTRAL, ASTRID, MP-EST, NJst, and others)*

• **Co-estimation methods**: Co-estimate gene trees and species trees *(TOO EXPENSIVE)*

• **Site-based methods**: estimate the species tree from the concatenated alignment, and do not estimate gene trees *(NOT WELL STUDIED)*
Statistically consistent methods (??)

• **Coalescent-based summary methods**: Estimate gene trees, and then combine together (ASTRAL, ASTRID, MP-EST, NJst, and others)

• **Co-estimation methods**: Co-estimate gene trees and species trees (TOO EXPENSIVE)

• **Site-based methods**: estimate the species tree from the concatenated alignment, and do not estimate gene trees (NOT WELL STUDIED)
Future Directions

• Theory: Determine which species tree estimation methods are statistically consistent when the number of sites per locus is bounded, and heterogeneity between loci is not constrained

• Practice: Understand why concatenation performs well, and develop better coalescent-based methods
• NP-hard optimization problems and large datasets
• Statistical estimation under stochastic models of evolution
• Probabilistic analysis of algorithms
• Graph-theoretic divide-and-conquer
• Chordal graph theory
• Combinatorial optimization
Acknowledgments

Roch and Warnow, Systematic Biology 2014
Mirarab and Warnow, Bioinformatics 2015
Molloy and Warnow, Systematic Biology 2017
Molloy and Warnow, RECOMB-CG 2018
Roch, Nute, and Warnow, arXiv:1803.02800 (accepted to Systematic Biology)
Papers available at http://tandy.cs.illinois.edu/papers.html

Funding: NSF, Grainger Foundation, and HHMI (to SM).