Progress and Challenges for Large-Scale Phylogeny Estimation

Tandy Warnow
Departments of Computer Science and Bioengineering
Carl R. Woese Institute for Genomic Biology
The University of Illinois at Urbana-Champaign
http://tandy.cs.illinois.edu
Species Tree

From the Tree of the Life Website,
University of Arizona
Phylogenies and Applications

Basic Biology:
How did life evolve?

Applications of phylogenies to:
protein structure and function
population genetics
human migrations
metagenomics
Phylogenomics

(Phylogenetic estimation from whole genomes)
Phylogenomic pipeline

- Select taxon set and markers
- Gather and screen sequence data, possibly identify orthologs
- Compute multiple sequence alignments for each locus
- Compute species tree or network:
  - Compute gene trees on the alignments and combine the estimated gene trees, OR
  - Perform “concatenation analysis” (aka “combined analysis”)
- 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**
- **Compute species tree or network:**
  - Compute gene trees on the alignments and combine the estimated gene trees, OR
  - Perform “concatenation analysis” (aka “combined analysis”)
- 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
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, 14,000 loci
• Jarvis, Mirarab, et al., Science 2014

Major challenges:
• Maximum likelihood analysis of concatenated alignment took 250 CPU years using supercomputers around the globe
• Massive gene tree heterogeneity consistent with incomplete lineage sorting
1kp: Thousand Transcriptome Project

- Plant Tree of Life based on transcriptomes
- First publication: PNAS 2014 (about 100 species and 800 genes)
- Second analysis underway, much larger dataset (~1200 species and ~1000 loci)

**Challenges:**
Massive gene tree conflict consistent with ILS
Alignment of datasets with > 100,000 sequences

Plus many many other people...
Phylogenetic Estimation: Big Data Challenges

**NP-hard problems**

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

“BigData” complexity
- missing data
- mixture models
- errors in input data
- model misspecification
- streaming data
Two of my favorite Challenges

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

• **Phylogenomics**: Methods for multi-locus species tree estimation that are robust to gene tree incongruence due to incomplete lineage sorting (ILS) and horizontal gene transfer (HGT)
Multiple Sequence Alignment (MSA): *another grand challenge*\(^1\)

\[
\begin{align*}
S_1 &= \text{AGGCTATCACCTGACCTCCA} & S_1 &= -\text{AGGCTATCACCTGACCTCCA} \\
S_2 &= \text{TAGCTATCAGGCCA} & S_2 &= \text{TAG-CTATCAGGCCA--} \\
S_3 &= \text{TAGCTGACCGC} & S_3 &= \text{TAG-CT--------GACCGC--} \\
& \vdots & & \vdots \\
S_n &= \text{TCACGACCGACA} & S_n &= \text{---------TCAC--GACCGACA}
\end{align*}
\]

*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
DNA Sequence Evolution

AAGACTT

-3 mil yrs

AAGGCCT

-2 mil yrs

AGGGCAT

-1 mil yrs

TAGCCCT

today

TGGACTT

AGCACTT

AGGGCAT

TAGCCCA

TAGACTT

AGCACAA

AGCGCTT
Phylogeny Problem

AGGGCAT  TAGCCCCA  TAGACTT  TGCACAA  TGCCTGCTT

U  V  W  X  Y

U  V  W

X  Y
Quantifying Error

TRUE TREE

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

INFERRED TREE
Statistical Consistency

Maximum likelihood is statistically consistent under standard models (e.g., GTR)
Mathematical 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 impact of model misspecification?
• What is the computational complexity of an estimation problem?
The Classical Phylogeny Problem

AGGGGCAT  TAGCCCCA  TAGACCTT  TGCACAA  TGCGBCTT
Much is known about this problem from a mathematical and empirical viewpoint
However…

AGGGGCATGA  AGAT  TAGACTT  TGCACAA  TGCAGCTT

U  V  W  X  Y

---

U  V  W

---

U  X  Y
Indels (insertions and deletions)
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 = TAGCTATCACGACCCGC
S3 = TAGCTGACCCGC
S4 = TCACGACCGACA
Phase 1: Alignment

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

S1 = -AGGCTATCACCTGACCTCCA
S2 = TAGCTATCAC--GACCGC--
S3 = TAGCTGACCGC--
S4 = ------TCAC--GACCGACAGACA
Phase 2: Construct tree

S1 = AGGCTATCACCTGACCTCCA  S1 = -AGGCTATCACCTGACCTCCA
S2 = TAGCTATCAGGACCCGC  S2 = TAGCTATCAGGACCCGC
S3 = TAGCTGACCCGC  S3 = TAGCTGACCCGC
S4 = TCAGGACCCGACA  S4 = TCAGGACCCGACA
Co-estimation would be much better!!!

S1 = AGGCTATCACCTGACCTCCA  S1 = -AGGCTATCACCTGACCTCCA
S2 = TAGCTATCAGACCGC       S2 = TAG-CTATCAC--GACCGC--
S3 = TAGCTGACCGC        S3 = TAG-CT--------GACCGC--
S4 = TCACGACCAGCA   S4 = --------TCAC--GACCGACGACA
Simulation Studies

\[ S_1 = \text{-AGGCTATCACCTGACCTCCA} \]
\[ S_2 = \text{TAGCTATCACGACCGC} \]
\[ S_3 = \text{TAGCTGACCGC} \]
\[ S_4 = \text{TCACGACCGACA} \]

Unaligned Sequences

True tree and alignment

Compare

Estimated tree and alignment

Unaligned Sequences

S1 = \text{-AGGCTATCACCTGACCTCCA}
S2 = \text{TAGCTATCACGACCGC--}
S3 = \text{TAGCTGACCGC--}
S4 = \text{T---C-A-CGACCGA----CA}
Quantifying Error

**FALSE NEGATIVE (FN)**: missing edge

**FALSE POSITIVE (FP)**: 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
Two-phase estimation

- Alignment methods
  - Clustal
  - POY (and POY*)
  - Probcons (and Probtree)
  - 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
1000-taxon models, ordered by difficulty (Liu et al., 2009)
SATé “Family” of methods

• Iterative divide-and-conquer methods
  – Each iteration re-aligns the sequences using the current tree, running preferred MSA methods on small local subsets, and merging subset alignments
  – Each iteration computes an ML tree on the current alignment, under the GTR (Generalized Time Reversible) Markov model of evolution

• Note: these methods are “MSA boosters”, designed to improve accuracy and/or scalability of the base method

• We show results using MAFFT-I-ins-i to align subsets
Re-aligning on a tree

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

Obtain initial alignment and estimated ML tree

Estimate ML tree on new alignment

Use tree to compute new alignment

Repeat until termination condition, and return the alignment/tree pair with the best ML score
1000-taxon models, ordered by difficulty – rate of evolution generally increases from left to right

SATé-1 24 hour analysis, on desktop machines

(Similar improvements for biological datasets)

SATé-1 can analyze up to about 8,000 sequences.
1000-taxon models ranked by difficulty

1000-taxon models ranked by difficulty

SATé-1: up to 8K
SATé-2: up to ~50K

SATé-1 and SATé-2 (Systematic Biology, 2012)
Tree Error – Simulated data

- 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)
PASTA Running Time and Scalability

![Graph showing running time vs. number of sequences]

- One iteration
- Using
  - 12 cpus
  - 1 node on Lonestar TACC
  - Maximum 24 GB memory
- Showing wall clock running time
  - ~ 1 hour for 10k taxa
  - ~ 17 hours for 200k taxa
1kp: Thousand Transcriptome Project

G. Ka-Shu Wong  
U Alberta

J. Leebens-Mack  
U Georgia

N. Wickett  
Northwestern

N. Matasci  
iPlant

T. Warnow,  
UIUC

S. Mirarab,  
UT-Austin

N. Nguyen,  
UT-Austin

Challenge:
Massive gene tree conflict consistent with ILS

Alignment of datasets with > 100,000 sequences

- Plant Tree of Life based on transcriptomes of ~1200 species
- More than 13,000 gene families (most not single copy)

Plus many many other people…
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.
1kp: Thousand Transcriptome Project

- Plant Tree of Life based on transcriptomes of ~1200 species
- More than 13,000 gene families (most not single copy)

Challenge:
Alignment of datasets with > 100,000 sequences with many fragmentary sequences
UPP

UPP = “Ultra-large multiple sequence alignment using Phylogeny-aware Profiles”


Purpose: highly accurate large-scale multiple sequence alignments, even in the presence of fragmentary sequences.
UPP

UPP = “Ultra-large multiple sequence alignment using Phylogeny-aware Profiles”


Purpose: highly accurate large-scale multiple sequence alignments, even in the presence of fragmentary sequences.

Uses an ensemble of HMMs
Simple idea (not UPP)

• Select random subset of sequences, and build “backbone alignment”

• Construct a Hidden Markov Model (HMM) on the backbone alignment

• Add all remaining sequences to the backbone alignment using the HMM
One Hidden Markov Model for the entire alignment?
Simple idea (not UPP)

• Select random subset of sequences, and build "backbone alignment"

• Construct a Hidden Markov Model (HMM) on the backbone alignment

• Add all remaining sequences to the backbone alignment using the HMM
This approach works well if the dataset is small or has low evolutionary rates, but can have reduced accuracy otherwise.

- Select random subset of sequences, and build “backbone alignment”
- Construct a Hidden Markov Model (HMM) on the backbone alignment
- Add all remaining sequences to the backbone alignment using the HMM
One Hidden Markov Model for the entire alignment?
Or 2 HMMs?
Or 4 HMMs?
Or all 7 HMMs?
1. Select random subset of full-length sequences, and build “backbone alignment”

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

3. Add all remaining sequences to the backbone alignment using the Ensemble of HMMs
Evaluation

• Simulated datasets (some have fragmentary sequences):
  – 10K to 1,000,000 sequences in RNASim – complex RNA sequence evolution simulation
  – 1000-sequence nucleotide datasets from SATé papers
  – 5000-sequence AA datasets (from FastTree paper)
  – 10,000-sequence Indelible nucleotide simulation

• Biological datasets:
  – Proteins: largest BaliBASE and HomFam
  – RNA: 3 CRW datasets up to 28,000 sequences
RNASeq Million Sequences: alignment error

Notes:
- We show alignment error using average of SP-FN and SP-FP.
- UPP variants have better alignment scores than PASTA.
- (Not shown: Total Column Scores – PASTA more accurate than UPP)
- No other methods tested could complete on these data
Using 12 processors:

- UPP(Fast,NoDecomp) took 2.2 days,
- UPP(Fast) took 11.9 days, and
- PASTA took 10.3 days
UPP vs. PASTA: impact of fragmentation

Under high rates of evolution, PASTA is badly impacted by fragmentary sequences (the same is true for other methods).

Under low rates of evolution, PASTA can still be highly accurate (data not shown).

UPP continues to have good accuracy even on datasets with many fragments under all rates of evolution.

Performance on fragmentary datasets of the 1000M2 model condition
Summary so far

- Standard MSA methods do not provide good accuracy on large datasets with high rates of evolution or with fragmentary sequences.
- Alignment error results in error in estimated phylogenies and associated questions (topologies, branch lengths, selection, etc.)
- PASTA generally better than UPP for tree error on datasets without fragmentary sequences, but UPP is more robust to fragmentary sequences.
- More generally, divide-and-conquer can improve the scalability of MSA methods, and maintain good accuracy to very large datasets.
**Ongoing Challenges and Opportunities**

- Statistical co-estimation of alignments and trees
- Phylogeny estimation given indels
- Exploring alignment uncertainty
- Estimating better alignments by combining estimated alignments

BUT: sequences evolve with duplications, rearrangements, and other events – not just substitutions and indels.

We need better statistical models of sequence evolution and scalable methods for estimating under these models.
Phylogeny + genomics = genome-scale phylogeny estimation
Gene tree discordance

Incomplete Lineage Sorting (ILS) is a dominant cause of gene tree heterogeneity
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, 14,000 loci
• Jarvis, Mirarab, et al., Science 2014

Major challenges:
• Maximum likelihood analysis of concatenated alignment took 250 CPU years using supercomputers around the globe
• Massive gene tree heterogeneity consistent with incomplete lineage sorting
1kp: Thousand Transcriptome Project

- Plant Tree of Life based on transcriptomes
- First publication: PNAS 2014 (about 100 species and 800 genes)
- Second analysis underway, much larger dataset (~1200 species and ~1000 loci)

Challenges:
- Massive gene tree conflict consistent with ILS
- Alignment of datasets with > 100,000 sequences

Plus many many other people…
Incomplete Lineage Sorting (ILS)

• Confounds phylogenetic analysis for many groups: Hominids, Birds, Yeast, Animals, Toads, Fish, Fungi, etc.

• There is substantial debate about how to analyze phylogenomic datasets in the presence of ILS, focused around statistical consistency guarantees (theory) and performance on data.
Part 2: Species tree estimation under the MSC

- Gene tree heterogeneity due to incomplete lineage sorting, modelled by the multi-species coalescent (MSC)
- Statistically consistent estimation of species trees under the MSC, and the impact of gene tree estimation error
- ASTRAL (Bioinformatics 2014, 2015): coalescent-based species tree estimation method that has high accuracy on large datasets (1000 species and genes)
- Statistical binning (Science 2015 and PLOS One 2015)
- Other techniques
Gorilla and Orangutan are not siblings in the species tree, but they are in the gene tree.
Main 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
Traditional approach: concatenation

- Statistically **inconsistent** and can even be positively misleading (proved for unpartitioned maximum likelihood) [Roch and Steel, Theo. Pop. Gen., 2014]
- **Mixed accuracy in simulations** [Kubatko and Degnan, Systematic Biology, 2007] [Mirarab, et al., Systematic Biology, 2014]
What about summary methods?
What about summary methods?

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

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 species tree?

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 (Aho et al.): The rooted tree on n species can be computed from its set of 3-taxon rooted subtrees in polynomial time.
1KP: Thousand Transcriptome Project

- 103 plant transcriptomes, 400-800 single copy “genes”
- Next phase will be much bigger
- Wickett, Mirarab et al., *PNAS* 2014

**Challenges:**
- Massive gene tree heterogeneity consistent with ILS
- At the time, MP-EST was the leading method.
- However, we could not use MP-EST due to missing data (many gene trees could not be rooted) and large number of species.
Species tree estimation from unrooted gene trees

Theorem (Allman et al.): Under the multi-species coalescent model, for any four taxa A, B, C, and D, the most probable unrooted gene tree on \{A,B,C,D\} is identical to the unrooted species tree induced on \{A,B,C,D\}.

Hence: Statistically consistent methods for estimating unrooted species trees from unrooted gene trees:
- BUCKy-pop (Larget et al., 2010) and
- ASTRAL (Mirarab et al., 2014 and Mirarab and Warnow 2015)
Minimum Quartet Distance

Input: Set of gene trees, \(t_1, t_2, \ldots, t_k\), each on leafset \(S\)

Output: Tree \(T\) on leafset \(S\) that minimizes

\[\Sigma_t d(t,T)\]

where \(d(t,T)\) is the quartet distance between the two trees, \(t\) and \(T\).

Notes:

• The computational complexity of this problem is open.

• An exact solution to this problem is a statistically consistent method for species tree estimation under the multi-species coalescent model.
**ASTRAL and ASTRAL-2**

- Exactly solves the minimum quartet distance problem subject to a constraint on the solution space, given by input set $X$ of bipartitions (find tree $T$ taking its bipartitions from a set $X$, that minimizes the quartet distance).
- **Theorem:** ASTRAL is statistically consistent under the MSC, even when solved in constrained mode (drawing bipartitions from the input gene trees)
- The constrained version of ASTRAL runs in polynomial time
- Open source software at [https://github.com/smirarab](https://github.com/smirarab)
- Published in Bioinformatics 2014 and 2015
- Used in 1KP analysis (Wickett, Mirarab et al., PNAS 2014) and many other studies
Simulation study

- Variable parameters:
  - Number of species: 10 – 1000
  - Number of genes: 50 – 1000
  - Amount of ILS: low, medium, high
  - Deep versus recent speciation

- 11 model conditions (50 replicas each) with heterogenous gene tree error

- Compare to NJst, MP-EST, concatenation (CA-ML)

- Evaluate accuracy using FN rate: the percentage of branches in the true tree that are missing from the estimated tree

Used SimPhy, Mallo and Posada, 2015
Tree accuracy when varying the number of species

1000 genes, “medium” levels of recent ILS
Tree accuracy when varying the number of species

1000 genes, “medium” levels of recent ILS
Tree accuracy when varying the number of species

1000 genes, “medium” levels of recent ILS
Running time when varying the number of species

1000 genes, “medium” levels of recent ILS
1KP: Thousand Transcriptome Project

- 103 plant transcriptomes, 400-800 single copy "genes"
- Next phase will be much bigger
- Wickett, Mirarab et al., *PNAS* 2014

Two trees presented: one based on concatenation, and the other based on ASTRAL.

The two trees were highly consistent.
Insights on biological data

• Main question: The placement of Amborella at the base of angiosperms

• Xi et al. (2014) used a collection of 310 genes sampled from 46 species.

• Conflicting results:
  • Concatenation puts Amborella at the base (H1)
  • MP-EST puts Amobrella + water lilies at the base (H2)
  • Xi et al. conclude ILS is the cause
  • ASTRAL like many other recent studies (e.g., 1KP) recovers H1
    • ILS is not necessarily the cause
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)
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, 14,000 loci
• Jarvis, Mirarab, et al., Science 2014

Major challenge:
• Massive gene tree heterogeneity consistent with incomplete lineage sorting
• Very poor resolution in the 14,000 gene trees (average bootstrap support 25%)
• Standard coalescent-based species tree estimation methods contradicted concatenation analysis and prior studies
Statistical Consistency for summary methods

Data are gene trees, presumed to be randomly sampled true gene trees.
Avian Phylogenomics Project

- Approx. 50 species, whole genomes, 14,000 loci
- Published Science 2014

**Most gene trees had very low bootstrap support, suggestive of gene tree estimation error**
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, 14,000 loci

Solution: Statistical Binning

• Improves coalescent-based species tree estimation by improving gene trees (Mirarab, Bayzid, Boussau, and Warnow, Science 2014)

• Avian species tree estimated using Statistical Binning with MP-EST (Jarvis, Mirarab, et al., Science 2014)
Ideas behind statistical binning

• “Gene tree” error tends to decrease with the number of sites in the alignment

  Number of sites in an alignment

• Concatenation (even if not statistically consistent) tends to be reasonably accurate when there is not too much gene tree heterogeneity
The statistical binning pipeline for estimating species trees from gene trees. Loci are grouped into bins based on a statistical test for combinability, before estimating gene trees.

Note: Supergene trees computed using fully partitioned maximum likelihood Vertex-coloring graph with balanced color classes is NP-hard; we used heuristic.
Sta=s=cal	binning	vs.	unbinned

Datasets: 11-taxon strongILS datasets with 50 genes from Chung and Ané, Systematic Biology
Binning produces bins with approximate 5 to 7 genes each

Average FN rate

<table>
<thead>
<tr>
<th>Method</th>
<th>MP-EST</th>
<th>MDC*(75)</th>
<th>MRP</th>
<th>MRL</th>
<th>GC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unbinned</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Statistical-75</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Statistical binning vs. unbinned
Theorem 3 (PLOS One, Bayzid et al. 2015): Unweighted statistical binning pipelines are not statistically consistent under GTR+MSC

As the number of sites per locus increase:
• All estimated gene trees converge to the true gene tree and have bootstrap support that converges to 1 (Steel 2014)
• For each bin, with probability converging to 1, the genes in the bin have the same tree topology (but can have different numeric parameters), and there is only one bin for any given tree topology
• For each bin, a fully partitioned maximum likelihood (ML) analysis of its supergene alignment converges to a tree with the common gene tree topology.

As the number of loci increase:
• every gene tree topology appears with probability converging to 1.

Hence as both the number of loci and number of sites per locus increase, with probability converging to 1, every gene tree topology appears exactly once in the set of supergene trees.

It is impossible to infer the species tree from the flat distribution of gene trees!
Fig 1. Pipeline for unbinned analyses, unweighted statistical binning, and weighted statistical binning.
Theorem 2 (PLOS One, Bayzid et al. 2015): WSB pipelines are statistically consistent under GTR+MSC

Easy proof:
As the number of sites per locus increase
• All estimated gene trees converge to the true gene tree and have bootstrap support that converges to 1 (Steel 2014)
• For every bin, with probability converging to 1, the genes in the bin have the same tree topology
• Fully partitioned GTR ML analysis of each bin converges to a tree with the common topology of the genes in the bin

Hence as the number of sites per locus and number of loci both increase, WSB followed by a statistically consistent summary method will converge in probability to the true species tree. Q.E.D.
Table 1. Model trees used in the Weighted Statistical Binning study. We show number of taxa, species tree branch length (relative to base model), and average topological discordance between true gene trees and true species tree.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Species tree branch length scaling</th>
<th>Average Discordance (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avian (48)</td>
<td>2X</td>
<td>35</td>
</tr>
<tr>
<td>Avian (48)</td>
<td>1X</td>
<td>47</td>
</tr>
<tr>
<td>Avian (48)</td>
<td>0.5X</td>
<td>59</td>
</tr>
<tr>
<td>Mammalian (37)</td>
<td>2X</td>
<td>18</td>
</tr>
<tr>
<td>Mammalian (37)</td>
<td>1X</td>
<td>32</td>
</tr>
<tr>
<td>Mammalian (37)</td>
<td>0.5X</td>
<td>54</td>
</tr>
<tr>
<td>10-taxon</td>
<td>“Lower ILS&quot;</td>
<td>40</td>
</tr>
<tr>
<td>10-taxon</td>
<td>“Higher ILS&quot;</td>
<td>84</td>
</tr>
<tr>
<td>15-taxon</td>
<td>“High ILS&quot;</td>
<td>82</td>
</tr>
</tbody>
</table>

doi:10.1371/journal.pone.0129183.t001
Weighted Statistical Binning: empirical

WSB generally benign to highly beneficial for moderate to large datasets:

- Improves gene tree estimation
- Improves species tree topology
- Improves species tree branch length
- Reduces incidence of highly supported false positive branches

• WSB can be hurtful on very small datasets with very high ILS levels.
Binning can improve species tree topology estimation

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

Comparing Binned and Un-binned MP-EST on the Avian Dataset

Binned MP-EST is largely consistent with the ML concatenation analysis.

The trees presented in Science 2014 were the ML concatenation and Binned MP-EST

Other techniques

• Disk-covering methods to improve coalescent-based species tree estimation methods (Bayzid, Hunt, and Warnow 2014)

• Co-estimation of gene trees and species trees:
  – *BEAST (Heled and Drummond): computationally intensive
  – BBCA (Zimmerman and Warnow): randomly divides genes into subsets, uses *BEAST on each subset to compute gene trees, then combines gene trees

• Single-site methods:
  – Supermatrix Rooted Triples (SMRT-ML) (DeGiorgio and Degnan 2010)
  – SNAPP (Bryant et al. 2012)
  – SVDquartets (Chifman and Kubatko 2014)
  – METAL (Dasarathy, Nowak, and Roch 2015)
Questions

• Impact of other sources of gene tree discord (e.g., gene flow, gene duplication and loss, horizontal gene transfer, introgression)

• Impact of missing data and limited number of sites per gene
What about performance on bounded number of sites?

• Question: Do any summary methods converge to the species tree as the number of loci increase, but where each locus has only a constant number of sites?

• Answers: Roch & Warnow, Syst Biol, March 2015:
  – Strict molecular clock: Yes for some new methods, even for a single site per locus
  – No clock: Unknown for all methods, including MP-EST, ASTRAL, etc.

S. Roch and T. Warnow. "On the robustness to gene tree estimation error (or lack thereof) of coalescent-based species tree methods", Systematic Biology, 64(4):663-676, 2015, [PDF]
Overall Observations

- Statistical methods – especially Bayesian MCMC methods – are typically computationally very intensive. But these are likely to be the most accurate.
- Relative accuracy on small to moderate datasets does not always extend to large datasets, or datasets with substantial heterogeneity or weak phylogenetic signal.
- Divide-and-conquer techniques have been useful for improving the accuracy and scalability of methods to challenging datasets.
Scientific challenges:

- Ultra-large multiple-sequence alignment
- Gene tree estimation
- Metagenomic classification
- Alignment-free phylogeny estimation
- Supertree estimation
- Estimating species trees from many gene trees
- Genome rearrangement phylogeny
- Reticulate evolution
- Visualization of large trees and alignments
- Data mining techniques to explore multiple optima
- Theoretical guarantees under Markov models of evolution

Techniques: applied probability theory, graph theory, supercomputing, and heuristics

Testing: simulations and real data
Acknowledgments

Papers available at [http://tandy.cs.illinois.edu/papers.html](http://tandy.cs.illinois.edu/papers.html)
PASTA, UPP, ASTRAL and statistical binning software at [https://github.com/smirarab](https://github.com/smirarab)

**Funding:** NSF, Guggenheim Foundation, David Bruton Jr. Centennial Professorship, TACC (Texas Advanced Computing Center), BlueWaters, Grainger Foundation, and HHMI (to SM)
• “Nothing in biology makes sense except in the light of evolution”

• “...... nothing in evolution makes sense except in the light of phylogeny ...”