Phylogenomics
(Phylogenetic estimation from whole genomes)
Species Tree

From the Tree of the Life Website,
University of Arizona
The Tree of Life: **Multiple Challenges**

- Large-scale statistical phylogeny estimation
- Ultra-large multiple-sequence alignment
- Estimating species trees from incongruent gene trees
- Supertree estimation
- Genome rearrangement phylogeny
- Reticulate evolution
- Visualization of large trees and alignments
- Data mining techniques to explore multiple optima

Large datasets:
- 100,000+ sequences
- 10,000+ genes

“BigData” complexity
The Tree of Life: Multiple Challenges

Large datasets: 100,000+ sequences
10,000+ genes
“BigData” complexity

Large-scale statistical phylogeny estimation
Ultra-large multiple-sequence alignment
Estimating species trees from incongruent gene trees
Supertree estimation
Genome rearrangement phylogeny
Reticulate evolution
Visualization of large trees and alignments
Data mining techniques to explore multiple optima
Topics

• Gene tree estimation and statistical consistency
• Gene tree conflict due to incomplete lineage sorting
• The multi-species coalescent model
  – Identifiability and statistical consistency
  – The challenge of gene tree estimation error
  – The challenge of dataset size
• New methods for coalescent-based estimation
  – Statistical binning (Mirarab et al., 2014, Bayzid et al. 2014) – used in Avian tree
  – ASTRAL (Mirarab et al., 2014, Mirarab and Warnow 2015) – used in Plant tree
DNA Sequence Evolution (Idealized)
Markov Model of Site Evolution

Simplest (Jukes-Cantor, 1969):

• The model tree $T$ is binary and has substitution probabilities $p(e)$ on each edge $e$.
• The state at the root is randomly drawn from $\{A,C,T,G\}$ (nucleotides).
• If a site (position) changes on an edge, \textit{it changes with equal probability to each of the remaining states}.
• The evolutionary process is Markovian.

The different sites are assumed to evolve \textit{i.i.d.} (independently and identically) down the tree (with rates that are drawn from a gamma distribution).

More complex models (such as the General Markov model) are also considered, often with little change to the theory.
Maximum Likelihood Phylogeny Estimation

Input: Sequence set $S$
Output: Jukes-Cantor model tree $T$ (with substitution probabilities on edges) such that $\Pr(S|T)$ is maximized

ML tree estimation is usually performed under other more realistic models (e.g., the Generalized Time Reversible model)
Quantifying Error

** FN: false negative  
(missing edge) 

** FP: false positive  
(incorrect edge) 

50% error rate

** DNA SEQUENCES **

$S_1$  ACAATTAGAAC  
$S_2$  ACCCTTAGAAC  
$S_3$  ACCATTCCAAC  
$S_4$  ACCAGACCAAC  
$S_5$  ACCAGACCGGA  

** TRUE TREE **

** INFERRED TREE **

Maximum Likelihood is Statistically Consistent
Maximum Likelihood is Statistically Consistent

Data are sites in an alignment
Species Tree

Orangutan  Gorilla  Chimpanzee  Human

From the Tree of Life Website,
University of Arizona
Sampling multiple genes from multiple species

From the Tree of the Life Website,
University of Arizona
Using multiple genes

<table>
<thead>
<tr>
<th>gene 1</th>
<th>gene 2</th>
<th>gene 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_1$</td>
<td>$S_4$</td>
<td>$S_1$</td>
</tr>
<tr>
<td>TCTAATGGAA</td>
<td>GGTAACCCTC</td>
<td>TATTGATACA</td>
</tr>
<tr>
<td>$S_2$</td>
<td>$S_5$</td>
<td>$S_3$</td>
</tr>
<tr>
<td>GCTAAGGGAA</td>
<td>GCTAAACCTC</td>
<td>TCTTGATACC</td>
</tr>
<tr>
<td>$S_3$</td>
<td>$S_6$</td>
<td>$S_4$</td>
</tr>
<tr>
<td>TCTAAGGGAA</td>
<td>GGTGACCATC</td>
<td>TAGT GatGCA</td>
</tr>
<tr>
<td>$S_4$</td>
<td>$S_7$</td>
<td>$S_7$</td>
</tr>
<tr>
<td>TCTAACGGAA</td>
<td>GCTAAACCTC</td>
<td>TAGT GatGCA</td>
</tr>
<tr>
<td>$S_7$</td>
<td></td>
<td>$S_8$</td>
</tr>
<tr>
<td>TCTAATGGAC</td>
<td></td>
<td>CAT T CAT ACC</td>
</tr>
<tr>
<td>$S_8$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Concatenation

<table>
<thead>
<tr>
<th></th>
<th>gene 1</th>
<th>gene 2</th>
<th>gene 3</th>
</tr>
</thead>
</table>
| $S_1$ | TCTAATGGAA | ???????????? | TATTGATA
| $S_2$ | GCTAAGGGAA | ???????????? | ???????????? |
| $S_3$ | TCTAAGGGAA | ???????????? | TCTTGATA
| $S_4$ | TCTAACGGAA | GGTAACCTC | TAGTGAT
| $S_5$ | ???????????? | GCTAAACCTC | ???????????? |
| $S_6$ | ???????????? | GGTGACCATC | ???????????? |
| $S_7$ | TCTAATGGAC | GCTAAACCTC | TAGTGAT
| $S_8$ | TATAACGGAA | ???????????? | CATTCA

Diagram showing the concatenation of sequences from different genes.
Red gene tree ≠ species tree (green gene tree okay)
Avian Phylogenomics Project

E Jarvis, HHMI

MTP Gilbert, Copenhagen

G Zhang, BGI

T. Warnow, UT-Austin

S. Mirarab, UT-Austin

Md. S. Bayzid, UT-Austin

Plus many many other people...

- Approx. 50 species, whole genomes
- 8000+ genes, UCEs
- Gene sequence alignments computed using SATé (Liu et al., Science 2009 and Systematic Biology 2012)

Science 2014, Jarvis, Mirarab, et al.
1KP: Thousand Transcriptome Project

- 1200 plant transcriptomes
- More than 13,000 gene families (most not single copy)
- Multi-institutional project (10+ universities)
- iPLANT (NSF-funded cooperative)
- Gene sequence alignments and trees computed using SATe (Liu et al., Science 2009 and Systematic Biology 2012)

Proceedings of the National Academy of Sciences, Wickett, Mirarab et al., 2014
Gene Tree Incongruence

• Gene trees can differ from the species tree due to:
  – Duplication and loss
  – Horizontal gene transfer
  – Incomplete lineage sorting (ILS)
Incomplete Lineage Sorting (ILS)

• 1000+ papers in 2013 alone
• Confounds phylogenetic analysis for many groups:
  – Hominids
  – Birds
  – Yeast
  – Animals
  – Toads
  – Fish
  – Fungi
• There is substantial debate about how to analyze phylogenomic datasets in the presence of ILS.
Species tree estimation: difficult, even for small datasets!

From the Tree of Life Website,
University of Arizona
The Coalescent

Past

Present

Courtesy James Degnan
Gene tree in a species tree

Courtesy James Degnan
Lineage Sorting

• Population-level process, also called the “Multi-species coalescent” (Kingman, 1982)

• Gene trees can differ from species trees due to short times between speciation events or large population size; this is called “Incomplete Lineage Sorting” or “Deep Coalescence”.
Using multiple genes

<table>
<thead>
<tr>
<th>gene 1</th>
<th>gene 2</th>
<th>gene 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_1$</td>
<td>TCTAATGGAA</td>
<td>$S_1$</td>
</tr>
<tr>
<td>$S_2$</td>
<td>GCTAAGGGAA</td>
<td>$S_3$</td>
</tr>
<tr>
<td>$S_3$</td>
<td>TCTAAGGGAA</td>
<td>$S_4$</td>
</tr>
<tr>
<td>$S_4$</td>
<td>TCTAACGGAA</td>
<td>$S_7$</td>
</tr>
<tr>
<td>$S_7$</td>
<td>TCTAATGGAC</td>
<td>$S_8$</td>
</tr>
<tr>
<td>$S_8$</td>
<td>TATAACGGAA</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S_4$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S_5$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S_6$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S_7$</td>
</tr>
</tbody>
</table>

|        | $S_3$  |         |
|        | $S_4$  | TAGTGTGCA |
|        | $S_7$  | TAGTGTGCA |
|        | $S_8$  | CATTCATAAC |
How to compute a species tree?
Inconsistent methods

- MDC (Parsimony-style method, Minimize Deep Coalescence)
- Greedy Consensus
- MRP (supertree method)
- Concatenation under maximum likelihood

In other words, all the usual approaches are not consistent – and some can be positively misleading!
Species tree estimation

1- **Concatenation**: statistically inconsistent (Roch & Steel 2014)

2- **Summary methods**: can be statistically consistent

3- **Co-estimation methods**: too slow for large datasets
Two competing approaches

Analyze separately

Species

gene 1  gene 2  ...  gene \( k \)

Concatenation

Summary Method
Key observation:
Under the multi-species coalescent model, the species tree defines a *probability distribution on the gene trees, and is identifiable from the distribution on gene trees*
How to compute a species tree?
How to compute a species tree?

Techniques:
Most frequent gene tree?
Under the multi-species coalescent model, the species tree defines a probability distribution on the gene trees.

Theorem (Degnan et al., 2006, 2009): Under the multi-species coalescent model, for any three taxa A, B, and C, the most probable rooted gene tree on \{A,B,C\} is identical to the rooted species tree induced on \{A,B,C\}. 

Courtesy James Degnan
Theorem: The most probable rooted gene tree on three species is topologically identical to the species tree.

Proof: Let the species tree have topology \(((A,B),C)\), and probability \(p^*\) of coalescence on the edge \(e^*\) above LCA\((A,B)\), with \(0<p^*<1\). We wish to show that

\[
\Pr(gt=((A,B),C)) > \Pr(gt=((A,C),B)) = \Pr(gt=((B,C),A)).
\]
Theorem: The most probable rooted gene tree on three species is topologically identical to the species tree.

Proof: Let the species tree have topology \(((A,B),C)\), and probability \(p^*\) of coalescence on the edge \(e^*\) above LCA\((A,B)\), with \(0<p^*<1\). We wish to show that

\[
Pr(gt=((A,B),C)) > Pr(gt=((A,C),B)) = Pr(gt=((B,C),A)).
\]

(1) If the lineages from \(A\) and \(B\) coalesce on \(e^*\), then the gene tree is topologically identical to the species tree.
Theorem: The most probable rooted gene tree on three species is topologically identical to the species tree.

Proof: Let the species tree have topology ((A,B),C), and probability $p^*$ of coalescence on the edge $e^*$ above $\text{LCA}(A,B)$, with $0<p^*<1$. We wish to show that

$$\Pr(gt=((A,B),C)) > \Pr(gt=((A,C),B)) = \Pr(gt=((B,C),A)).$$

(1) If the lineages from A and B coalesce on $e^*$, then the gene tree is topologically identical to the species tree.

(2) If they do not coalesce, then all three lineages enter the edge above the root – and any pair can coalesce with equal probability.
Theorem: The most probable rooted gene tree on three species is topologically identical to the species tree.

Proof: Let the species tree have topology \(((A,B),C)\), and probability \(p^*\) of coalescence on the edge \(e^*\) above LCA\((A,B)\), with \(0<p^*<1\). We wish to show that

\[
\Pr(\text{gt}=((A,B),C)) > \Pr(\text{gt}=((A,C),B)) = \Pr(\text{gt}=((B,C),A)).
\]

(1) If the lineages from A and B coalesce on \(e^*\), then the gene tree is topologically identical to the species tree.

(2) If they do not coalesce, then all three lineages enter the edge above the root – and any pair can coalesce with equal probability.

Hence,

\[
\Pr(\text{gt}=((B,C),A)) = \Pr(\text{gt}=((A,C),B)) = (1-p^*)/3 < 1/3.
\]
Theorem: The most probable rooted gene tree on three species is topologically identical to the species tree.

Proof: Let the species tree have topology ((A,B),C), and probability p* of coalescence on the edge e* above LCA(A,B), with 0<p*<1. We wish to show that

\[ \Pr(gt=((A,B),C)) > \Pr(gt=((A,C),B)) = \Pr(gt=((B,C),A)). \]

(1) If the lineages from A and B coalesce on e*, then the gene tree is topologically identical to the species tree.

(2) If they do not coalesce, then all three lineages enter the edge above the root – and any pair can coalesce with equal probability.

Hence, \[ \Pr(gt=((B,C),A)) = \Pr(gt=((A,C),B)) = (1-p*)/3 < 1/3. \]
And also, \[ \Pr(gt = ((A,B),C) = p* + (1-p*)/3 > 1/3 \]
Theorem: The most probable unrooted gene tree on four species is topologically identical to the species tree.

Courtesy James Degnan
Theorem: The most probable unrooted gene tree on four species is topologically identical to the species tree.

Proof: The rooted species tree on \{A,B,C,D\} has one of two shapes: either balanced or “pectinate” (caterpillar). We show the proof when the species tree is balanced – it is a homework problem to do the other case.
Theorem: The most probable unrooted gene tree on four species is topologically identical to the species tree.

Proof: The rooted species tree on \{A,B,C,D\} has one of two shapes: either balanced or “pectinate” (caterpillar). We show the proof when the species tree is balanced – it is a homework problem to do the other case.

Let the rooted species tree have topology \(((A,B),(C,D))\), and probabilities \(p_1, p_2\) of coalescence on the edges \(e_1\) and \(e_2\) above \(\text{LCA}(A,B)\) and \(\text{LCA}(C,D)\), respectively, with \(0<p_1,p_2<1\). We wish to show that as *unrooted* gene trees,

\[
\Pr(\text{gt}=((A,B),(C,D))) > \Pr(\text{gt}=((A,C),(B,D))) = \Pr(\text{gt}=((B,C),(A,D))).
\]
Theorem: The most probable unrooted gene tree on four species is topologically identical to the species tree.

Proof: The rooted species tree on \{A,B,C,D\} has one of two shapes: either balanced or “pectinate” (caterpillar). We show the proof when the species tree is balanced – it is a homework problem to do the other case.

Let the rooted species tree have topology \(((A,B),(C,D))\), and probabilities \(p_1, p_2\) of coalescence on the edges \(e_1\) and \(e_2\) above \(LCA(A,B)\) and \(LCA(C,D)\), respectively, with \(0<p_1,p_2<1\). We wish to show that as *unrooted* gene trees,

\[
Pr(gt=((A,B),(C,D))) > Pr(gt=((A,C),(B,D))) = Pr(gt=((B,C),(A,D))).
\]

(1) If the lineages from A and B coalesce on \(e_1\), then the gene tree is topologically identical to the species tree as an unrooted tree.
Theorem: The most probable **unrooted** gene tree on **four species** is topologically identical to the species tree.

Proof: The rooted species tree on \{A,B,C,D\} has one of two shapes: either balanced or “pectinate” (caterpillar). We show the proof when the species tree is balanced – it is a homework problem to do the other case.

Let the rooted species tree have topology \(((A,B),(C,D))\), and probabilities \(p_1, p_2\) of coalescence on the edges \(e_1\) and \(e_2\) above LCA\((A,B)\) and LCA\((C,D)\), respectively, with \(0<p_1,p_2<1\). We wish to show that as *unrooted* gene trees,

\[
\Pr(gt=((A,B),(C,D))) > \Pr(gt=((A,C),(B,D))) = \Pr(gt=((B,C),(A,D))).
\]

(1) If the lineages from A and B coalesce on \(e_1\), then the gene tree is topologically identical to the species tree as an unrooted tree.

(2) If the lineages from C and D coalesce on \(e_2\), then the gene tree is topologically identical to the species tree as an unrooted tree.
Theorem: The most probable unrooted gene tree on four species is topologically identical to the species tree.

Proof: The rooted species tree on \{A,B,C,D\} has one of two shapes: either balanced or “pectinate” (caterpillar). We show the proof when the species tree is balanced – it is a homework problem to do the other case.

Let the rooted species tree have topology (\((A,B),(C,D)\)), and probabilities \(p_1, p_2\) of coalescence on the edges \(e_1\) and \(e_2\) above LCA(A,B) and LCA(C,D), respectively, with \(0 < p_1, p_2 < 1\). We wish to show that as *unrooted* gene trees,

\[
\Pr(\text{gt}=(A,B),(C,D)) > \Pr(\text{gt}=(A,C),(B,D)) = \Pr(\text{gt}=(B,C),(A,D)).
\]

(1) If the lineages from A and B coalesce on \(e_1\), then the gene tree is topologically identical to the species tree as an unrooted tree.

(2) If the lineages from C and D coalesce on \(e_2\), then the gene tree is topologically identical to the species tree as an unrooted tree.

(3) If neither pair coalesces on these edges, then all four lineages enter the edge above the root – and any pair can coalesce with equal probability.
Theorem: The most probable unrooted gene tree on four species is topologically identical to the species tree.

Proof: The rooted species tree on \{A,B,C,D\} has one of two shapes: either balanced or “pectinate” (caterpillar). We show the proof when the species tree is balanced – it is a homework problem to do the other case.

Let the rooted species tree have topology ((A,B),(C,D)), and probabilities p1, p2 of coalescence on the edges e1 and e2 above LCA(A,B) and LCA(C,D), respectively, with 0<p1,p2<1. We wish to show that as *unrooted* gene trees,

\[
\Pr(gt=((A,B),(C,D))) > \Pr(gt=((A,C),(B,D))) = \Pr(gt=((B,C),(A,D))).
\]

(1) If the lineages from A and B coalesce on e1, then the gene tree is topologically identical to the species tree as an unrooted tree.

(2) If the lineages from C and D coalesce on e2, then the gene tree is topologically identical to the species tree as an unrooted tree.

(3) If neither pair coalesces on these edges, then all four lineages enter the edge above the root – and any pair can coalesce with equal probability.

Hence, \[
\Pr(gt=((B,C),(A,D))) = \Pr(gt=((A,C),(B,D))) = (1-p1)(1-p2)/3 < 1/3.
\]
And so \[
\Pr(gt = ((A,B),(C,D))) > 1/3 \]
Theorem: The most probable rooted gene tree on four species may not be the rooted species tree.
Theorem: The most probable **rooted** gene tree on **four** species may not be the rooted species tree.

Proof: Let the species tree $T^*$ have topology $((A, (B, (C, D)))$, and probability $p_1, p_2$ of coalescence on the edges $e_1$ and $e_2$ above $LCA(C, D)$ and $LCA(B, C)$, respectively, with $0 < p_1, p_2 < 1$. We wish to show that we can pick $p_1, p_2$ so that as *rooted* gene trees, $Pr(gt=((A, B), (C, D))) > Pr(gt=T^*=(A, (B, (C, D))))$. 
Theorem: The most probable rooted gene tree on four species may not be the rooted species tree.

Proof: Let the species tree $T^*$ have topology $((A,(B,(C,D))),$ and probability $p_1$, $p_2$ of coalescence on the edges $e_1$ and $e_2$ above $LCA(C,D)$ and $LCA(B,C)$, respectively, with $0<p_1,p_2<1$. We wish to show that we can pick $p_1$, $p_2$ so that as *rooted* gene trees, $Pr(gt= ((A,B),(C,D))) > Pr(gt = T^* = (A,(B,(C,D))))$.

(1) Let $\varepsilon>0$ be given. Pick $p_1$ and $p_2$ small enough so that with probability at least $1-\varepsilon$, all four lineages enter the edge above the root (i.e., $(1-p_1)(1-p_2) > 1-\varepsilon$).
Theorem: The most probable **rooted** gene tree on **four** species may not be the rooted species tree.

Proof: Let the species tree $T^*$ have topology $((A,(B,(C,D))),$ and probability $p_1, p_2$ of coalescence on the edges $e_1$ and $e_2$ above $\text{LCA}(C,D)$ and $\text{LCA}(B,C),$ respectively, with $0<p_1,p_2<1$. We wish to show that we can pick $p_1, p_2$ so that as *rooted* gene trees, $\text{Pr}(\text{gt}=((A,B),(C,D))) > \text{Pr}(\text{gt}=T^*=(A,(B,(C,D))))$.

(1) Let $\varepsilon>0$ be given. Pick $p_1$ and $p_2$ small enough so that with probability at least $1-\varepsilon,$ all four lineages enter the edge above the root (i.e., $(1-p_1)(1-p_2) > 1-\varepsilon$).

(2) Then, all pairs of lineages have equal probability of coalescing. Each rooted tree is defined by the order of coalescent events –
Theorem: The most probable rooted gene tree on four species may not be the rooted species tree.

Proof: Let the species tree $T^*$ have topology $((A,(B,(C,D))),$ and probability $p_1,$ $p_2$ of coalescence on the edges $e_1$ and $e_2$ above $\text{LCA}(C,D)$ and $\text{LCA}(B,C),$ respectively, with $0<p_1,p_2<1$. We wish to show that we can pick $p_1,$ $p_2$ so that as *rooted* gene trees, $\Pr(gt=((A,B),(C,D))) > \Pr(gt=T^*=(A,(B,(C,D))))$.

1. Let $\varepsilon>0$ be given. Pick $p_1$ and $p_2$ small enough so that with probability at least $1-\varepsilon,$ all four lineages enter the edge above the root (i.e., $(1-p_1)(1-p_2) > 1-\varepsilon$).
2. Then, all pairs of lineages have equal probability of coalescing. Each rooted tree is defined by the order of coalescent events –
   1. $(A,(B,(C,D)))$ is produced by the sequence of coalescences $A&B,$ then $AB&C,$ then $ABC&D.$
Theorem: The most probable **rooted** gene tree on **four species** may not be the rooted species tree.

Proof: Let the species tree $T^*$ have topology $((A,(B,(C,D)))$, and probability $p_1, p_2$ of coalescence on the edges $e_1$ and $e_2$ above LCA$(C,D)$ and LCA$(B,C)$, respectively, with $0<p_1,p_2<1$. We wish to show that we can pick $p_1, p_2$ so that as *rooted* gene trees, $Pr(gt=(((A,B),(C,D))) > Pr(gt=T^*=(A,(B,(C,D))))$.

(1) Let $\epsilon>0$ be given. Pick $p_1$ and $p_2$ small enough so that with probability at least $1-\epsilon$, all four lineages enter the edge above the root (i.e., $(1-p_1)(1-p_2) > 1-\epsilon$).

(2) Then, all pairs of lineages have equal probability of coalescing. Each rooted tree is defined by the order of coalescent events –

(1) $((A,(B,(C,D))))$ is produced by the sequence of coalescences $A&B$, then $AB&C$, then $ABC&D$.

(2) $((A,B),(C,D))$ is produced by two sequences of coalescence events – $A&B$, then $C&D$, then $AB&CD$, and also by $C&D$, then $A&B$, then $AB&CD$. 
Theorem: The most probable rooted gene tree on four species may not be the rooted species tree.

Proof: Let the species tree $T^*$ have topology $((A,(B,(C,D))),$ and probability $p_1,$ $p_2$ of coalescence on the edges $e_1$ and $e_2$ above $\text{LCA}(C,D)$ and $\text{LCA}(B,C),$ respectively, with $0<p_1,p_2<1$. We wish to show that we can pick $p_1,$ $p_2$ so that as *rooted* gene trees, $\Pr(\text{gt}=((A,B),(C,D))) > \Pr(\text{gt}=T^*=(A,(B,(C,D)))).$

(1) Let $\varepsilon>0$ be given. Pick $p_1$ and $p_2$ small enough so that with probability at least $1-\varepsilon,$ all four lineages enter the edge above the root (i.e., $(1-p_1)(1-p_2) > 1-\varepsilon$).

(2) Then, all pairs of lineages have equal probability of coalescing. Each rooted tree is defined by the order of coalescent events –

(1) $((A,B),(C,D)))$ is produced by the sequence of coalescences $A&B,$ then $AB&C,$ then $ABC&D.$

(2) $((A,B),(C,D)))$ is produced by two sequences of coalescence events – $A&B,$ then $C&D,$ then $AB&CD,$ and also by $C&D,$ then $A&B,$ then $AB&CD.$ Hence, the rooted tree $((A,B),(C,D)))$ is twice as likely as the rooted tree $(A,(B,(C,D)))$, which is the species tree $T^*$!
Theorem: The most probable rooted gene tree on four species may not be the rooted species tree.

Proof: Let the species tree $T^*$ have topology $((A,(B,(C,D))),$ and probability $p_1$, $p_2$ of coalescence on the edges $e_1$ and $e_2$ above $\text{LCA}(C,D)$ and $\text{LCA}(B,C)$, respectively, with $0<p_1,p_2<1$. We wish to show that we can pick $p_1$, $p_2$ so that as *rooted* gene trees,

$$\Pr(gt= ((A,(B,(C,D)))) > \Pr(gt= T^* = (A,(B,(C,D))))).$$

(1) Let $\varepsilon>0$ be given. Pick $p_1$ and $p_2$ small enough so that with probability at least $1-\varepsilon$, all four lineages enter the edge above the root (i.e., $(1-p_1)(1-p_2) > 1-\varepsilon$).

(2) Then, all pairs of lineages have equal probability of coalescing. Each rooted tree is defined by the order of coalescent events –

(1) $((A,(B,(C,D))))$ is produced by the sequence of coalescences $A$&$B$, then $AB$&$C$, then $ABC$&$D$.

(2) $((A,B),(C,D))$ is produced by two sequences of coalescence events – $A$&$B$, then $C$&$D$, then $AB$&$CD$, and also by $C$&$D$, then $A$&$B$, then $AB$&$CD$. Hence, the rooted tree $((A,B),(C,D))$ is twice as likely as the rooted tree $(A,(B,(C,D)))$, which is the species tree $T^*$!

Hence, for some model species tree, the rooted species tree topology is not the most likely rooted gene tree topology.
Problem 1: Let the species tree have topology \((A, (B, (C, D)))\). Show that for any probabilities \(p_1\) and \(p_2\) of coalescence on the internal edges of the species tree, then considering *unrooted* gene trees,

\[
\Pr(gt=((A, B), (C, D))) > \Pr(gt=((A, C), (B, D))) = \Pr(gt=((B, C), (A, D))).
\]

Problem 2: Let the species tree have topology \(((A, B), (C, D))\). Show that for any probabilities of coalescence on the internal edges of the tree, \(p_1\) and \(p_2\) (both strictly between 0 and 1), the most probable *rooted* gene tree is topologically identical to the rooted species tree.
How to compute a species tree?

Technique:
Most frequent gene tree?
How to compute a species tree?

Technique:
Most frequent gene tree?

**YES** if you have only three species

**YES** if you have four species and are content with the unrooted species tree

Otherwise **NO!**
How to compute a rooted species tree from rooted gene trees?
How to compute a rooted species tree from rooted gene trees?

Theorem (Degnan et al., 2006, 2009):
Under the multi-species coalescent model, for any three taxa A, B, and C, the most probable rooted gene tree on \( \{A, B, C\} \) is identical to the rooted species tree induced on \( \{A, B, C\} \).
How to compute a rooted species tree from rooted gene trees?

Estimate species tree for every 3 species

Theorem (Degnan et al., 2006, 2009): Under the multi-species coalescent model, for any three taxa A, B, and C, the most probable rooted gene tree on \{A,B,C\} is identical to the rooted species tree induced on \{A,B,C\}.
How to compute a rooted species tree from rooted gene trees?

Estimate species tree for every 3 species

Theorem (Aho et al.): The rooted tree on n species can be computed from its set of 3-taxon rooted subtrees in polynomial time.
How to compute a rooted species tree from rooted gene trees?

Estimate species tree for every 3 species

Combine rooted 3-taxon trees

Theorem (Aho et al.): The rooted tree on n species can be computed from its set of 3-taxon rooted subtrees in polynomial time.
How to compute a rooted species tree from rooted gene trees?

Theorem (Degnan et al., 2009): Under the multi-species coalescent, the rooted species tree can be estimated correctly (with high probability) given a large enough number of true rooted gene trees.

Combine rooted 3-taxon trees
How to compute an unrooted species tree from unrooted gene trees?

(Pretend these are unrooted trees)
How to compute an unrooted species tree from unrooted gene trees?

(Pretend these are unrooted trees)

Theorem (Allman et al., 2011, and others): For every four leaves \{a,b,c,d\}, the most probable unrooted quartet tree on \{a,b,c,d\} is the true species tree.
How to compute an unrooted species tree from unrooted gene trees?

Theorem (Allman et al., 2011, and others): For every four leaves \{a,b,c,d\}, the most probable unrooted quartet tree on \{a,b,c,d\} is the true species tree.
How to compute an unrooted species tree from unrooted gene trees?

(Pretend these are unrooted trees)  Estimate species tree for every 4 species

Theorem (Allman et al., 2011, and others): For every four leaves \{a,b,c,d\}, the most probable unrooted quartet tree on \{a,b,c,d\} is the true species tree.

Use the All Quartets Method to construct the species tree, based on the most frequent gene trees for each set of four species.
How to compute an unrooted species tree from unrooted gene trees?

(Pretend these are unrooted trees)

Estimate species tree for every 4 species

Combine unrooted 4-taxon trees

Theorem (Allman et al., 2011, and others): For every four leaves \{a,b,c,d\}, the most probable unrooted quartet tree on \{a,b,c,d\} is the true species tree.

Use the All Quartets Method to construct the species tree, based on the most frequent gene trees for each set of four species.
Statistically consistent methods

Methods that require rooted gene trees:
• SRSTE (Simple Rooted Species Tree Estimation) - see textbook
• **MP-EST** (Liu et al. 2010): maximum pseudo-likelihood estimation

Methods that work from unrooted gene trees:
• SUSTE (Simple Unrooted Species Tree Estimation) - see textbook
• **BUCKy-pop** (Ané and Larget 2010): quartet-based Bayesian species tree estimation
• **ASTRAL** (Mirarab et al., 2014) – quartet-based estimation
• **NJst** (Liu and Yu, 2011): distance-based method

Methods that work from sequence alignments
• BEST (Liu 2008) and **BEAST** (Heled, and Drummond): Bayesian co-estimation of gene trees and species trees
• SVD quartets (Kubatko): quartet-based method
• SNAPP and others...
Summary methods are statistically consistent species tree estimators

Here, the “data” are true gene trees
Results on 11-taxon datasets with weak ILS

*BEAST more accurate than summary methods (MP-EST, BUCKy, etc)
CA-ML: concatenated analysis) most accurate

Datasets from Chung and Ané, 2011
Bayzid & Warnow, Bioinformatics 2013
Results on 11-taxon datasets with strong ILS

*BEAST more accurate than summary methods (MP-EST, BUCKy, etc)

CA-ML: (concatenated analysis) also very accurate

Datasets from Chung and Ané, 2011
Bayzid & Warnow, Bioinformatics 2013
*BEAST co-estimation produces more accurate gene trees than Maximum Likelihood

11-taxon weak ILS datasets

17-taxon (very high ILS) datasets

11-taxon datasets from Chung and Ané, Syst Biol 2012
17-taxon datasets from Yu, Warnow, and Nakhleh, JCB 2011

Bayzid & Warnow, Bioinformatics 2013
Impact of Gene Tree Estimation Error on MP-EST

MP-EST has **no error on true gene trees**, but MP-EST has **9% error on estimated gene trees**

Datasets: 11-taxon strongILS conditions with 50 genes

Similar results for other summary methods (MDC, Greedy, etc.).
Problem: poor gene trees

• Summary methods combine estimated gene trees, not true gene trees.
Problem: poor gene trees

• Summary methods combine estimated gene trees, not true gene trees.

• The individual gene sequence alignments in the 11-taxon datasets have poor phylogenetic signal, and result in poorly estimated gene trees.
Problem: poor gene trees

- Summary methods combine estimated gene trees, not true gene trees.
- The individual gene sequence alignments in the 11-taxon datasets have poor phylogenetic signal, and result in poorly estimated gene trees.
- Species trees obtained by combining poorly estimated gene trees have poor accuracy.
Summary methods combine estimated gene trees, not true gene trees.

The individual gene sequence alignments in the 11-taxon datasets have poor phylogenetic signal, and result in poorly estimated gene trees.

Species trees obtained by combining poorly estimated gene trees have poor accuracy.
Addressing gene tree estimation error

- Get better estimates of the gene trees
- Restrict to subset of estimated gene trees
- Model error in the estimated gene trees
- Modify gene trees to reduce error
- Develop methods with greater robustness to gene tree error
Addressing gene tree estimation error

- Get better estimates of the gene trees
- Restrict to subset of estimated gene trees
- Model error in the estimated gene trees
- Modify gene trees to reduce error
- Develop methods with greater robustness to gene tree error
  - ASTRAL. Bioinformatics 2014 (Mirarab et al.)
  - Statistical binning. Science 2014 (Mirarab et al.)
Avian Phylogenomics Project

E Jarvis, HHMI
MTP Gilbert, Copenhagen
G Zhang, BGI
T. Warnow UT-Austin
S. Mirarab UT-Austin
Md. S. Bayzid, UT-Austin

Plus many many other people...

• Approx. 50 species, whole genomes
• 8000+ genes, UCEs
• Gene sequence alignments computed using SATé (Liu et al., Science 2009 and Systematic Biology 2012)

Species tree estimated using Statistical Binning with MP-EST (Jarvis, Mirarab, et al., Science 2014)
Avian whole genomes phylogenies

[Jarvis*, Mirarab*, et al., Science, 2014]

- International team of more than 100 researchers

- Whole genomes for 48 bird species (~100 million years of evolution)

- Goal: a phylogeny of major bird lineages
  - Extremely challenging due to rampant gene tree incongruence
  - Implications for traits such as vocal learning
  - 14,000 “genes” (typically short and relatively conserved)
Gene trees on the avian dataset

A measure of confidence in estimated gene tree branches

14,000 genes from avian genome-scale data
[Jarvis*, Mirarab*, et al., Science, 2014]
Gene trees on the avian dataset

14,000 genes from avian genome-scale data
[Jarvis*, Mirarab*, et al., Science, 2014]

Avian-like simulations (1000 genes)
Gene trees on the avian dataset

Gene tree error matters

[Ané, et al, MBE, 2007]
[Patel, et al, MBE, 2013]
[Mirarab, et al., Systematic Biology, 2014]

14,000 genes from avian genome-scale data
[Jarvis*, Mirarab*, et al., Science, 2014]

Avian-like simulations (1000 genes)
• The individual gene sequence alignments in the avian datasets have poor phylogenetic signal, and result in poorly estimated gene trees.
• The individual gene sequence alignments in the avian datasets have poor phylogenetic signal, and result in poorly estimated gene trees.

• Species trees obtained by combining poorly estimated gene trees have poor accuracy.
• The individual gene sequence alignments in the avian datasets have poor phylogenetic signal, and result in poorly estimated gene trees.

• Species trees obtained by combining poorly estimated gene trees have poor accuracy.

• There are no theoretical guarantees for summary methods except for perfectly correct gene trees.
• The individual gene sequence alignments in the avian datasets have poor phylogenetic signal, and result in poorly estimated gene trees.

• Species trees obtained by combining poorly estimated gene trees have poor accuracy.

• There are no theoretical guarantees for summary methods except for perfectly correct gene trees.

COMMON PHYLOGENOMICS PROBLEM: many poor gene trees
Idea: combine best aspects of concatenation and summary methods

- Concatenation (fully partitioned) works fine when the concatenated data evolve under identical (or very similar) trees
- Some pairs of genes are not discordant (at least in topology)
- Concatenate “combinable” sets of genes into “supergenes” to increase the phylogenetic signal
- But how do we know which genes are combinable if we cannot estimate them correctly?
Statistical binning

Input: estimated gene trees with bootstrap support, and minimum support threshold $t$

Step 1: partition of the estimated gene trees into sets, so that no two gene trees in the same set are strongly incompatible, and the sets have approximately the same size.

Step 2: estimate “supergene” trees on each set using concatenation (maximum likelihood)

Step 3: combine supergene trees using coalescent-based method

Note: Step 1 requires solving the NP-hard “balanced vertex coloring problem”, for which we developed a good heuristic (modified 1979 Brelaz algorithm)
Statistical binning vs. unbinned

Binning produces bins with approximate 5 to 7 genes each

Datasets: 11-taxon strongILS datasets with 50 genes, Chung and Ané, Systematic Biology
Avian Simulation: Impact of binning with MP-EST

(a) Varying gene tree estimation error

(b) Varying the level of ILS
Comparing Binned and Un-binned MP-EST on the Avian Dataset

Unbinned MP-EST strongly rejects Columbea, a major finding by Jarvis, Mirarab, et al.
Summary so far

- Standard coalescent-based methods (such as MP-EST) have poor accuracy in the presence of gene tree error.

- Statistical binning improves the estimation of gene tree distributions, and so:
  - Improves species tree estimation
Summary so far

- Standard coalescent-based methods (such as MP-EST) have poor accuracy in the presence of gene tree error.
- Statistical binning improves the estimation of gene tree distributions, and so:
  - Improves species tree estimation
  - Improves species tree branch lengths
Summary so far

- Standard coalescent-based methods (such as MP-EST) have poor accuracy in the presence of gene tree error.

- Statistical binning improves the estimation of gene tree distributions, and so:
  - Improves species tree estimation
  - Improves species tree branch lengths
  - Reduces incidence of strongly supported false positive branches
1KP: Plant whole transcriptomes

[Wickett*, Mirarab*, et al., PNAS, 2014]

Phylotranscriptomic analysis of the origin and early diversification of land plants

Norman J. Wickett\textsuperscript{a,b,1,2}, Siavash Mirarab\textsuperscript{c,1}, Nam Nguyen\textsuperscript{1}, Tandy Warnow\textsuperscript{1}, Eric Carpenter\textsuperscript{d}, Naim Matasci\textsuperscript{e,f}, Saravanaraj Ayyampalayam\textsuperscript{g}, Michael S. Barker\textsuperscript{f}, J. Gordon Burleigh\textsuperscript{h}, Matthew A. Gitzendanner\textsuperscript{h,i}, Brad R. Ruhfel\textsuperscript{h,j,k}, Eric Wafulla\textsuperscript{1}, Joshua P. Der\textsuperscript{l}, Sean W. Graham\textsuperscript{m,n}, Sarah Mathews\textsuperscript{n}, Michael Melkonian\textsuperscript{o}, Douglas E. Soltis\textsuperscript{p,q}, Pamela S. Soltis\textsuperscript{r,s}, Nicholas W. Miles\textsuperscript{t}, Carl J. Rothfels\textsuperscript{h,u}, Lisa Pokorny\textsuperscript{t,v}, A. Jonathan Shaw\textsuperscript{t}, Lisa DeGironimo\textsuperscript{t}, Dennis W. Stevenson\textsuperscript{t}, Barbara Surek\textsuperscript{t}, Juan Carlos Villarreal\textsuperscript{t}, Béatrice Roure\textsuperscript{t}, Hervé Philippe\textsuperscript{w,x}, Claude W. dePamphilis\textsuperscript{t}, Tao Chen\textsuperscript{y}, Michael K. Deyholos\textsuperscript{t}, Regina S. Baucom\textsuperscript{t}, Toni M. Kutchan\textsuperscript{t}, Megan M. Augustin\textsuperscript{t}, Jum Wang\textsuperscript{t}, Yong Zhang\textsuperscript{t}, Zhijian Tian\textsuperscript{t}, Zhixiang Yan\textsuperscript{t}, Xiaolei Wu\textsuperscript{t}, Xiao Sun\textsuperscript{t}, Gane Ka-Shu Wong\textsuperscript{c,1,2}, and James Leebens-Mack\textsuperscript{c,2}

- Whole transcriptomes for 103 plant species
- 1,200 in the next phase
- 400-800 single copy “genes”
- Spans ~1 billion years of evolution
- Many unanswered questions about plant evolution
Summary methods on the 1KP data (103 plants)

- Existing summary methods produced species trees with low support and unbelievable relationships
- .. despite having gene trees with relatively high bootstrap support

400 genes from 1KP data
[Wickett*, Mirarab*, et al., PNAS, 2014]
Summary methods on the 1KP data (103 plants)

• Existing summary methods produced species trees with low support and unbelievable relationships

• ... despite having gene trees with relatively high bootstrap support

• Our simulation studies showed that the reason had to do with the number of taxa

400 genes from 1KP data [Wickett*, Mirarab*, et al., PNAS, 2014]

1000 simulated genes, “medium” levels of ILS [Mirarab and Warnow, ISMB, 2015]
Summary methods on the 1KP data (103 plants)

- Existing summary methods produced species trees with low support and unbelievable relationships

- Despite having gene trees with relatively high bootstrap support

- Our simulation studies showed that the reason had to do with the number of taxa

400 genes from 1KP data
[Wickett*, Mirarab*, et al., PNAS, 2014]

1000 simulated genes, “medium” levels of ILS
[Mirarab and Warnow, ISMB, 2015]
Summary methods on the 1KP data (103 plants)

- Existing summary methods produced species trees with low support and unbelievable relationships
- .. despite having gene trees with relatively high bootstrap support
- Our simulation studies showed that the reason had to do with the number of taxa

The problem size (# species) matters too!

400 genes from 1KP data [Wickett*, Mirarab*, et al., PNAS, 2014]

1000 simulated genes, “medium” levels of ILS [Mirarab and Warnow, ISMB, 2015]
1KP: Thousand Transcriptome Project

- 1200 plant transcriptomes
- More than 13,000 gene families (most not single copy)
- Gene sequence alignments and trees computed using SATe (Liu et al., Science 2009 and Systematic Biology 2012)

Species tree estimated using ASTRAL (Bioinformatics, 2014)

1KP paper by Wickett, Mirarab et al., PNAS 2014
ASTRAL

• Accurate Species Trees Algorithm
• Mirarab et al., ECCB 2014 and Bioinformatics 2014
• Statistically-consistent estimation of the species tree from unrooted gene trees
ASTRAL’s approach

• Input: set of unrooted gene trees $T_1, T_2, \ldots, T_k$
• Output: Tree $T^*$ maximizing the total quartet-similarity score to the unrooted gene trees

Theorem:

• An exact solution to this problem would be a statistically consistent algorithm in the presence of ILS
ASTRAL’s approach

- **Input:** set of unrooted gene trees $T_1, T_2, \ldots, T_k$
- **Output:** Tree $T^*$ maximizing the total quartet-similarity score to the unrooted gene trees

**Theorem:**
- An exact solution to this problem is NP-hard

**Comment:** unknown computational complexity if all trees $T_i$ are on the same leaf set
ASTRAL’s approach

• Input: set of unrooted gene trees $T_1, T_2, \ldots, T_k$ and set $X$ of bipartitions on species set $S$

• Output: Tree $T^*$ maximizing the total quartet-similarity score to the unrooted gene trees, subject to $\text{Bipartitions}(T^*)$ drawn from $X$

Theorem:

• An exact solution to this problem is achievable in polynomial time!
ASTRAL’s approach

• Input: set of unrooted gene trees $T_1, T_2, \ldots, T_k$ and set $X$ of bipartitions on species set $S$

• Output: Tree $T^*$ maximizing the total quartet-similarity score to the unrooted gene trees, subject to $\text{Bipartitions}(T^*)$ drawn from $X$

Theorem:

• Letting $X$ be the set of bipartitions from the input gene trees is statistically consistent and polynomial time.
ASTRAL vs. MP-EST and Concatenation

200 genes, 500bp

Mammalian Simulation Study, Varying ILS level

Less ILS
ASTRAL on plants dataset

- The ASTRAL tree:
  - High support
  - Similar to concatenation with some interesting differences (e.g., recovered bryophytes)
  - ASTRAL took only about 10 minutes (serial running time) on 103 taxa and 400 genes

[Wickett*, Mirarab*, et al., PNAS, 2014]
1kp: Thousand Transcriptome Project

Upcoming Challenges (~1200 species, ~400 loci):
• Species tree estimation under the multi-species coalescent from hundreds of conflicting gene trees on >1000 species; we will use ASTRAL-2 (Mirarab and Warnow, 2015)

• Multiple sequence alignment of >100,000 sequences (with lots of fragments!) – we will use UPP (Nguyen et al., in press)
ASTRAL-II on biological datasets (ongoing collaborations)

- 1200 plants with ~ 400 genes (1KP consortium)
- 250 avian species with 2000 genes (with LSU, UF, and Smithsonian)
- 200 avian species with whole genomes (with Genome 10K, international)
- 250 suboscine species (birds) with ~2000 genes (with LSU and Tulane)
- 140 Insects with 1400 genes (with U. Illinois at Urbana-Champaign)
- 50 Hummingbird species with 2000 genes (with U. Copenhagen and Smithsonian)
- 40 raptor species (birds) with 10,000 genes (with U. Copenhagen and Berkeley)
- 38 mammalian species with 10,000 genes (with U. of Bristol, Cambridge, and Nat. Univ. of Ireland)
Summary

• Gene tree estimation error (e.g., due to insufficient phylogenetic signal) is a typical occurrence in phylogenomics projects.

• Standard summary methods for coalescent-based species tree estimation (e.g., MP-EST) are impacted by gene tree error.

• Statistical binning improves the estimation of gene tree distributions – and hence leads to improved species tree estimation when gene trees have insufficient accuracy.

• ASTRAL is a statistically consistent summary method that is quartet-based, and can analyze very large datasets (1000 genes and 1000 species).
Papers and Software

- ASTRAL (Mirarab et al., ECCB and Bioinformatics 2014 (paper 115)
- ASTRAL-2 (Mirarab and Warnow, ISMB 2015, not yet available, paper 125)
- Statistical Binning (Mirarab et al., Science 2014, paper 122)

- SVDQuartets (Chifman and Kubatko, Bioinformatics 2014)
- BUCKy (Larget et al., Bioinformatics 26(22):2910–2911, 2010)

Open source software for ASTRAL and Statistical Binning available at github.

Datasets (simulated and biological) available online.
Research Projects

• Many summary methods are either triplet-based or quartet-based methods, and so could potentially be improved by using better triplet- or quartet-amalgamation techniques.

• NJst runs NJ on a distance matrix it computes. It might be improved by using a better distance-based technique!

• Many summary methods are not designed for large numbers of species, and so improvements in running time might be easily obtained through better algorithm design (and re-implementation). Perhaps even HPC implementations.
Homework

• Statistical binning is an attempt to improve gene tree estimation, given a set of multiple sequence alignments on different “genes” (loci), and uses bootstrap support to determine if two genes are combinable. What other techniques could be used instead of bootstrap support? What are the pros and cons of the different approaches?

• ASTRAL is a quartet-based method for estimating a species tree from unrooted gene trees. It is constrained by how the set “X” is defined. Discuss the problems involved with computing X – the consequences if X is too small, or if it is too large. Suggest a technique for computing X, and explain why you like the technique. Is ASTRAL run with your technique for computing X statistically consistent under the multi-species coalescent model?

• ASTRAL could be re-implemented to work with rooted gene trees instead of unrooted gene trees. Discuss the advantages and disadvantages of this change.

• Discuss how you would assess confidence in branches of a species tree estimated using a coalescent method.