2020 Phylogenomics Symposium and Software School

Tandy Warnow
The University of Illinois
NSF grant ABI-1458652
http://tandy.cs.illinois.edu/ssb-standalone-2020.html
Talks/Software Downloads

http://tandy.cs.illinois.edu/ssb-standalone-2020.html
Today

• Tutorials on: ASTRAL, PhyloNet, and PASTA
• Talks about species tree estimation:
  – ASTRAL-pro: species trees addressing GDL
  – ASTRID: species trees addressing ILS
  – TreeMerge: scaling tree estimation methods to large datasets
  – INSTRAL: adding genomes into trees
  – TreeShrink: finding errors in data
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
Orthology

- Determining which homologs are orthologs and which are paralogs is not that easy.
Orthology

- Determining which homologs are orthologs and which are paralogs is not that easy.
- A big question is: if you aren’t certain about this, what should you do with your data?
Orthology

• Determining which homologs are orthologs and which are paralogs is not that easy.
• A big question is: if you aren’t certain about this, what should you do with your data?

— See talk by Siavash Mirarab about ASTRAL-PRO
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
Indels (insertions and deletions)

...ACGGTG\textcolor{magenta}{CAGT}TACCA...

\textcolor{red}{Deletion} \quad \textcolor{red}{Mutation}

...ACC\textcolor{magenta}{CAGT}C\textcolor{red}{ACCA}...

...ACC\textcolor{magenta}{CAGT}C\textcolor{red}{ACCA}...
The true multiple alignment

- Reflects historical substitution, insertion, and deletion events
- Defined using transitive closure of pairwise alignments computed on edges of the true tree
Input: unaligned sequences

S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC
S3 = TAGCTGACCGC
S4 = TCACGACCGACA
Phase 1: Alignment

S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC
S3 = TAGCTGACCGC
S4 = TCACGACCGACA

S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAC--GACCGC--
S3 = TAG-CT-------GACCGC--
S4 = --------TCAC--GACCGACA
Phase 2: Construct tree

S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC
S3 = TAGCTGACCGC
S4 = TCACGACCGACA

S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC
S3 = TAGCTGACCGC
S4 = TCACGACCGACA
1000-taxon models, ordered by difficulty (Liu et al., 2009)
1KP dataset: more than 100,000 p450 amino-acid sequences, many fragmentary
1KP dataset: more than 100,000 p450 amino-acid sequences, many fragmentary.

All standard multiple sequence alignment methods we tested performed poorly on datasets with fragments.
Multiple Sequence Alignment (MSA): a scientific grand challenge

\[
\begin{align*}
S1 &= AGGCTATCACCTGACCTCCA & S1 &= -AGGCTATCACCTGACCTCCA \\
S2 &= TAGCTATCACGACCGC & S2 &= TAG-CTATCAC--GACCGC-- \\
S3 &= TAGCTGACCGC & S3 &= TAG-CT--------GACCGC-- \\
\cdots & & \cdots \\
Sn &= TCACGACCGACACA & Sn &= ---------TCAC--GACCGACACA
\end{align*}
\]

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
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 - see Tandy’s talk about MSA estimation, and then Kodi’s PASTA tutorial
• 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
# Two-phase estimation

## Alignment methods
- Clustal
- POY (and POY*)
- Probcons (and Proftree)
- Probalign
- MAFFT
- Muscle
- Di-align
- T-Coffee
- Prank (PNAS 2005, Science 2008)
- Opal (ISMB and Bioinf. 2007)
- Infernal (Bioinf. 2009)
- Etc.

## Phylogeny methods
- Bayesian MCMC
- Maximum parsimony
- **Maximum likelihood**
- Neighbor joining
- FastME
- UPGMA
- Quartet puzzling
- Etc.

*RAxML*: heuristic for large-scale ML optimization
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 interesting models (e.g., Ghost and PoMo)
  – **PhyML**: also 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, cannot handle very “long” alignments, and degrades on alignments with fragmentary sequences)
Challenges for ML tree estimation

• Computational difficulties:
  – Large numbers of sequences (tips)
  – Large numbers of sites
• Rogue taxa
• Fragmentary sequences
• Models too simple (e.g., assume a single substitution rate matrix across the tree)
Challenges for ML tree estimation

• Computational difficulties:
  – Large numbers of sequences (tips)
  – Large numbers of sites
• Rogue taxa
• Fragmentary sequences
• Models too simple (e.g., assume a single substitution rate matrix across the tree)

Not covered in today’s tutorials/talks, but we have ideas about how to address these issues.
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
phylogenomics

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.
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 (Roch and Steel)
Main competing approaches

Gene 1  Gene 2  ...  Gene k

Species

Concatenation

Analyze separately

Summary Method
ASTRAL

- Mirarab and Warnow, Bioinformatics 2014
- https://github.com/smirarab/ASTRAL

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.
Limitations of these methods

- Only address ILS - See ASTRAL-pro
Limitations of these methods

- Only address ILS - See ASTRAL-pro
- Only construct trees – See PhyloNet tutorial
Limitations of these methods

• Only address ILS - See ASTRAL-pro
• Only construct trees – See PhyloNet tutorial
• Errors in input can impact phylogeny estimation – See TreeShrink
Limitations of these methods

• Only address ILS - See ASTRAL-pro
• Only construct trees – See PhyloNet tutorial
• Errors in input can impact phylogeny estimation – See TreeShrink
• Not sufficiently scalable to large datasets
  – See INSTRAL and TreeMerge
Downloads

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:30-10:00</td>
<td>ASTRAL-Pro</td>
<td>Siavash Mirarab</td>
</tr>
<tr>
<td>10:00-10:30</td>
<td>Coffee break</td>
<td></td>
</tr>
<tr>
<td>10:30-10:45</td>
<td>ASTRID</td>
<td>Pranjal Vachaspati</td>
</tr>
<tr>
<td>10:45-11:00</td>
<td>TreeMerge</td>
<td>Erin Molloy</td>
</tr>
<tr>
<td>11:00-12:00</td>
<td>ASTRAL tutorial</td>
<td>Siavash Mirarab</td>
</tr>
<tr>
<td>12:00-1:30</td>
<td>Lunch break</td>
<td></td>
</tr>
<tr>
<td>1:30-2:30</td>
<td>Phylonet tutorial</td>
<td>Huy Ogilvie, Zhen Cao, and Zhi Yan</td>
</tr>
<tr>
<td>2:30-3:00</td>
<td>Coffee Break</td>
<td></td>
</tr>
<tr>
<td>3:00-3:15</td>
<td>INSTRAL</td>
<td>Maryam Rabiee</td>
</tr>
<tr>
<td>3:15-3:30</td>
<td>TreeShrink</td>
<td>Uyen Mai</td>
</tr>
<tr>
<td>3:30-3:45</td>
<td>Intro to MSA</td>
<td>Tandy</td>
</tr>
<tr>
<td>3:45-4:45</td>
<td>PASTA tutorial</td>
<td>Kodi Taraszka</td>
</tr>
<tr>
<td>4:45-5:00</td>
<td>Discussion and closing</td>
<td>(All speakers)</td>
</tr>
</tbody>
</table>