Mathematical and Computational Challenges in Reconstructing Evolution

Tandy Warnow
The University of Illinois
Phylogeny
(phylogenetic tree)

- Orangutan
- Gorilla
- Chimpanzee
- Human

From the Tree of the Life Website,
University of Arizona
Phylogeny + genomics = genome-scale phylogeny estimation.

Nature Reviews | Genetics
• “Nothing in biology makes sense except in the light of evolution”

• “...... nothing in evolution makes sense except in the light of phylogeny ...”
“Big Data”:
• Heterogeneous
• Large
• Noisy
• Error-ridden
• Streaming
• Model-misspecification

Approaches:
• 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
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
  – TreeMerge: scaling species tree methods to large datasets
• Discussion and Future directions
Phyogenomic Pipeline

• Assemble and annotate genomes (e.g., determine orthologs)
• Compute multiple sequence alignments of individual loci
• Construct gene trees
• Construct species tree
• Perform post-tree analyses (e.g., estimate dates, infer selection, etc.)
Phyogenomic Pipeline

• Assemble and annotate genomes (e.g., determine orthologs)
• Compute multiple sequence alignments of individual loci
  • Construct gene trees
  • Construct species tree
• Perform post-tree analyses (e.g., estimate dates, infer selection, etc.)
DNA Sequence Evolution (Idealized)

AAGGCCT
-3 mil yrs
AGGGCAT
-2 mil yrs
TAGCCCA
-1 mil yrs
TAGACTT
today
AGCACAA
AGCGCTT
AAGACTT
TGGACTT
Phylogeny Problem

AGGGCAT  TAGCCCA  TAGACTT  TGCACAA  TGCGCTT
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 Generalized Time Reversible model) are also considered, often with little change to the theory.
Phylogeny Problem

AGGGCAT
TAGCCCA
TAGACTT
TGCACAA
TGCGGCTT
FN: false negative  
(missing edge)  
FP: false positive  
(incorrect edge)

50% error rate
Statistical Consistency/Identifiability

error

Data
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)?
• What are the computational issues?
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 model identifiability!*
Computational issues

- **Maximum likelihood (ML):** NP-hard, and tree-space grows exponentially with the number of leaves
- **Bayesian estimation:** need to run to convergence (may fail)
- Parallelism helps but is not enough
Genome-scale data?
Phylogenomics

Phylogeny + genomics = genome-scale phylogeny estimation.

Nature Reviews | Genetics
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.

Courtesy James Degnan
Gene Trees inside the Species Tree (Coalescent Process)

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

Gorilla and Orangutan are not siblings in the species tree, but they are in the gene 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.
Hierarchical Model: MSC+GTR

• Multi-locus data, generated by a hierarchical model
  – Species tree generates gene trees under MSC
  – Gene trees generate sequences under the Generalized Time Reversible (GTR) model
Gene evolution model

Species tree

Gorilla  Human  Chimp  Orangutan

Gene tree

Chimp  Human  Orang.
Gorilla  Human  Chimp

Sequence data
(Alignments)

ACTGCACACCG  CTGAGCATCG
ACTGC–CCCCCG  CTGAGC–TCG
AATGC–CCCCCG  ATGAGC–TC–
–CTGCACACCGG  CTGA–CAC–G

Sequence evolution model

Sequence data
(Alignments)

1 1

AGCAGCATCGTG  CAGGCACGCACGAA
AGCAGC–TCGTG  AGC–CACGC–CATA
AGCA–TC–TG  ATGGCACGC–C–TA
C–TA–CACGGTG  AGCTAC–CACGGAT
Big picture challenge

• Multi-locus data, generated by a hierarchical model
  – Species tree generates gene trees
  – Gene trees generate sequences

• How can we estimate the species tree from the sequence data?
Big picture challenge

• Multi-locus data, generated by a hierarchical model
  – Species tree generates gene trees
  – Gene trees generate sequences

• How can we estimate the species tree from the sequence data?

• Suppose the number of genes and the sequence data per gene both go to infinity?
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)
Gene evolution model

Sequence evolution model

Species tree

Orangutan

Gorilla

Human

Chimp

Gene tree

Gene tree

Gene tree

Gene tree

Sequence data (Alignments)

ACTGCACACCG
ACTGC–CCCCCG
AATGC–CCCCCG
–CTGCACACCGG

CTGAGCATCG
CTGAGC–TCG
ATGAGC–T–
CTGACACCGG

AGCAGCATCGTG
AGCAGC–TCGTG
AGCAGC–TC–TG
AGCTAC–CACGGTG

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

Gene tree

Gene tree

Gene tree

Gene tree

Sequence data (Alignments)
Step 1: infer gene trees (traditional methods)

ACTGCACACCG
ACTGC–CCCCCG
AATGC–CCCCCG
–CTGCACACCGG

Gene tree

Gene tree

Gene tree

Gene tree

Gorilla Human Orang. Chimp

Gorilla Human Orang. Chimp

Gorilla Human Orang. Chimp

Gorilla Human Orang. Chimp

Step 2: infer species trees

Gene tree

Gene tree

Gene tree

Gene tree

Gorilla Human Orang. Chimp

Gorilla Human Orang. Chimp

Gorilla Human Orang. Chimp

Gorilla Human Orang. Chimp

ACTGACATCG
CTGAGCATCG
AGCAGCATCGTG
CAGGCACGCACGAA

ACTGC–TCGTG
AGCAG–TCGTG
AGCAGC–TC–TG
AGC–CACGC–CATA

AATGC–TC–
AGCAGC–TC–TG
AGC–CACGC–C–TA
ATGGCACGC–C–TA

CTGAC–C–
CTGAC–C–
C–TA–CAGGAT
AGCTAC–CACGGAT
What about summary methods?
What about summary methods?

Techniques:
  Most frequent gene tree?
  Consensus of gene trees?
  Other?
Quartet tree estimation

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\}.
Quartet tree estimation

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\}. 

\[ \theta_1=70\% \quad \theta_2=15\% \quad \theta_3=15\% \]
Species tree estimation from unrooted gene trees

Corollary: Under the multi-species coalescent model, the species tree is identifiable from the gene tree distribution.
Species tree estimation from unrooted gene trees

Corollary: Under the multi-species coalescent model, the species tree is identifiable from the gene tree distribution.

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

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

  a gene tree

  Set of quartet trees induced by T

  all input gene trees

• Theorem: Statistically consistent under the multispecies coalescent model when solved exactly
ASTRAL on biological datasets

- 1KP: **103** plant species, 400-800 genes
- Yang, et al. **96** Caryophyllales species, 1122 genes
- Dentinger, et al. **39** mushroom species, 208 genes
- Giarla and Esselstyn. **19** Philippine shrew species, 1112 genes
- Laumer, et al. **40** flatworm species, 516 genes
- Grover, et al. **8** cotton species, 52 genes
- Hosner, Braun, and Kimball. **28** quail species, 11 genes
- Simmons and Gatesy. **47** angiosperm species, 310 genes
- Prum et al, **198** avian species, 259 genes

Dissecting Molecular Evolution in the Highly Diverse Plant Clade Caryophyllales Using Transcriptome Sequencing

The Challenges of Resolving a Rapid, Recent Radiation: Empirical and Simulated Phylogenomics of Philippine Shrews

Nuclear genomic signals of the 'microturbellarian' roots of platyhelminth evolutionary innovation

Re-evaluating the phylogeny of allopolyploid *Gossypium* L.

A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing
ASTRAL

- Statistically consistent under the MSC, and runs in polynomial time
- Solves constrained version of the NP-hard Maximum Quartet Support problem using dynamic programming
  - Input: Gene trees and set X of allowed bipartitions
  - Output: Species tree T that maximizes the quartet support criterion, subject to drawing its bipartitions from the set X
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
Running time as function of # species

1000 genes, “medium” levels of ILS, simulated species trees
[Mirarab and Warnow, ISMB, 2015]
Running time as function of # species

But ASTRAL can fail to return a tree within 48 hrs on some very large datasets with high ILS.

1000 genes, “medium” levels of ILS, simulated species trees

[Mirarab and Warnow, ISMB, 2015]
Impact of Gene Tree Estimation Error (from Molloy and Warnow 2017)

Error is fraction of bipartitions that are not recovered

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

<table>
<thead>
<tr>
<th>Species</th>
<th>gene 1</th>
<th>gene 2</th>
<th>...</th>
<th>gene k</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Analyse separately

Concatenation

Summary Method
What about statistical consistency 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, Syst. Biol. 2019
  • 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, Syst Biol. 2019

- 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)

Summary (so far)

• Concatenation Analyses (e.g., CA-ML using RAxML):
  – most commonly used method, not statistically consistent, sometimes more accurate than summary methods
  – computationally intensive (e.g., 250 CPU years for the Avian Phylogenomics project with only 48 species) and do not scale to large numbers of species

• Summary methods (e.g., ASTRAL):
  – increasingly used in practice, consistent given true gene trees, sometimes more accurate than concatenation, but impacted by gene tree estimation error
  – very fast for many datasets (faster than concatenation) but ASTRAL can fail on large enough datasets
Summary (so far)

- Statistical consistency complicated issue
- The better summary methods (e.g., ASTRAL, ASTRID) are faster than concatenation analyses (e.g., CA-ML), and are more accurate when gene tree error is low
- CA-ML can be more accurate than all other methods when gene tree error is high
- But ASTRAL can fail to complete in 48 hrs on some datasets and CA-ML is extremely expensive (years of CPU time)
Decompose species set into pairwise disjoint subsets.

Build a tree on each subset

Full species set

Auxiliary Info (e.g., distance matrix)

Tree on full species set

Compute tree on entire set of species using “Disjoint Tree Merger” method

D&C Pipeline using Disjoint Tree Mergers (DTMs)
Disjoint Tree Merger techniques

All proven statistically consistent under MSC+GTR (when #sites per gene and #genes increase):

- **NJMerge** (Molloy and Warnow, RECOMB-CG 2018 and Alg Mol Biol, to appear)
- **TreeMerge** (Molloy and Warnow, ISMB 2019 and Bioinformatics 2019)
- **Constrained-INC** (Zhang, Rao, and Warnow, WABI 2018 and Alg Mol Biol 2019, and Le et al., AlCoB 2019)
- **GTM** (Smirnov and Warnow, RECOMB-CG 2019, and submitted to Alg Mol Biol)
TreeMerge

- Molloy and Warnow, ISMB 2019 and *Bioinformatics*, Volume 35, Issue 14, July 2019, Pages i417–i426
- TreeMerge is a direct improvement to NJMerge (which it uses on pairs of trees)
- Github site: [https://github.com/ekmolloy/treemerge](https://github.com/ekmolloy/treemerge)

Algorithmic strategy:
- divide species set into disjoint subsets,
- compute species trees on the subsets using selected species tree method, and
- merge subset trees using a distance-based method
DTMs Merge Subset Trees

Notes:
• Subset trees are requirements (constraint trees)
• Blending is permitted!
TreeMerge study

• Datasets: 1000 species and 1000 genes (exons and introns), two levels of ILS (low/mod and very high)
• Base methods: ASTRAL and with RAxML
• Criteria: Impact on base method
  – How many replicates complete?
  – Species tree error
  – Running time?
Impact of using TreeMerge with RAxML on 1000 species and 1000 genes
Impact of using TreeMerge with ASTRAL-III on 1000 species and 1000 genes

![Graph showing the impact of using TreeMerge with ASTRAL-III on 1000 species and 1000 genes. The graph compares species tree error and running time for exons and introns under low/moderate and very high ILS conditions. The results show that TreeMerge-fast has a lower species tree error and shorter running time compared to ASTRAL-III.](image-url)
Summary for TreeMerge

- **TreeMerge**
  - Improves scalability (ASTRAL and RAxML can fail to complete in 48 hours, but not when run with TreeMerge)
  - Reduces running time, sometimes dramatically
  - Maintains accuracy
  - Statistically consistent under MSC+GTR when used with ASTRAL (or other statistically consistent base method)
Summary

• The best tree estimation methods are computationally intensive, and tree-space grows exponentially
Summary

• The best tree estimation methods are computationally intensive, and tree-space grows exponentially

• Statistical consistency is important but not sufficient

• Divide-and-conquer improves scalability, maintains statistical consistency, and can maintain accuracy (or only lose a small amount)

• Divide-and-conquer is highly parallelizable
Summary

- The best tree estimation methods are computationally intensive, and tree-space grows exponentially.
- Statistical consistency is important but not sufficient.
- Parallel implementations of expensive methods are helpful but not enough.
- Divide-and-conquer improves scalability, maintains statistical consistency, and can maintain accuracy (or only lose a small amount).
- Divide-and-conquer is highly parallelizable.
Summary

• The best tree estimation methods are computationally intensive, and tree-space grows exponentially
• Statistical consistency is important but not sufficient
• Parallel implementations of expensive methods are helpful but not enough
• Divide-and-conquer improves scalability, maintains statistical consistency, and can maintain accuracy (or only lose a small amount)
Summary

• The best tree estimation methods are computationally intensive, and tree-space grows exponentially
• Statistical consistency is important but not sufficient
• Parallel implementations of expensive methods are helpful but not enough
• Divide-and-conquer improves scalability, maintains statistical consistency, and can maintain accuracy (or only lose a small amount)
• Divide-and-conquer is highly parallelizable
Acknowledgments

Papers available at http://tandy.cs.illinois.edu/papers.html
Presentations available at http://tandy.cs.illinois.edu/talks.html

**Funding:** NSF (CCF 1535977 and also NSF Graduate Fellowship to Erin Molloy)
**Supercomputers:** Blue Waters and Campus Cluster, both supported by NCSA
Future Work

- Complex genome-scale phylogeny estimation (heterotachy across tree, gene duplication and loss, horizontal gene transfer, genome rearrangements, etc.)
- Sequence analysis from 3rd generation sequencing technologies, applications to microbiomes, metagenomic sequence analysis, ecology, etc.
- High-performance computing
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\(^8\), 1000 gene trees estimated from GTR simulated sequences using FastTree-2\(^7\)

\(^7\) Price, Dehal, Arkin 2015
\(^8\) Fletcher, Yang 2009

Davidson et al., RECOMB-CG, BMC Genomics 2015