Advancing Genome-Scale Phylogenomic Analysis

Tandy Warnow
Department of Computer Science
Carl R. Woese Institute for Genomic Biology
National Center for Supercomputing Applications
The University of Illinois at Urbana-Champaign
http://tandy.cs.illinois.edu
Species Tree

From the Tree of the Life Website, University of Arizona
Phylogenomic Pipelines

- Finding related genomic regions (homology detection)
- Multiple sequence alignment
- Maximum Likelihood phylogeny estimation for single genes
- Species tree estimation from multiple conflicting genes
- Answer biological questions
Computational Grand Challenges

- NP-hard problems
- Exact solutions infeasible – heuristics necessary
- Parallelism helpful but does not address scalability with number of species
- Large datasets:
  - 100,000+ sequences
  - 10,000+ genes
- “BigData” complexity
DNA Sequence Evolution

AAGACTT

TG

GACTT

AAG

G

C

C

T

-3 mil yrs

-2 mil yrs

-1 mil yrs

today

AAGGCCT

AGGGCAT

AGGGCAT

TAGCCCT

TAGACTT

AGCACAA

AGCGCTT
Phylogeny Estimation

AGGGGCAT    TAGCCCCA    TAGACTTT    TGCACAA    TGCACCTT

U   V   W   X   Y

U   V   W   X   Y
Maximum Likelihood Phylogeny Estimation: A Grand Challenge

- Methods: RAxML, ExaML, PhyML, IQTree, FastTree-2 (and others)
- Only FastTree-2 seems feasible for very large numbers of species
- Large numbers of sites also impacts running time and accuracy
- Models matter
Phylogeny + genomics = genome-scale phylogeny estimation
Gene tree discordance

Incomplete Lineage Sorting (ILS) is a dominant cause of gene tree heterogeneity
Avian Phylogenomics Project

- Approx. 50 species, whole genomes, 14,000 loci
- Jarvis, Mirarab, et al., Science 2014

• Challenge #1: Maximum likelihood analysis of concatenated multiple sequence alignments took 250 CPU years using supercomputers around the globe.
• Challenge #2: Massive gene tree heterogeneity
1kp: Thousand Transcriptome Project

- PNAS 2014 (about 100 species and 800 genes)
- Second analysis underway, much larger dataset (~1200 species and ~1000 loci)

- Challenge #1: Construct multiple sequence alignment and tree for more than 100,000 sequences
- Challenge #2: Massive gene tree heterogeneity
Three of my favorite Grand Challenges

• **Multiple Sequence Alignment**: Methods for large-scale MSA (up to 1,000,000 sequences, including fragments)

• **Phylogenomics**: Methods for multi-locus species tree estimation that are robust to gene tree incongruence due to incomplete lineage sorting (ILS) and horizontal gene transfer (HGT)

• **Supertree estimation**: Methods that combine smaller trees into larger trees (useful for divide-and-conquer)
MSA estimation

• Small numbers of sequences (e.g., 50-100): MAFFT

• Larger numbers of sequences:
  – PASTA (improvement over SATé)
  – UPP (handles sequence length heterogeneity)

• New methods under development
PASTA and UPP: boosters of MSA methods

• PASTA
  – Combines iteration and divide-and-conquer to “boost” a preferred MSA method to large datasets; we showed results based on MAFFT

• UPP
  – Step 1: Constructs a “backbone” tree and an alignment on a small random subset of the sequences
  – Step 2: Aligns all the remaining sequences to the backbone alignment
  – We showed results where default PASTA computed the backbone alignment and tree.
Re-aligning on a tree (boosting an MSA method)

1. Decompose dataset
2. Align subsets
3. Estimate ML tree on merged alignment
4. Merge sub-alignments
SATé and PASTA Algorithms

Obtain initial alignment and estimated ML tree

Estimate ML tree on new alignment

Use tree to compute new alignment

Repeat until termination condition, and return the alignment/tree pair with the best ML score
PASTA: even better than SATé-2

- Simulated RNASim datasets from 10K to 200K taxa
- Limited to 24 hours using 12 CPUs
- Not all methods could run (missing bars could not finish)
UPP Algorithmic Approach

1. Select small random subset of full-length sequences, and build “backbone alignment”

2. Construct an “Ensemble of Hidden Markov Models” on the backbone alignment

3. Add all remaining sequences to the backbone alignment using the Ensemble of HMMs
Notes:
• We show alignment error using average of SP-FN and SP-FP.
• UPP variants have better alignment scores than PASTA.
• (Not shown: Total Column Scores – PASTA more accurate than UPP)
• No other methods tested could complete on these data
RNASim Million Sequences: tree error

Using 12 TACC processors:

- UPP(Fast,NoDecomp) took 2.2 days,
- UPP(Fast) took 11.9 days, and
- PASTA took 10.3 days
Species tree estimation

- Concatenation (e.g., ExaML)
- ILS-aware methods:
  - ASTRAL (Mirarab et al.)
  - ASTRID (Vachaspati and Warnow)
  - *BEAST and BBCA (improves scalability of *BEAST Zimmermann and Warnow)
  - SVDquartets (site-based method for quartet trees, may need may loci to provide good accuracy)
    - Needs quartet amalgamation method
    - PAUP* standard implementation
    - SVDquest is another implementation
Tree accuracy when varying the number of species

1000 genes, “medium” levels of recent ILS
Both ASTRAL and ASTRID substantially outperform MP-EST

Avian simulated dataset

Mammalian simulated dataset
ASTRID is very fast

On the ASTRAL-2 dataset with 1000 taxa, 1000 genes, ASTRID-FastME takes 33 minutes, ASTRAL takes 12 hours.
Supertree estimation

- MRP (matrix representation with parsimony) – standard approach
- MRL (matrix representation with likelihood)
- Robinson-Foulds Supertrees (e.g., FastRFS by Vachaspati and Warnow)
- Maximum Likelihood and Bayesian Supertrees

Applications to Divide-and-Conquer method, such as DACTAL
Other issues

• Very short loci (low information content): may not work well for “summary methods” such as ASTRAL and ASTRID
• No need to filter genes on the basis of missing data
• Filtering on the basis of gene tree estimation error may be beneficial

See Molloy and Warnow, Systematic Biology 2017
Results:

- $M_{iid}$ missing data:

26-taxon simulation with ILS, Nute et al., submitted
Results:

- $M_{clade}$ missing data:

26-taxon simulation with ILS, Nute et al., submitted
Species tree estimation error for MP-EST and ASTRAL, and also concatenation using ML, on avian simulated datasets: 48 taxa, moderately high ILS (AD=47%), 1000 genes, and varying gene sequence length.

Acknowledgments

Papers available at http://tandy.cs.illinois.edu/papers.html
PASTA, UPP, and ASTRAL at https://github.com/smirarab
ASTRID and FastRFS available at https://github.com/pranjalv123

Funding: NSF ABI-1458652 and III:AF:1513629, a Founder Professorship from the Grainger Foundation, and HHMI (to S.M.)
Computational support: BlueWaters and TACC