Advances in Phylogenomic Estimation

Tandy Warnow
The University of Illinois
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.)
**Brief Comments on MSA Estimation**

- Many good methods for aligning moderate sized datasets, but choice of alignment method depends on type of data (e.g., nucleotide vs. protein). MAFFT is very good for both types of data, and methods like PROMALS, T-Coffee, and ContrAlign are very good for proteins.

- Statistical co-estimation of alignments and trees (e.g., using BAli-Phy) also appealing, but expensive.

- Large-scale multiple sequence alignment (MSA) is achievable using PASTA (Mirarab et al., J. Computational Biology 2014) – an improvement over SATe (Liu et al., Science 2009)
  
  - PASTA available at [https://github.com/smirarab/](https://github.com/smirarab/)
  
  - PASTA+BAli-Phy at [http://github.com/MGNute/pasta](http://github.com/MGNute/pasta)
PASTA: even better than SATé-2

PASTA: Mirarab, Nguyen, and Warnow, J Comp. Biol. 2015

- 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)
Brief Comments on Gene Tree Estimation

- Simulations suggest maximum likelihood is a preferred approach to gene tree estimation
- Some ML software to consider:
  - **RAxML** (and its derivatives): generally the most popular
  - **IQTree**: comparable to RAxML under some conditions (sometimes worse, sometimes better) and includes **PoMo** (and its derivatives) which addresses polymorphism
  - **PhyML**: lots of models!
  - **FastTree-2**: may be the only software that can run on datasets with many many sequences (but not as accurate as RAxML, doesn’t have good parallelism, and cannot handle very “long” alignments)
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.)
I’ll use the term “gene” to refer to “c-genes”: recombination-free orthologous stretches of the genome.
Gene tree discordance

Incomplete Lineage Sorting (ILS) is a dominant cause of gene tree heterogeneity
Multi-locus Species Tree Estimation

• The main challenges are:
  – Computational issues (both time and memory usage)
  – Heterogeneity across the genome due to various causes, including
    • Incomplete Lineage Sorting (ILS)
    • Gene duplication and loss (if failure to correctly identify orthologs)
Four Basic Approaches

Statistically consistent methods:

- Co-estimate species tree and gene trees: e.g., *BEAST (Heled and Drummond)
- Site-based methods: e.g., SVDquartets (Chifman and Kubatko, implemented in PAUP*)
- Methods that combine gene trees (summary methods): e.g., NJst, MP-EST, ASTRAL, ASTRID, STEM, etc.

And of course

- Concatenation, but this isn’t statistically consistent in the presence of ILS
Main competing approaches

Species

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

Concatenation

Analyze separately

Summary Method
New methods (all open source)

- **BBCA**: Improving scalability of *BEAST* (Zimmermann, Mirarab, and Warnow 2015) to large numbers of loci
- **SVDquest**: Improving accuracy and scalability for SVDquartets (Vachaspati and Warnow, 2018)
- **ASTRAL** (Mirarab et al. 2014, 2015, etc.) and **ASTRID** (Vachaspati and Warnow 2015): summary methods that can analyze datasets with thousands of species and loci with high accuracy
- **NJMerge**: Improving scalability of species tree estimation methods to large numbers of species (Molloy and Warnow, 2018)
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
Algorithmic approach:

• Given set of gene trees, find the species tree that agrees with the maximum number of quartet trees within a constrained search space.
• Polynomial time and statistically consistent in the presence of ILS.
ASTRID

• ASTRID: Accurate species trees using internode distances, Vachaspati and Warnow, RECOMB-CG 2015 and BMC Genomics 2015
• Github site: https://github.com/pranjalv123/ASTRID

Algorithmic design:
• Nearly the same as NJst (Liu and Yu, 2010)- computes a matrix of average leaf-to-leaf topological distances, and then computes a tree using FastME (more accurate than neighbor Joining and faster, too).
• Polynomial time and statistically consistent in the presence of 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.
NJMerge

• Molloy and Warnow, RECOMB-CG 2018
• Github site: https://github.com/ekmolloy/njmerge

Algorithmic strategy:
• Divide-and-conquer: divides species set into disjoint subsets, computes species trees on the subsets using selected species tree method (e.g., ASTRAL, RAxML, SVDquartets), and then merges subset trees using a distance-based method.
NOTE: ASTRAL completed on 16/20 datasets with 1000 taxa and very high ILS; NJMerge+ASTRAL completed on 20/20 of these datasets; box plots show all datasets on which at least one method completed.
NOTE: ASTRAL completed on 16/20 datasets with 1000 taxa and very high ILS; NJMerge+ASTRAL completed on 20/20 of these datasets; bar graphs show all datasets on which at least one method completed.
RAxML and NJMerge: Tree Error

NOTE: RAxML completed on 1/40 datasets with 1000 taxa due to “Out of Memory” errors; NJMerge+RAxML completed on 40/40 of these datasets; box plots show all datasets on which at least one method completed.
NOTE: RAxML completed on 1/40 datasets with 1000 taxa due to “Out of Memory” errors; NJMerge+RAxML completed on 40/40 of these datasets; bar graphs show all datasets on which at least one method completed.
Summary for species tree estimation

- NJMerge improves scalability (and in some cases accuracy) of species tree estimation.
- ASTRAL and ASTRID are very good summary methods and can analyze ultra-large datasets.
- Concatenation is often highly accurate, and can be better than alternative methods under some conditions (notably when gene tree estimation error is high, ILS is high, and number of loci is low).
- Site-based methods (such as SVDquartets and SVDquest) are expected to provide improved accuracy (compared to summary methods and concatenation) under some conditions.
- Choice of method depends on the data (degree and cause of gene tree heterogeneity, gene tree estimation error, and number of species).
- Generally not beneficial to remove loci (but removing fragmentary sequences is another issue).
- This is an active research area, and new methods are being developed.
Acknowledgments

Software: all open source, see http://tandy.cs.illinois.edu/software.html

NSF grants DBI-1461364 and ABI-1458652
NSF graduate fellowships to Pranjal Vachaspati and Erin Molloy
HHMI graduate fellowship to Siavash Mirarab
Computing done at Blue Waters (part of NCSA)

Papers available at http://tandy.cs.illinois.edu/papers.html
Should you filter?

- Filtering genes based on missing data?
  - Generally not beneficial (see Molloy and Warnow, Systematic Biology 2018)

- Filtering genes based on gene tree estimation error?
  - Depends on conditions (see Molloy and Warnow, Systematic Biology 2018)

- Filtering fragmentary sequences from genes while keeping the gene?
  - Often beneficial (see Sayyari, Whitfield, and Mirarab, MBE 2018)
Comparing default PASTA to PASTA+BAli-Phy on simulated datasets with 1000 sequences

Decomposition to 100-sequence subsets, one iteration of PASTA+BAli-Phy
SVDquartets and NJMerge: Tree Error

NOTE: SVDquartets ran on 0/40 datasets with 1000 taxa due to segmentation faults; NJMerge+SVDquartets completed on 40/40 of these datasets; box plots show all datasets on which at least one method completed.
NOTE: SVDquartets ran on 0/40 datasets with 1000 taxa due to segmentation faults; NJMerge+SVDquartets completed on 40/40 of these datasets; bar graphs show all datasets on which at least one method completed.
Tree accuracy when varying the number of species

1000 genes, “medium” levels of recent ILS
What method should I use on large datasets?

<table>
<thead>
<tr>
<th>Method</th>
<th>Species Tree Error</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Moderate ILS</strong></td>
<td></td>
</tr>
<tr>
<td>NJMerge+ASTRAL</td>
<td></td>
</tr>
<tr>
<td>NJMerge+RAxML</td>
<td></td>
</tr>
<tr>
<td>NJMerge+SVDquartets</td>
<td></td>
</tr>
<tr>
<td>NJst</td>
<td></td>
</tr>
<tr>
<td><strong>Very High ILS</strong></td>
<td></td>
</tr>
<tr>
<td>NJMerge+ASTRAL</td>
<td></td>
</tr>
<tr>
<td>NJMerge+RAxML</td>
<td></td>
</tr>
<tr>
<td>NJMerge+SVDquartets</td>
<td></td>
</tr>
<tr>
<td>NJst</td>
<td></td>
</tr>
</tbody>
</table>
Multiple Sequence Alignment (MSA): a scientific grand challenge

S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC
S3 = TAGCTGACCGC
...
Sn = TCACGACCGACA

S1 = -AGGCTATCACCCTGACCTCCA
S2 = TAG-CTATCAC--GACCGC--
S3 = TAG--CT---------GACCGC--
...
Sn = -------TCAC--GACCGACA

Novel techniques needed for scalability and accuracy
NP-hard problems and large datasets
Current methods do not provide good accuracy
Few methods can analyze even moderately large datasets

Many important applications besides phylogenetic estimation

¹ Frontiers in Massive Data Analysis, National Academies Press, 2013
1000-taxon models, ordered by difficulty (Liu et al., Science 324(5934):1561-1564, 2009)
Re-aligning on a tree

Decompose dataset

Align subsets

Estimate ML tree on merged alignment

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
1000 taxon models, ordered by difficulty, Liu et al., Science 324(5934):1561-1564, 2009

24 hour SATé-I analysis, on desktop machines

(Similar improvements for biological datasets)
SATé-2 better than SATé-1

1000-taxon models ranked by difficulty

SATé-1 (Liu et al., Science 2009): can analyze up to 8K sequences
SATé-2 (Liu et al., Systematic Biology 2012): can analyze up to ~50K sequences
Re-aligning on a tree

Decompose dataset

Align subsets: MAFFT

Merge sub-alignments

Estimate ML tree on merged alignment
Re-aligning on a tree

1. **Decompose dataset**
   - A
   - B
   - C
   - D

2. **Align subsets**
   - BA
   - li-
   - Phy??
   - A
   - B
   - C
   - D

3. **Merge sub-alignments**
   - ABCD

4. **Estimate ML tree on merged alignment**
   - ABCD