From Gene Trees to Species Trees

Tandy Warnow
The University of Texas at Austin
Phylogeny (evolutionary tree)

From the Tree of the Life Website,
University of Arizona
Estimating The Tree of Life: a Grand Challenge

Most well studied problem:

Given DNA sequences, find the Maximum Likelihood Tree

NP-hard, lots of software (RAxML, FastTree-2, GARLI, etc.)
Estimating The Tree of Life: a *Grand Challenge*

Novel techniques needed for scalability and accuracy: NP-hard problems and large datasets
Current methods not good enough on large datasets
HPC is necessary but not sufficient
Phylogenomics

(Phylogenetic estimation from whole genomes)
Sampling multiple genes from multiple species

From the Tree of 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>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>$S_5$</td>
</tr>
<tr>
<td>$S_7$</td>
<td>TCTAATGGAC</td>
<td>$S_6$</td>
</tr>
<tr>
<td>$S_8$</td>
<td>TATAACGGAA</td>
<td>$S_7$</td>
</tr>
<tr>
<td>$S_7$</td>
<td>GCTAAGGGAA</td>
<td>$S_8$</td>
</tr>
</tbody>
</table>
Two competing approaches

Species

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

Concatenation

Analyze separately

Summary Method
1kp: Thousand Transcriptome Project

- Plant Tree of Life based on transcriptomes of ~1200 species
- More than 13,000 gene families (most not single copy)
- Gene sequence alignments and trees computed using SATé (Liu et al., Science 2009 and Systematic Biology 2012)

Challenges:
- Multiple sequence alignments of > 100,000 sequences
- Gene tree incongruence
Avian Phylogenomics Project

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

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

Challenges:
Maximum likelihood on multi-million-site sequence alignments
Massive gene tree incongruence
The Tree of Life: **Multiple Challenges**

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

Also:
- 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
The Tree of Life: **Multiple Challenges**

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

“BigData” complexity

Also:

- **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

Tomorrow’s talk
The Tree of Life: **Multiple Challenges**

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

Also:  
- 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
This talk

Species tree estimation from multiple genes

– Mathematical foundations
– Algorithms
– Data challenges
– New statistical questions
– Avian Phylogenomics
Part I: Gene Tree Estimation
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, it changes with equal probability to each of the remaining states.
- The evolutionary process is Markovian.

More complex models (such as the General Markov model) are also considered, often with little change to the theory.
Quantifying Error

TRUE TREE

FN: false negative (missing edge)
FP: false positive (incorrect edge)

50% error rate

DNA SEQUENCES

S1  ACAATTAGAAC
S2  ACCCTTAGAAC
S3  ACCATTCGAC
S4  ACCACGACCGGA
S5  ACCAGACCAG

INFERRED TREE
Questions

• Is the model tree identifiable?
• Which estimation methods are statistically consistent under this model?
• How much data does the method need to estimate the model tree correctly (with high probability)?
• What is the computational complexity of an estimation problem?
Statistical Consistency

error

Data
Statistical Consistency

Data are sites in an alignment
Neighbor Joining (and many other distance-based methods) are statistically consistent under Jukes-Cantor.
Neighbor Joining on large diameter trees

Simulation study based upon fixed edge lengths, K2P model of evolution, sequence lengths fixed to 1000 nucleotides.

Error rates reflect proportion of incorrect edges in inferred trees.

[Nakhleh et al. ISMB 2001]
“Convergence rate” or sequence length requirement

The sequence length (number of sites) that a phylogeny reconstruction method \( M \) needs to reconstruct the true tree with probability at least \( 1 - \varepsilon \) depends on

- \( M \) (the method)
- \( \varepsilon \)
- \( f = \min p(e) \)
- \( g = \max p(e) \), and
- \( n = \) the number of leaves

We fix everything but \( n \).
Theorem (Erdos et al. 1999, Atteson 1999):

Various distance-based methods (including Neighbor joining) will return the true tree with high probability given sequence lengths that are \textit{exponential} in the evolutionary diameter of the tree (hence, \textit{exponential in n}).

Proof:

• the method returns the true tree if the estimated distance matrix is close to the model tree distance matrix

• the sequence lengths that suffice to achieve bounded error are exponential in the evolutionary diameter.
Afc methods (Warnow et al., 1999)

A method M is “absolute fast converging”, or afc, if for all positive f, g, and ε, there is a polynomial p(n) s.t. Pr(M(S)=T) > 1- ε, when S is a set of sequences generated on T of length at least p(n).

Notes:

1. The polynomial p(n) will depend upon M, f, g, and ε.
2. The method M is not “told” the values of f and g.
Statistical consistency, exponential convergence, and absolute fast convergence (afc)
Fast-converging methods (and related work)

- 1997: Erdos, Steel, Szekely, and Warnow (ICALP).
- 1999: Erdos, Steel, Szekely, and Warnow (RSA, TCS);
  Huson, Nettles and Warnow (J. Comp Bio.)
- **2001:** Warnow, St. John, and Moret (SODA);
  Nakhleh, St. John, Roshan, Sun, and Warnow (ISMB)
  Cryan, Goldberg, and Goldberg (SICOMP);
  Csuros and Kao (SODA);
- 2002: Csuros (J. Comp. Bio.)
- 2006: Daskalakis, Mossel, Roch (STOC),
  Daskalakis, Hill, Jaffe, Mihaescu, Mossel, and Rao (RECOMB)
- 2007: Mossel (IEEE TCBB)
- 2008: Gronau, Moran and Snir (SODA)
- 2010: Roch (Science)
- 2013: Roch (in preparation)
Neighbor Joining on large diameter trees

Simulation study based upon fixed edge lengths, K2P model of evolution, sequence lengths fixed to 1000 nucleotides.

Error rates reflect proportion of incorrect edges in inferred trees.

[Nakhleh et al. ISMB 2001]
DCM1-boosting distance-based methods

[Nakhleh et al. ISMB 2001]

• Theorem (Warnow et al., SODA 2001): DCM1-NJ converges to the true tree from polynomial length sequences. Hence DCM1-NJ is afc.

Proof: uses chordal graph theory and probabilistic analysis of algorithms
Questions

• Is the model tree identifiable?
• Which estimation methods are statistically consistent under this model?
• How much data does the method need to estimate the model tree correctly (with high probability)?
• What is the computational complexity of an estimation problem?
Answers?

• We know a lot about which site evolution models are identifiable, and which methods are statistically consistent.
Answers?

- We know a lot about which site evolution models are identifiable, and which methods are statistically consistent.
- Some polynomial time afc methods have been developed, and we know a little bit about the sequence length requirements for standard methods.
Answers?

• We know a lot about which site evolution models are identifiable, and which methods are statistically consistent.

• Some polynomial time afc methods have been developed, and we know a little bit about the sequence length requirements for standard methods.

• Just about everything is NP-hard, and the datasets are big.
Answers?

• We know a lot about which site evolution models are identifiable, and which methods are statistically consistent.
• Some polynomial time afc methods have been developed, and we know a little bit about the sequence length requirements for standard methods.
• Just about everything is NP-hard, and the datasets are big.
• Extensive studies show that even the best methods produce gene trees with some error.
In other words...

Statistical consistency doesn’t guarantee accuracy w.h.p. unless the sequences are long enough.
1. Why do we need whole genomes?
2. Will whole genomes make phylogeny estimation easy?
3. How hard are the computational problems?
4. Do we have sufficient methods for this?

(Phylogenetic estimation from whole genomes)
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>$S_5$</td>
</tr>
<tr>
<td>$S_7$</td>
<td>TCTAATGGAC</td>
<td>$S_6$</td>
</tr>
<tr>
<td>$S_8$</td>
<td>TATAACGGAA</td>
<td>$S_7$</td>
</tr>
</tbody>
</table>

---

<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>$S_5$</td>
</tr>
<tr>
<td>$S_7$</td>
<td>TCTAATGGAC</td>
<td>$S_6$</td>
</tr>
<tr>
<td>$S_8$</td>
<td>TATAACGGAA</td>
<td>$S_7$</td>
</tr>
</tbody>
</table>
## Concatenation

<table>
<thead>
<tr>
<th></th>
<th>gene 1</th>
<th>gene 2</th>
<th>gene 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_1$</td>
<td>TCTAATGGAA</td>
<td>??????????????</td>
<td>TATTGATA</td>
</tr>
<tr>
<td>$S_2$</td>
<td>GCTAAGGGAA</td>
<td>??????????????</td>
<td>???????????</td>
</tr>
<tr>
<td>$S_3$</td>
<td>TCTAAGGGAA</td>
<td>??????????????</td>
<td>TCTTGATA</td>
</tr>
<tr>
<td>$S_4$</td>
<td>TCTAACGGA</td>
<td>GGTAACCCTC</td>
<td>TAGTGATGCA</td>
</tr>
<tr>
<td>$S_5$</td>
<td>??????????</td>
<td>GCTAAACCTC</td>
<td>???????????</td>
</tr>
<tr>
<td>$S_6$</td>
<td>??????????</td>
<td>GGTGACCATC</td>
<td>???????????</td>
</tr>
<tr>
<td>$S_7$</td>
<td>TCTAATGGAC</td>
<td>GCTAAACCTC</td>
<td>TAGTGATGCA</td>
</tr>
<tr>
<td>$S_8$</td>
<td>TATAACGGAA</td>
<td>??????????????</td>
<td>CATTCAT</td>
</tr>
</tbody>
</table>

![Diagram showing the concatenation process]
Red gene tree ≠ species tree
(green gene tree okay)
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)
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)
Gene Tree Incongruence

- Gene trees can differ from the species tree due to:
  - Duplication and loss
  - Horizontal gene transfer
  - Incomplete lineage sorting (ILS)
Part II: Species Tree Estimation in the presence of ILS

- Mathematical model: Kingman’s coalescent
- “Coalescent-based” species tree estimation methods
- Simulation studies evaluating methods
- New techniques to improve methods
- Application to the Avian Tree of Life
Species tree estimation: difficult, even for small datasets!

From the Tree of the 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”.
Key observation:
Under the multi-species coalescent model, the species tree defines a *probability distribution on the gene trees*
Incomplete Lineage Sorting (ILS)

- 2000+ 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.
Two competing approaches

Gene 1   Gene 2   ...   Gene k
Species

Concatenation

Analyze separately

Summary Method
How to compute a species 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
How to compute a species tree?

Techniques:
- MDC?
- Most frequent gene tree?
- Consensus of gene trees?
- Other?
How to compute a species tree?

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 species tree?

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\}. 

Estimate species tree for every 3 species
How to compute a species tree?

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 species tree?

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 species tree?

Theorem (Degnan et al., 2009): Under the multispecies 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 a species tree?

Estimate species tree for every 3 species

Combine rooted 3-taxon 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.

Theorem (Allman et al., 2011): the unrooted species tree can be estimated from a large enough number of true unrooted gene trees.
How to compute a species tree?

Estimate species tree for every 3 species

Combine rooted 3-taxon 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.

Theorem (Allman et al., 2011): the unrooted species tree can be estimated from a large enough number of true unrooted gene trees.
Data are gene trees, presumed to be randomly sampled true gene trees.
Questions

• Is the model tree **identifiable**?
• Which estimation methods are **statistically consistent** under this model?
• **How much data** does the method need to estimate the model tree correctly (with high probability)?
• What is the **computational complexity** of an estimation problem?
Statistically consistent under ILS?

- MP-EST (Liu et al. 2010): maximum likelihood estimation of rooted species tree – YES
- BUCKy-pop (Ané and Larget 2010): quartet-based Bayesian species tree estimation – YES
- MDC – NO
- Greedy – NO
- Concatenation under maximum likelihood – open
- MRP (supertree method) – open
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
Gene Tree Estimation:
*BEAST vs. Maximum Likelihood

11-taxon weak ILS datasets
17-taxon (very high ILS) datasets

*BEAST produces more accurate gene trees than ML on gene sequence alignments

11-taxon datasets from Chung and Ané, Syst Biol 2012
17-taxon datasets from Yu, Warnow, and Nakhleh, JCB 2011
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 strong ILS 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.**
TYPICAL PHYLOGENOMICS PROBLEM: many 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.
Questions

• Is the model species tree identifiable?

• Which estimation methods are statistically consistent under this model?

• How much data does the method need to estimate the model species tree correctly (with high probability)?

• What is the computational complexity of an estimation problem?
Questions

• Is the model species tree identifiable?

• Which estimation methods are statistically consistent under this model?

• How much data does the method need to estimate the model species tree correctly (with high probability)?

• What is the computational complexity of an estimation problem?

• What is the impact of error in the input data on the estimation of the model species tree?
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
• “Bin-and-conquer”
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
• “Bin-and-conquer”
Technique #2: Bin-and-Conquer?

1. Assign genes to “bins”, creating “supergene alignments”

2. Estimate trees on each supergene alignment using maximum likelihood

3. Combine the supergene trees together using a summary method
Technique #2: Bin-and-Conquer?

1. Assign genes to “bins”, creating “supergene alignments”

2. Estimate trees on each supergene alignment using maximum likelihood

3. Combine the supergene trees together using a summary method

Variants:

• Naïve binning (Bayzid and Warnow, Bioinformatics 2013)
• Statistical binning (Mirarab, Bayzid, and Warnow, in preparation)
Statistical binning

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

Output: partition of the estimated gene trees into sets, so that no two gene trees in the same set are strongly incompatible.
Statistical binning

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

Output: partition of the estimated gene trees into sets, so that no two gene trees in the same set are strongly incompatible.

Vertex coloring problem (NP-hard),
but good heuristics are available (e.g., Brelaz 1979)

However, for statistical inference reasons, we need balanced vertex color classes
Balanced Statistical Binning

Mirarab, Bayzid, and Warnow, in preparation
Modification of Brelaz Heuristic for minimum vertex coloring.
Sta<cal	
  binning	
  vs.	
  unbinned

Mirarab, et al. in preparation
Datasets: 11-taxon strongI LS datasets with 50 genes, Chung and Ané, Systematic Biology
Mammalian Simulation Study

Observations:

Binning can improve accuracy, but impact depends on accuracy of estimated gene trees and phylogenetic estimation method.

Binned methods can be more accurate than RAxML (maximum likelihood), even when unbinned methods are less accurate.

Data: 200 genes, 20 replicate datasets, based on Song et al. PNAS 2012

Mirarab et al., in preparation
Mammalian simulation

Binning can improve summary methods, but amount of improvement depends on: method, amount of ILS, and accuracy of gene trees.

MP-EST is statistically consistent in the presence of ILS; Greedy is not, unknown for MRP and RAxML.
Data (200 genes, 20 replicate datasets) based on Song et al. PNAS 2012
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...

Gene Tree Incongruence
Avian Phylogeny

- **GTRGAMMA Maximum likelihood analysis (RAxML) of 37 million basepair alignment (exons, introns, UCEs)** – highly resolved tree with near 100% bootstrap support.

- **More than 17 years of compute time**, and used 256 GB. Run at HPC centers.

Avian Phylogenomics Project, in preparation
Avian Phylogeny

- GTRGAMMA Maximum likelihood analysis (RAxML) of 37 million basepair alignment (exons, introns, UCEs) – highly resolved tree with near 100% bootstrap support.

- More than 17 years of compute time, and used 256 GB. Run at HPC centers.

- Unbinned MP-EST on 14000+ genes: highly incongruent with the concatenated maximum likelihood analysis, poor bootstrap support.

- Sta**c**al binning: faster than concatenated analysis, highly parallelized.

Avian Phylogenomics Project, in preparation
Avian Simulation – 14,000 genes

- **MP-EST:**
  - Unbinned  ~ 11.1% error

- **Greedy:**
  - Unbinned  ~ 26.6% error

- 8250 exon-like genes (27% avg. bootstrap support)
- 3600 UCE-like genes (37% avg. bootstrap support)
- 2500 intron-like genes (51% avg. bootstrap support)
Avian Simulation – 14,000 genes

- **MP-EST:**
  - Unbinned ~ 11.1% error
  - Binned ~ 6.6% error

- **Greedy:**
  - Unbinned ~ 26.6% error
  - Binned ~ 13.3% error

- 8250 exon-like genes (27% avg. bootstrap support)
- 3600 UCE-like genes (37% avg. bootstrap support)
- 2500 intron-like genes (51% avg. bootstrap support)
Avian Phylogeny

• GTRGAMMA Maximum likelihood analysis (RAxML) of 37 million basepair alignment (exons, introns, UCEs) – highly resolved tree with near 100% bootstrap support.

• More than 17 years of compute time, and used 256 GB. Run at HPC centers.

• Unbinned MP-EST on 14000+ genes: highly incongruent with the concatenated maximum likelihood analysis, poor bootstrap support.

Avian Phylogenomics Project, in preparation
Avian Phylogeny

• GTRGAMMA Maximum likelihood analysis (RAxML) of 37 million basepair alignment (exons, introns, UCEs) – highly resolved tree with near 100% bootstrap support.

• More than 17 years of compute time, and used 256 GB. Run at HPC centers.

• Unbinned MP-EST on 14000+ genes: highly incongruent with the concatenated maximum likelihood analysis, poor bootstrap support.

• Statistical binning version of MP-EST on 14000+ gene trees – highly resolved tree, largely congruent with the concatenated analysis, good bootstrap support

Avian Phylogenomics Project, in preparation
To consider

• Binning reduces the amount of data (number of gene trees) but can improve the accuracy of individual “supergene trees”. The response to binning differs between methods. Thus, there is a trade-off between data quantity and quality, and not all methods respond the same to the trade-off.

• We know very little about the impact of data error on methods. We do not even have proofs of statistical consistency in the presence of data error.
Basic Questions

• Is the model tree identifiable?
• Which estimation methods are statistically consistent under this model?
• How much data does the method need to estimate the model tree correctly (with high probability)?
• What is the computational complexity of an estimation problem?
Additional Statistical Questions

• Trade-off between data quality and quantity
• Impact of data selection
• Impact of data error
• Performance guarantees on finite data (e.g., prediction of error rates as a function of the input data and method)

We need a solid mathematical framework for these problems.
Summary

• DCM1-NJ: an absolute fast converging (afc) method, uses chordal graph theory and probabilistic analysis of algorithms to prove performance guarantees

• Binning: species tree estimation from multiple genes, can improve coalescent-based species tree estimation methods.

• New questions in phylogenetic estimation about impact of error in input data.
Warnow Laboratory

PhD students: Siavash Mirarab*, Nam Nguyen, and Md. S. Bayzid**
Undergrad: Keerthana Kumar
Lab Website: [http://www.cs.utexas.edu/users/phylo](http://www.cs.utexas.edu/users/phylo)

**Funding**: Guggenheim Foundation, Packard, NSF, Microsoft Research New England, David Bruton Jr. Centennial Professorship, and TACC (Texas Advanced Computing Center)

**TACC and UTCS computational resources**
* Supported by HHMI Predoctoral Fellowship
** Supported by Fulbright Foundation Predoctoral Fellowship
Metazoa Dataset from Salichos & Rokas - Nature 2013
225 genes and 21 species

- UnBinned MP-EST compared to Concatenation using RAxML
  - Poor bootstrap support
  - Substantial conflict with concatenation (red is conflict - green/black is congruence)
RAxML on combined datamatrix

Binned MP-EST

MP-EST unbinned
Binned vs. unbinned analyses

- 75%-threshold for binning
- Number of species: 21 for both
- Number of “genes”
  - Unbinned: 225 genes
  - Binned: 17 supergenes
- Gene tree average bootstrap support
  - Unbinned: 47%
  - Binned: 78%
- Species tree bootstrap support
  - Unbinned: avg 83%, 11 above 75%, 10 above 90%
  - Binned: avg 89%, 15 above 75%, 12 above 90%
Naïve binning vs. unbinned: 50 genes

Bayzid and Warnow, Bioinformatics 2013
11-taxon strong ILS datasets with 50 genes, 5 genes per bin
Naïve binning vs. unbinned, 100 genes

*BEAST did not converge on these datasets, even with 150 hours. With binning, it converged in 10 hours.
Naïve binning vs. unbinned: 50 genes

Bayzid and Warnow, Bioinformatics 2013
11-taxon strongILS datasets with 50 genes, 5 genes per bin
Avian Simulation study – binned vs. unbinned, and RAxML
Mammals Simulation
Avian Simulation

![Diagram showing tree error for different datasets and methods: Greedy, MRP, MP-EST, RA-ML. The datasets include 1X, 200 genes, 250bp, 1X, 200 genes, 500bp, 1X, 200 genes, 1000bp, 1X, 600 genes, Mixed, 0.5X, 200 genes, 500bp, 2X, 200 genes, 500bp. The error rates are measured as a percentage, and the methods are compared for their performance.]