Challenges and Advances for Phylogenomic Estimation

Tandy Warnow
The University of Illinois
(joint work with Erin Molloy)
Phylogenomics

Phylogeny + genomics = genome-scale phylogeny estimation.
• “Nothing in biology makes sense except in the light of evolution”

• “...... nothing in evolution makes sense except in the light of phylogeny ...”
1KP: Thousand Transcriptome Project

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

**Major Challenges:**
- 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 challenge:
• Massive gene tree heterogeneity consistent with ILS
• ExaML analysis took 250 CPU years

and many others!
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

“Big Data”:

- Heterogeneous
- Large
- Noisy
- Error-ridden
- Streaming
- Model-misspecification
“Resolving the Tree of Life is unquestionably among the most complex scientific problems facing biology and presents challenges much greater than sequencing the human genome.”

This talk

• Methods to address gene tree heterogeneity
• A new divide-and-conquer approach for phylogeny estimation
• Other challenges
  – Addressing violations of GTR model assumptions
  – Gene duplication and loss
  – Reticulate phylogeny
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.
FN: false negative  
(missing edge)

FP: false positive  
(incorrect edge)

50% error rate
Phylogeny Problem

AGGGCAT  TAGCCCA  TAGACTT  TGCACAA  TGCGGCTT

U  V  W  X  Y
Statistical Consistency/Identifiability
Genome-scale data?
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)

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

Past

Present

Courtesy James Degnan
Three major approaches

• Concatenation using Maximum likelihood (CA-ML): NP-hard, and not consistent in the presence of ILS

• Summary methods (e.g., ASTRAL, ASTRID, MP-EST): consistent, and much faster than CA-ML, but impacted by gene tree estimation error

• Site-based methods (e.g., SVDquartets, SVDquest): consistent but not very efficient
Main competing approaches

- Analyze separately
- Concatenation
- Summary Method
Traditional approach: concatenation

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

Main competing approaches

- Gene 1, Gene 2, ..., Gene k
- Species
- Concatenation
- Analyze separately
- Summary Method
Gene evolution model

Sequence evolution model

Species tree

Gene tree

Sequence data (Alignments)

ACTGACACCCCG
ACTGC-CCCCCG
AATGC-CCCCCG
-CTGCACACCG

Gene tree

CTGAGCATCG
CTGAGC-TCG
ATAGGC-TC-T
CTGAC-CAC-G

Gene tree

Orang. Human
Gorilla Human
Chimp Human
Orang. Chimp

Gene tree

Chimp Human
Gorilla Human
Orang. Human
Orang. Gorilla

Gene tree

Chimp Human
Orang. Human
Chimp Orang.

Sequence data (Alignments)

1

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

CAGG-CACGCACGAA
AGC-CACGC-CATA
ATGGC-CACGC-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
- AATGCACGCACG
- CTGAGCATCGTG
- CTGAGC-TCGTG
- AGCAGC-TC-TG
- C-TCACACGGAT
- CAGGCACGCACGAA
- AGCAGC-TCGTG
- ATGGCACGC-C-TA
- AGCTAC-CACGGAT

Step 2: infer species 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

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

• Theorem: Statistically consistent under the multi-species coalescent model when solved exactly
Tree accuracy when varying the number of species

1000 genes, “medium” levels of recent ILS
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
ASTRAL – great, but…

• The good: ASTRAL is
  – increasingly used in practice
  – Statistically 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)

• The bad: ASTRAL can fail to complete on large enough datasets within reasonable time frames (days of computation)
The alternatives are worse

- Concatenation Analyses (e.g., 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

- Co-estimation of gene trees and species trees: too expensive

- Other statistically consistent methods: not as accurate as ASTRAL
D&C Pipeline using **Disjoint Tree Mergers (DTMs)**

Decompose species set into *pairwise disjoint* subsets.

1. Full species set
2. Auxiliary Info (e.g., distance matrix)
3. Tree on full species set
4. Build a tree on each subset
5. Compute tree on entire set of species using "Disjoint Tree Merger" method
DTMs Merge Subset Trees

Notes:
- Subset trees are requirements (*constraint trees*)
- Blending is permitted!
DTM methods

• NJMerge (Molloy and Warnow, 2018)
• INC (Zhang, Rao, and Warnow, 2019)
• TreeMerge (Molloy and Warnow, 2019)
• Guide Tree Merger (Smirnov and Warnow, 2019)
DTM methods

• NJMerge (Molloy and Warnow, 2018)
• INC (Zhang, Rao, and Warnow, 2019)
• TreeMerge (Molloy and Warnow, 2019)
• Guide Tree Merger (Smirnov and Warnow, 2019)
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

Theorems:
- TreeMerge is polynomial time
- Pipelines using TreeMerge are consistent under MSC+GTR (e.g., TreeMerge+ASTRAL, TreeMerge+SVDquest)
- TreeMerge+RAxML consistent under GTR
D&C Pipeline using Disjoint Tree Mergers (DTMs)

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.
Decompose species set into pairwise disjoint subsets.

Full species set

Use LOGDET matrix

Auxiliary Info (e.g., distance matrix)

Tree on full species set

Use TreeMerge

Build a tree on each subset

Use RAxML

Compute tree on entire set of species using “Disjoint Tree Merger” method
Impact of using TreeMerge-fast with RAxML on 1000 species and 1000 genes
Decompose species set into pairwise disjoint subsets.

Full species set

Use NJst matrix

Auxiliary Info (e.g., distance matrix)

Tree on full species set

Use ASTRAL

Build a tree on each subset

Use TreeMerge

Compute tree on entire set of species using “Disjoint Tree Merger” method
Impact of using TreeMerge with ASTRAL-III on 1000 species and 1000 genes
**TreeMerge Summary**

**Theorem:** TreeMerge enables polynomial time and statistically consistent species tree estimation pipelines.

**Empirical:** TreeMerge has been tested with SVDquartets, ASTRAL, and RAxML – and improves scalability, reduces running time, and maintains (or improves) accuracy.

**Running time:** TreeMerge uses only minutes to merge subset trees on 1000 species.

Molloy and Warnow, ISMB 2019.
Decompose species set into pairwise disjoint subsets.

Compute tree on entire set of species using "Disjoint Tree Merger" method.
Gene tree discordance

Multiple causes for discord, including
- Incomplete Lineage Sorting (ILS),
- Gene Duplication and Loss (GDL), and
- Horizontal Gene Transfer (HGT)
Gene Family Trees

Note that the gene family tree has two copies of the species tree!

The species tree has one duplication (at the root), which produces a gene family tree that has two copies of the species tree!

Multi-copy trees: MUL-trees
Problem: Given set of MUL-trees, infer the species tree

(a) Species tree $T^*$

(b) Gene tree $M_1$ with one duplication.

(c) Gene tree $M_2$ with one duplication and two losses.

(d) Gene tree with one duplication and three losses.
Problem: Given set of MUL-trees, infer the species tree

Many methods, but until Fall 2019, none proven statistically consistent under GDL

Now – ASTRAL-multi and FastMulRFS proven consistent!
FastMulRFS vs MulRF, ASTRAL-multi, and DupTree

Results on data (100 species):

- ASTRAL-multi shown
- FastMulRFS and MulRF tied for best in terms of accuracy
- FastMulRFS is by far the fastest

Results on 100-species datasets with moderate GDL, moderately high ILS, and high GTEE
FastMulRFS vs MulRF, ASTRAL-multi, and DupTree

Results on 100-species datasets with moderate GDL, moderately high ILS, and high GTEE
Summary

Under MSC+GTR:
• Substantial advances in estimating large species trees, addressing gene tree heterogeneity due to ILS (e.g., ASTRAL)

Under GDL+GTR:
• New methods (very recent, most unpublished) for species tree estimation addressing gene duplication and loss (FastMulRFS, ASTRAL-multi, and ASTRAL-pro)

In general:
• New divide-and-conquer approaches scaling methods to large datasets (e.g., TreeMerge)
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.
But biological evolution doesn’t follow these simplistic models

- Heterotachy
- Non-stationarity (e.g., GC content varying across the tree)
- Selection
Decompose species set into pairwise disjoint subsets.

Tree on full species set

Auxiliary Info (e.g., distance matrix)

Build a tree on each subset

Use TreeMerge

Use ANYTHING

Use ANYTHING

Compute tree on entire set of species using “Disjoint Tree Merger” method
More to do!

• Continue to advance species tree estimation in the presence of heterogeneity due to GDL (e.g., ASTRAL-Pro) and when more than one process is operating

• Heterotachy and other types of model misspecification (e.g., GC content varying across the tree) – so that standard maximum likelihood approaches are inadequate

• Genome-scale changes (e.g., rearrangements, duplications, etc.)

• Phylogenetic network estimation needed due to reticulate evolution (hybridization, horizontal gene transfer, etc.)
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 Graduate Fellowship to Erin), HHMI (to Siavash), and Debra and Ira Cohen (to Erin)

**Supercomputers:** Blue Waters and Campus Cluster, both supported by NCSA
Accuracy in the presence of HGT + ILS

200 Estimated Gene Trees

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