New methods for very-large scale maximum likelihood tree estimation

Tandy Warnow
The University of Illinois
Joint work with: Paul Zaharias and Minhyuk Park
Phylogeny + genomics = genome-scale phylogeny estimation.
Phylogenomic pipeline

- Select taxon set and markers
- Gather and screen sequence data, possibly identify orthologs
- Compute multiple sequence alignments for each locus, and construct gene trees
- Compute species tree or network:
  - Combine the estimated gene trees, OR
  - Estimate a tree from a concatenation of the multiple sequence alignments
- Get statistical support on each branch (e.g., bootstrapping)
- Estimate dates on the nodes of the phylogeny
- Use species tree with branch support and dates to understand biology
Phylogenomic pipeline

• Select taxon set and markers
• Gather and screen sequence data, possibly identify orthologs
• Compute multiple sequence alignments for each locus, and construct gene trees
• Compute species tree or network:
  • Combine the estimated gene trees, OR
  • Estimate a tree from a concatenation of the multiple sequence alignments
• Get statistical support on each branch (e.g., bootstrapping)
• Estimate dates on the nodes of the phylogeny
• Use species tree with branch support and dates to understand biology
1KP: Thousand Transcriptome Project

- 2014 *PNAS* study: 103 plant transcriptomes, 400-800 single copy “genes”
- 2019 *Nature* study: much larger!

**Major Challenges:**
- Large alignments (and sequence length heterogeneity)
- Multi-copy genes omitted (9500 -> 400)
- Massive gene tree heterogeneity consistent with ILS
Avian Phylogenomics Project

Erich Jarvis, HHMI
MTP Gilbert, Copenhagen
Guojie Zhang, BGI
Siavash Mirarab, Texas
Tandy Warnow, Texas and UIUC

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

Major challenges:
• Multi-copy genes omitted
• Massive gene tree heterogeneity consistent with ILS
• Concatenation analysis took 250 CPU years
Large datasets are difficult

• Two dimensions:
  • Number of loci
  • Number of species (or individuals)
• Missing data
• Heterogeneity
• Many analytical pipelines involve Maximum likelihood and Bayesian estimation
What I hope to convince you of:

• Great progress in large-scale phylogeny estimation (both for gene trees and species trees)
• "Disjoint tree mergers" (DTMs) are generic methods, that can be used with any phylogeny estimation method (for any kind of data), and enable scalability to large datasets.
  • The Guide Tree Merger (GTM) is the current leading DTM technique, based on empirical performance.
  • GTM improves maximum likelihood gene tree estimation and also species tree estimation.
  • However, GTM does NOT allow blending, and so should be able to be improved.
• Open problem: Develop a better DTM approach that allows blending.
This talk

- Part I: Models of sequence evolution and maximum likelihood
- Part II: Divide-and-conquer methods for maximum likelihood tree estimation
- Part III: Applications of techniques to species tree estimation, and open problems
Part I

• Models of evolution
• Maximum likelihood tree estimation
DNA Sequence Evolution (Idealized)
Phylogeny Problem

AGGGCAT  TAGCCCCA  TAGACTT  TGCACAA  TGCCTGCTT
FN: false negative
(missing edge)
FP: false positive
(incorrect edge)

50% error rate
Is method M statistically consistent under model G?

Question answered by mathematical proof
Markov Models of Sequence Evolution

The different sites are assumed to evolve *i.i.d.* down the model tree, so it suffices to model a single site.

Jukes-Cantor, 1969 (simplest DNA site evolution model):

- The state at the root is randomly drawn from \{A,C,T,G\} (nucleotides).
- The model tree \(T\) is binary and has substitution probabilities \(p(e)\) on each edge \(e\), with \(0 < p(e) < 3/4\).
- 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 (e.g., Generalized Time Reversible) are also considered, often with little change to the theory.
Questions

• Is the model tree identifiable?
• Which estimation methods are statistically consistent under this model?
• What is the sample complexity of the method (i.e., 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 sample complexity for standard methods.

*Take home message: need to limit (or not allow) heterogeneity to get model identifiability!*
Part III: Large-scale maximum likelihood trees
Maximum likelihood tree estimation

• **Input**: multiple sequence alignment and “model” (e.g., GTR, Jukes-Cantor)

• **Output**: Model tree (rooted binary tree with numeric parameters) that maximizes the probability of producing the alignment

Other optimization problems also used, such as maximum parsimony, and various distance-based optimization.
Bayesian methods also used.
Maximum likelihood tree estimation

• Theory:
  • Statistically consistent under standard models
  • Low sample complexity (Roch & Sly, Prob. Theory and Related Fields, 2017): phase transition (logarithmic then polynomial)
  • NP-hard

• Empirical (based on heuristics) – using RAxML (leading ML heuristic)
  • Outstanding accuracy on simulated data
  • Challenging on large datasets (best methods can take CPU years or fail to run on large datasets)
Maximum Likelihood Software (heuristics)

- RAxML-ng (probably the best?)
- IQ-TREE2 (possibly competitive with RAxML-ng)
- FastTree 2 (extremely fast, not as accurate)
- And others, but none competitive with RAxML-ng

These use hill-climbing and randomness to get out of local optima.
None (other than FastTree 2) are designed really for ML on large datasets (many sequences)
Decompose species set into pairwise disjoint subsets.

Full species set

Auxiliary Info (e.g., distance matrix)

Tree on full species set

Build a tree on each subset

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

Note: use most accurate method on subsets, and treat as absolute constraints

Erin Molloy, Introduced this approach
DTMs Merge Subset Trees

Notes:
• Subset trees are requirements (constraint trees)
• Blending is permitted!
Decompose species set into pairwise disjoint subsets.

Build a tree on each subset.

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

Note: use most accurate method on subsets, and treat as absolute constraints.

Erin Molloy, Introduced this approach

Theorem: If the subtree method is statistically consistent, then many DTM methods are statistically consistent.
Disjoint Tree Mergers (DTMs)

- NJMerge (Molloy and Warnow, Alg Mol Biol 2019)
- TreeMerge (Molloy and Warnow, Bioinf 2019)
- Constrained-IN (Zhang, Rao, and Warnow, Alg Mol Biol 2019)
- Guide Tree Merger (Smirnov and Warnow, 2020)
Guide Tree Merger

• Input:
  • set \( \mathcal{T} \) of trees \( T_i \) on leafset \( S_i \) (disjoint sets)
  • “guide tree” \( T \) on union of \( S_i \)
• Output: Tree \( T^* \) that induces each \( T_i \) and minimizes the bipartition distance to \( T \)

• NP-hard
• If we constrain \( T^* \) to be formed by adding edges between the trees \( T_i \) (i.e., no blending allowed), then solvable in polynomial time.
• Smirnov and Warnow, BMC Genomics 2020
Decompose species set into pairwise disjoint subsets.

Full species set

Auxiliary Info (e.g., distance matrix)

Tree on full species set

Build a tree on each subset

Note: use most accurate method on subsets, and treat as absolute constraints

RAxML, IQ-TREE, etc

Guide Tree Merger

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

Divide-and-Conquer Gene Tree Estimation
Figure 2 from "Disjoint Tree Mergers for Large-Scale Maximum Likelihood Tree Estimation", Park et al., Algorithms 2021

GTM pipeline:
- starting tree is IQ-Tree or FastTree (smaller datasets),
- IQ-tree used to compute subset trees,
- Guide Tree = Starting Tree
GTM-pipeline:
- Scales to large datasets
- Is competitive with RAxML and IQ-TREE for accuracy
- Is only slightly slower than starting tree (but more accurate)
Trends
- On RNASim10k: GTM most accurate topology
- On RNASim50K:
  - IQTree failed
  - RAxML had nearly 100% error
  - GTM most accurate
What about biological data?

• We used the same technique but evaluated maximum likelihood scores on an MAGUS+EMMA alignment of the Recombinase dataset (~70,000 protein sequences) from Kelly Williams, restricting the alignment to approximately 1000 sites.

• Revised GTM pipeline: construct FastTree tree on full-length sequences, and add remaining sequences in using phylogenetic placement method BATCH-SCAMPP (with EPA-ng) – Eleanor Wedell et al. (2023).

• We let RAxML run with different starting trees: its default approach, using FastTree as a starting tree, and using our GTM tree as a starting tree.

• We compared these RAxML runs (different starting trees) to each other, using LG+Gamma(4) for the model.

• Unpublished analyses performed by Minhyuk Park.
Analysis of Kelly Williams dataset (Minhyuk Park et al., NYP)

Choice of starting tree matters!

RAxML continues to improve its ML score during the entire 8 day period (but most gains are in the first 4 days)

GTM takes a bit more than 24 hours
On this dataset,

- Default RAxML worst
- FastTree is a better starting tree
- GTM is much better

Large datasets need long running times and very good starting trees!
Part III: Species Tree Estimation

From the Tree of the Life Website,
University of Arizona
Gene tree discordance

Multiple causes for discord, including

- Incomplete Lineage Sorting (ILS),
- Gene Duplication and Loss (GDL), and
- Horizontal Gene Transfer (HGT)
Gene tree discordance

Multiple causes for discord, including

- Incomplete Lineage Sorting (ILS),
- Gene Duplication and Loss (GDL), and
- Horizontal Gene Transfer (HGT)
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.
Is method M statistically consistent under model G?

Question answered by mathematical proof

Error
in species tree inferred by method M

Amount of data
generated under model G and then given to method M as input
Genome-scale data?

![Graph showing the relationship between error and length of the genome. The error decreases as the length of the genome increases.]
MSC+GTR Hierarchical Model

1. Gene trees evolve within the species tree (under the Multi-Species Coalescent model)
2. Sequences evolve down the gene trees (under GTR model)
Traditional approach: concatenation

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

Main Approaches for Species Tree Estimation under ILS

- **Gene Concatenation**: e.g., RAxML
- **Analyzing Separately**: e.g., ASTRAL
- **Summary Method**
Decompose species set into pairwise disjoint subsets.

Full species set

Auxiliary Info (e.g., distance matrix)

Tree on full species set

Build a tree on each subset

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

Note: use most accurate method on subsets, and treat as absolute constraints

Erin Molloy, Introduced this approach

Theorem: If the subtree method is statistically consistent, then many DTM methods are statistically consistent.
Disjoint Tree Mergers for Species Tree Estimation

Decompose species set into pairwise disjoint subsets.

Build a tree on each subset.

Note: use most accurate method on subsets, and treat as absolute constraints.

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

Erin Molloy, Introduced this approach

Use ASTRAL or Concatenation for subtree construction!

Combine with DTM method.
GTM+ASTRAL: faster and more accurate than ASTRAL

Table 3 Comparison of average runtime (seconds) of GTM+ASTRAL vs ASTRAL for high ILS conditions with introns on 1000 species. The value for $n$ is the number of replicates being compared (i.e., where ASTRAL trees are available). Pre-GTM covers computing gene trees using FastTree, the NJst starting tree, and ASTRAL subset trees; the gap between “total” and “ASTRAL” for the right hand column reflects the time to compute gene trees using FastTree, which is 3.9 seconds per gene. Results for the 1000-gene ASTRAL trees are taken from the NJMerge study [2].

<table>
<thead>
<tr>
<th></th>
<th>GTM+ASTRAL</th>
<th>ASTRAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 Genes ($n=18$)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-GTM</td>
<td>97.4</td>
<td>n.a.</td>
</tr>
<tr>
<td>ASTRAL</td>
<td>n.a.</td>
<td>8,617.0</td>
</tr>
<tr>
<td>GTM</td>
<td>0.4</td>
<td>n.a.</td>
</tr>
<tr>
<td>Total</td>
<td>97.8</td>
<td>8,656.0</td>
</tr>
<tr>
<td>25 Genes ($n=20$)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-GTM</td>
<td>174.7</td>
<td>n.a.</td>
</tr>
<tr>
<td>ASTRAL</td>
<td>n.a.</td>
<td>5,441.4</td>
</tr>
<tr>
<td>GTM</td>
<td>0.4</td>
<td>n.a.</td>
</tr>
<tr>
<td>Total</td>
<td>175.1</td>
<td>5,539.4</td>
</tr>
<tr>
<td>1000 Genes ($n=16$)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre-GTM</td>
<td>7,948.9</td>
<td>n.a.</td>
</tr>
<tr>
<td>ASTRAL</td>
<td>n.a.</td>
<td>149,145.9</td>
</tr>
<tr>
<td>GTM</td>
<td>0.4</td>
<td>n.a.</td>
</tr>
<tr>
<td>Total</td>
<td>7,949.3</td>
<td>153,045.9</td>
</tr>
</tbody>
</table>
Summary and open problem

• Great progress in large-scale phylogeny estimation (both for gene trees and species trees)

• “Disjoint tree mergers” (DTMs) are generic methods, that can be used with any phylogeny estimation method (for any kind of data).
  • DTMs enable scalability to large datasets.
  • DTMs improve Maximum Likelihood gene tree estimation as well as ASTRAL (species tree estimation).
  • GTM is the current leading DTM technique, based on empirical performance. However, because it does NOT allow blending, it is unlikely GTM is the best that can be done.

• Open problem: Develop a better DTM approach that allows blending.
Overall summary

• Large-scale phylogenetic tree estimation is becoming truly feasible!
  • Large numbers of sequences no longer a major impediment
  • Heterogeneity across the genome presents challenges, but methods are being
developed that address biological heterogeneity

• Not discussed here (and still needs work):
  • Phylogenetic networks
  • Genome rearrangement phylogeny
  • Multiple whole genome alignment
Acknowledgments

Papers available at [http://tandy.cs.illinois.edu/papers.html](http://tandy.cs.illinois.edu/papers.html)
Presentations available at [http://tandy.cs.illinois.edu/talks.html](http://tandy.cs.illinois.edu/talks.html)
Software on github, links at [http://tandy.cs.illinois.edu/software.html](http://tandy.cs.illinois.edu/software.html)

**Funding:** NSF (CCF 1535977, ABI-1458652, 2006069, Graduate Fellowship to Erin Molloy), the Grainger Foundation, the Ira and Debra Cohen Fellowship to Vlad Smirnov, and [Sandia National Laboratories-Livermore (LDRD)](http://tandy.cs.illinois.edu/)

**Supercomputers:** Blue Waters and Campus Cluster, both supported by NCSA

Write to me: warnow@illinois.edu