Constrained Exact Optimization in Phylogenetics

Tandy Warnow
The University of Illinois at Urbana-Champaign
Phylogeny
(evolutionary tree)

From the Tree of Life Website,
University of Arizona
DNA Sequence Evolution

AAGACTT

TG

AAG

C

C

T

-3 mil yrs

-2 mil yrs

-1 mil yrs

today

AAGGCCT

AAGGCAT

AGGGCAT

AGGGCAT

TAGCCCA

TAGACTT

AGCACA

AGCGCTT

TAGCCCT

AGGCCT

TGGACTT

AAGACTT

AGCACTT

AGGA
 Phylogeny Problem

AGTGCAT  TAGCCCCA  TAGACTTT  TGCACAA  TGCCTTT
Phylogenomics

Phylogeny + genomics = genome-scale phylogeny estimation.
Main competing approaches

<table>
<thead>
<tr>
<th>Species</th>
<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>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Concatenation
- Analyze separately
- Summary Method
Phylogenomic pipeline

- Select taxon set and markers
- Gather and screen sequence data, possibly identify orthologs
- Compute multiple sequence alignment and “gene tree” for each locus
- Compute species tree or phylogenetic network from the gene trees or alignments
- Get statistical support on each branch (e.g., bootstrapping)
- Estimate dates on the nodes of the phylogeny
- Use species tree with branch support and dates to understand biology
Just about everything worth doing is NP-hard

- Multiple sequence alignment
- Maximum likelihood gene tree estimation
- Species tree estimation by combining gene trees
- Supertree estimation
  - and hence divide-and-conquer strategies
Just about everything worth doing is NP-hard

- Multiple sequence alignment
- Maximum likelihood gene tree estimation
- Species tree estimation by combining gene trees
- Supertree estimation
  - and hence divide-and-conquer strategies

And Bayesian methods are even more computationally intensive
Local search heuristics

Standard heuristic search:
TBR moves through “treespace”
Randomness to exit local optima
But there are \((2n-5)!!\) trees on \(n\) leaves
Solving maximum likelihood (and other hard optimization problems) is… unlikely

<table>
<thead>
<tr>
<th># of Taxa</th>
<th># of Unrooted Trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>6</td>
<td>105</td>
</tr>
<tr>
<td>7</td>
<td>945</td>
</tr>
<tr>
<td>8</td>
<td>10395</td>
</tr>
<tr>
<td>9</td>
<td>135135</td>
</tr>
<tr>
<td>10</td>
<td>2027025</td>
</tr>
<tr>
<td>20</td>
<td>$2.2 \times 10^{20}$</td>
</tr>
<tr>
<td>100</td>
<td>$4.5 \times 10^{190}$</td>
</tr>
<tr>
<td>1000</td>
<td>$2.7 \times 10^{2900}$</td>
</tr>
</tbody>
</table>
Avian Phylogenomics Project

- Approx. 50 species, whole genomes
- 14,000 loci
- Multi-national team (100+ investigators)
- 8 papers published in special issue of Science 2014
Only 48 species, but heuristic ML took ~300 CPU years on multiple supercomputers and used 1Tb of memory

Jarvis, Mirarab, et al., examined 48 bird species using 14,000 loci from whole genomes. Two trees were presented.

1. A single dataset maximum likelihood concatenation analysis used ~300 CPU years and 1Tb of distributed memory, using TACC and other supercomputers around the world.

2. However, every locus had a different tree — suggestive of “incomplete lineage sorting” — and the noisy genome-scale data required the development of a new method, “statistical binning”.

RESEARCH ARTICLE

Whole-genome analyses resolve early branches in the tree of life of modern birds


To better determine the history of modern birds, we performed a genome-scale phylogenetic analysis of 48 species representing all orders of Neoaves using phylogenomic methods created to handle genome-scale data. We recovered a highly resolved tree that confirms
1KP: Thousand Transcriptome Project

- 103 plant transcriptomes, 400-800 single copy “genes”
- Wickett, Mirarab et al., *PNAS* 2014
- Next phase will be much bigger
  - ~1000 species and ~1000 genes,
  - and will require the inference of multiple sequence alignments and trees on more than 100,000 sequences
Species sequenced by year

- Viruses
- Eukaryotes
- Prokaryotes

Muir, 2016
This Talk

• Model-based tree estimation and NP-hard problems
• How divide-and-conquer can improve tree estimation
• Constrained optimization
• Open problems
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.

The different sites are assumed to evolve 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.
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), and how well do the methods perform in practice?
Statistical Consistency

error

Data
Answers?

• We know a lot about which site evolution models are identifiable, and which methods are statistically consistent.
• We know a little bit about the sequence length requirements for standard methods.
• Extensive studies show that even the best methods produce gene trees with some error.
Statistically consistent phylogenetic reconstruction methods

1. Hill-climbing heuristics for NP-hard optimization criteria (e.g., Maximum Likelihood)

![Graph showing cost vs. phylogenetic trees with local and global optimum]

2. Polynomial time distance-based methods: Neighbor Joining, FastME, etc.

3. Bayesian methods
Distance-based estimation

TRUE TREE

S1 S2 S3 S4

DNA SEQUENCES

S1 ACAATTAGAAC
S2 ACCCTTAGAAC
S3 ACCATTCCAAC
S4 ACCAGACCAAC

METHODS SUCH AS NEIGHBOR JOINING

INFERRED TREE

S1 2
S2 1
S3 1
S4 2

DISTANCE MATRIX

<table>
<thead>
<tr>
<th></th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>0</td>
<td>3</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>S2</td>
<td>0</td>
<td>5</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>S3</td>
<td>0</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S4</td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>
Quantifying Error

TRUE TREE

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

50% error rate

DNA SEQUENCES

S₁  ACAATTAGAAC
S₂  ACCCTTAGAAC
S₃  ACCATTCCAAAC
S₄  ACCAGACCCAAC
S₅  ACCAGACCGGA

INFERRED TREE
Neighbor Joining (NJ) on large 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]
In other words...

Statistical consistency doesn’t guarantee accuracy w.h.p. unless the sequences are long enough.
“Boosting” phylogeny reconstruction methods

- DCMs “boost” the performance of phylogeny reconstruction methods.
Divide-and-conquer for phylogeny estimation
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.

The different sites are assumed to evolve 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.
Sequence length requirements

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

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

We fix everything but $n$. 
Statistical consistency, exponential convergence, and absolute fast convergence (afc)
Distance-based estimation

TRUE TREE

S1  S2  S3  S4

INFERRED TREE

S1  2
S2  1  3
S3  1
S4  2

METHODS SUCH AS NEIGHBOR JOINING

DNA SEQUENCES

S1  ACAATTAGAAC
S2  ACCCTTAGAAC
S3  ACCATTCCAAC
S4  ACCAGACCAAC

DISTANCE MATRIX

<table>
<thead>
<tr>
<th></th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>0</td>
<td>3</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>S2</td>
<td></td>
<td>0</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>S3</td>
<td></td>
<td></td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>S4</td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>

STATISTICAL ESTIMATION OF PAIRWISE DISTANCES
Neighbor Joining’s sequence length requirement is exponential!

- Atteson 1999: Let $T$ be a Jukes-Cantor model tree defining additive matrix $D$. Then Neighbor Joining will reconstruct the true tree with high probability from sequences that are of length $O(\lg n e^{\max D_{ij}})$.

- Lacey and Chang 2009: Matching lower bound
Divide-and-conquer for phylogeny estimation

1. Construct subset trees
2. Supertree Step
3. Refinement Step
DCM1 Decompositions

**Input:** Set $S$ of sequences, distance matrix $d$, threshold value $q \subseteq \{d_{ij}\}$

1. Compute threshold graph
   
   $$G_q = (V, E), V = S, E = \{(i, j) : d(i, j) \leq q\}$$

2. Perform minimum weight triangulation (note: if $d$ is an additive matrix, then the threshold graph is provably triangulated).

DCM1 decomposition:

Compute maximal cliques
DCM1-boosting:
Warnow, St. John, and Moret,
SODA 2001

- The DCM1 phase produces a collection of trees (one for each threshold), and the SQS phase picks the “best” tree.
- For a given threshold, the base method is used to construct trees on small subsets (defined by the threshold) of the taxa. These small trees are then combined into a tree on the full set of taxa.
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
DCM1-boosting

• DCM1-boosting: reducing sequence length requirements for gene tree accuracy from exponential to polynomial
DCM1-boosting

• DCM1-boosting: reducing sequence length requirements for gene tree accuracy from exponential to polynomial

• Key algorithmic ingredients:
  – Construct triangulated graph
  – Apply tree construction methods to subsets
  – Combine subset trees together using supertree method
  – Select best tree from a set of trees
Other applications of divide-and-conquer

• DACTAL-boosting:
  – almost alignment-free tree estimation
  – Improving multi-locus species tree estimation

• DCM2-boosting: improving heuristic searches for maximum likelihood

• Superfine-boosting: improving supertree methods
Divide-and-conquer for phylogeny estimation

Construct subset trees

Supertree Step

Refinement Step
Supertree Problems

- Quartet Median Supertree
- Robinson-Foulds Supertree
- Matrix Representation with Parsimony
- Matrix Representation with Likelihood
- Etc.

All are NP-hard because even testing compatibility of unrooted trees is NP-complete.
Supertree Problems

• **Quartet Median Supertree**
• Robinson-Foulds Supertree
• Matrix Representation with Parsimony
• Matrix Representation with Likelihood
• Etc.

All are NP-hard because even testing compatibility of unrooted trees is NP-complete.
I’ll use the term “gene” to refer to “c-genes”: recombination-free orthologous stretches of the genome
Gene tree discordance

Incomplete Lineage Sorting (ILS) is a dominant cause of gene tree heterogeneity
Gorilla and Orangutan are not siblings in the species tree, but they are in the gene tree.
Incomplete Lineage Sorting (ILS)

• Confounds phylogenetetic 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.
Anomaly zone

- An anomalous gene tree (AGT) is one that is more probable than the true species tree under the multi-species coalescent model.

- Theorem (Hudson 1983): There are no rooted 3-leaf AGTs.

- Theorem (Allman et al. 2011, Degnan 2013): There are no unrooted 4-leaf AGTs.

- Theorem (Degnan 2013, Rosenberg 2013): For n > 3, there are model species trees with rooted AGTs, and for n > 4 there are model species trees with unrooted AGTs.
Anomaly zone

- An anomalous gene tree (AGT) is one that is more probable than the true species tree under the multi-species coalescent model.
- Theorem (Hudson 1983): There are no rooted 3-leaf AGTs.
Anomaly zone

• An anomalous gene tree (AGT) is one that is more probable than the true species tree under the multi-species coalescent model.

• Theorem (Hudson 1983): There are no rooted 3-leaf AGTs.

• Theorem (Allman et al. 2011, Degnan 2013): There are no unrooted 4-leaf AGTs.
Anomaly zone

• An anomalous gene tree (AGT) is one that is more probable than the true species tree under the multi-species coalescent model.

• Theorem (Hudson 1983): There are no rooted 3-leaf AGTs.

• Theorem (Allman et al. 2011, Degnan 2013): There are no unrooted 4-leaf AGTs.

• Theorem (Degnan 2013, Rosenberg 2013): For $n>3$, there are model species trees with rooted AGTs, and for $n>4$ there are model species trees with unrooted AGTs.
Anomaly zone

• An anomalous gene tree (AGT) is one that is more probable than the true species tree under the multi-species coalescent model.

• Theorem (Hudson 1983): There are no rooted 3-leaf AGTs.

• Theorem (Allman et al. 2011, Degnan 2013): There are no unrooted 4-leaf AGTs.

• Theorem (Degnan 2013, Rosenberg 2013): For n>3, there are model species trees with rooted AGTs, and for n>4 there are model species trees with unrooted AGTs.
• Optimization Problem (NP-Hard):

Find the species tree with the maximum number of induced quartet trees shared with the collection of input gene trees

\[ \text{Score}(T) = \sum_{t \in \mathcal{T}} |Q(T) \cap Q(t)| \]

• Theorem: Statistically consistent under the multispecies coalescent model when solved exactly
Constrained Maximum Quartet Support Tree

- **Input:** Set \( T = \{t_1, t_2, ..., t_k\} \) of unrooted gene trees, with each tree on set S with n species, and **set X of allowed bipartitions**
- **Output:** Unrooted tree \( T \) on leafset S, maximizing the total quartet tree similarity to \( T \), **subject to T drawing its bipartitions from X.**

- **Theorems:**
  - Mirarab et al. 2014: If X contains the bipartitions from the input gene trees (and perhaps others), then an exact solution to this problem is statistically consistent under the MSC.
  - Mirarab and Warnow 2015: The constrained MQST problem can be solved in \( O(|X|^2 nk) \) time. (We use dynamic programming, and build the unrooted tree from the bottom-up, based on "allowed clades" – halves of the allowed bipartitions.)
Constrained Maximum Quartet Support Tree

- Input: Set $\mathcal{T} = \{t_1, t_2, \ldots, t_k\}$ of unrooted gene trees, with each tree on set $S$ with $n$ species, and set $X$ of allowed bipartitions.
- Output: Unrooted tree $T$ on leafset $S$, maximizing the total quartet tree similarity to $\mathcal{T}$, subject to $T$ drawing its bipartitions from $X$.

Theorems:
- Mirarab et al. 2014: If $X$ contains the bipartitions from the input gene trees (and perhaps others), then an exact solution to this problem is statistically consistent under the MSC.
Constrained Maximum Quartet Support Tree

- **Input:** Set $\mathcal{T} = \{t_1, t_2, ..., t_k\}$ of unrooted gene trees, with each tree on set $S$ with $n$ species, and set $X$ of allowed bipartitions
- **Output:** Unrooted tree $T$ on leafset $S$, maximizing the total quartet tree similarity to $\mathcal{T}$, subject to $T$ drawing its bipartitions from $X$.

**Theorems:**
- Mirarab et al. 2014: If $X$ contains the bipartitions from the input gene trees (and perhaps others), then an exact solution to this problem is statistically consistent under the MSC.
- Mirarab and Warnow 2015: The constrained MQST problem can be solved in $O(|X|^2nk)$ time. (We use dynamic programming, and build the unrooted tree from the bottom-up, based on “allowed clades” – halves of the allowed bipartitions.)
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
Running time as function of # species

1000 genes, "medium" levels of ILS, simulated species trees
[Mirarab and Warnow, ISMB, 2015]
ASTRAL on biological datasets

- **1KP**: 103 plant species, 400-800 genes
- Yang, et al. 96 Caryophyllales species, 1122 genes
- Dentinger, et al. 39 mushroom species, 208 genes
- Giarla and Esselstyn. 19 Philippine shrew species, 1112 genes
- Laumer, et al. 40 flatworm species, 516 genes
- Grover, et al. 8 cotton species, 52 genes
- Hosner, Braun, and Kimball. 28 quail species, 11 genes
- Simmons and Gatesy. 47 angiosperm species, 310 genes
- Prum et al, 198 avian species, 259 genes
Other polynomial time algorithms for constrained optimization problems

• Minimize Duplication/Loss Supertree (Hallett and Lagergren 2000)
• Quartet Support (Bryant and Steel 2001)
• Gene tree estimation under Duplication and Loss (Szöllősi et al. 2013)
• Constrained Robinson-Foulds Supertree (Vachaspati and Warnow 2016)
Summary

• NP-hard optimization problems abound in phylogeny reconstruction, and in computational biology in general, and need very accurate solutions.
Summary

• NP-hard optimization problems abound in phylogeny reconstruction, and in computational biology in general, and need very accurate solutions.

• Divide-and-conquer techniques can provide greatly improved accuracy and scalability, and excellent statistical performance guarantees. But divide-and-conquer methods depend on having good supertree methods.
Summary

• NP-hard optimization problems abound in phylogeny reconstruction, and in computational biology in general, and need very accurate solutions.

• Divide-and-conquer techniques can provide greatly improved accuracy and scalability, and excellent statistical performance guarantees. But divide-and-conquer methods depend on having good supertree methods.

• Constrained exact optimization has been surprisingly beneficial for phylogenomic analysis.
Open Problems

• What other optimization problems are tractable, given constraint sets?
• How should we define the set $X$ of constraints from a set of source trees, especially when the source trees are incomplete?
• Are there other ways of constraining the search space that are effective and useful?
• What are other effective ways of approaching truly large-scale phylogeny estimation?
• Can divide-and-conquer be employed with Bayesian methods? (Or, more generally, how can we make Bayesian methods more scalable?)
Acknowledgments

NSF grant DBI-1461364 (joint with Noah Rosenberg at Stanford and Luay Nakhleh at Rice):
http://tandy.cs.illinois.edu/PhylogenomicsProject.html

Papers available at http://tandy.cs.illinois.edu/papers.html

**ASTRAL**: Available at https://github.com/smirarab
**FastRFS**: Available at https://github.com/pranjalv123/FastRFS
Constrained Robinson-Foulds Supertree

• Input: set $\mathcal{T}$ of source trees, and set $X$ of allowed bipartitions

• Output: tree $T$ that minimizes the total Robinson-Foulds distance to the input source trees, and that draws its bipartitions from $X$.

Theorem (Vachaspati and Warnow 2016): The constrained RF Supertree can be solved in $O(|X|^2 nk)$ time, where there are $k$ source trees and $n$ species.
Avian Phylogenomics Project

- Approx. 50 species, whole genomes
- 14,000 loci
- Multi-national team (100+ investigators)
- 8 papers published in special issue of Science 2014
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
1KP: Thousand Transcriptome Project

- 103 plant transcriptomes, 400-800 single copy “genes”
- Wickett, Mirarab et al., PNAS 2014
- Next phase will be much bigger
  - ~1000 species and ~1000 genes,
  - and will require the inference of multiple sequence alignments and trees on more than 100,000 sequences

And many others
Constrained optimization

• First proposed in Hallet and Lagergren 2000, for the duplication-loss species tree problem

• Algorithms for constrained optimization use dynamic programming to find optimal solutions, constructing the best tree from the “bottom-up”

• The set X is usually defined to be the bipartitions from the source trees.
New supertree method: SuperFine

Existing Method: RAxML(MAFFT)

Unaligned Sequences

pRecDCM3

Overlapping subsets

A tree for the entire dataset

A tree for each subset

BLAST-based
Results on three biological datasets – 6000 to 28,000 sequences. We show results with 5 DACTAL iterations.
DACTAL

Construct supertree

Construct subset trees

Decompose using tree

Overlapping subsets

A tree for each subset

A tree for the entire dataset

Arbitrary input
Triangulated Graphs

- Definition: A graph is triangulated if it has no simple cycles of size four or more.
Sampling multiple genes from multiple species

From the Tree of the Life Website,
University of Arizona
Standard heuristic search

- Random perturbation
- Hill-climbing
Problems with current techniques for MP

Shown here is the performance of the TNT software for maximum parsimony on a real dataset of almost 14,000 sequences. The required level of accuracy with respect to MP score is no more than 0.01% error (otherwise high topological error results). (“Optimal” here means best score to date, using any method for any amount of time.)
Solving NP-hard problems exactly is ... unlikely

- Number of (unrooted) binary trees on $n$ leaves is $(2n-5)!!$
- If each tree on 1000 taxa could be analyzed in 0.001 seconds, we would find the best tree in 2890 millennia

<table>
<thead>
<tr>
<th>#leaves</th>
<th>#trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>6</td>
<td>105</td>
</tr>
<tr>
<td>7</td>
<td>945</td>
</tr>
<tr>
<td>8</td>
<td>10395</td>
</tr>
<tr>
<td>9</td>
<td>135135</td>
</tr>
<tr>
<td>10</td>
<td>2027025</td>
</tr>
<tr>
<td>20</td>
<td>$2.2 \times 10^{20}$</td>
</tr>
<tr>
<td>100</td>
<td>$4.5 \times 10^{190}$</td>
</tr>
<tr>
<td>1000</td>
<td>$2.7 \times 10^{2900}$</td>
</tr>
</tbody>
</table>
Gene tree discordance
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

Challenges:
• Concatenation analysis used >200 CPU years

But also
• Massive gene tree conflict suggestive of ILS, and most gene trees had very low bootstrap support
• Coalescent-based analysis using MP-EST produced tree that conflicted with concatenation analysis
Big datasets and hard problems

• Computationally intensive problems:
  – Multiple sequence alignment
  – Maximum likelihood gene tree estimation
  – Species tree estimation from multiple gene trees

• Fast methods are generally not sufficiently accurate

• Accurate methods are generally computationally intensive
Quantifying Error

**TRUE TREE**

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

50% error rate

**DNA SEQUENCES**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_1$</td>
<td>ACAATTAGAAC</td>
</tr>
<tr>
<td>$S_2$</td>
<td>ACCCTTAGAAC</td>
</tr>
<tr>
<td>$S_3$</td>
<td>ACCATTCCAAC</td>
</tr>
<tr>
<td>$S_4$</td>
<td>ACCAGACCAAC</td>
</tr>
<tr>
<td>$S_5$</td>
<td>ACCAGACCGGA</td>
</tr>
</tbody>
</table>

**INFERRED TREE**
Neighbor joining has poor performance on large diameter trees [Nakhleh et al. ISMB 2001]

Theorem (Atteson): Exponential sequence length requirement for Neighbor Joining!
Species Tree Estimation requires multiple genes!
Species tree estimation: difficult, even for small datasets!

From the Tree of the Life Website, University of Arizona
Major Challenges:
large datasets, fragmentary sequences

- **Multiple sequence alignment**: Few methods can run on large datasets, and alignment accuracy is generally poor for large datasets with high rates of evolution.

- **Gene Tree Estimation**: standard methods have *poor accuracy* on even moderately large datasets, and the most accurate methods are enormously *computationally intensive* (weeks or months, high memory requirements).

- **Species Tree Estimation**: gene tree incongruence makes accurate estimation of species tree challenging.

Both phylogenetic estimation and multiple sequence alignment are also impacted by *fragmentary data*. 
Major Challenges:
large datasets, fragmentary sequences

- **Multiple sequence alignment**: Few methods can run on large datasets, and alignment accuracy is generally poor for large datasets with high rates of evolution.

- **Gene Tree Estimation**: standard methods have *poor accuracy* on even moderately large datasets, and the most accurate methods are enormously *computationally intensive* (weeks or months, high memory requirements).

- **Species Tree Estimation**: gene tree incongruence makes accurate estimation of species tree challenging.

- **Phylogenetic Network Estimation**: Horizontal gene transfer and hybridization requires non-tree models of evolution

Both phylogenetic estimation and multiple sequence alignment are also impacted by *fragmentary data.*
Strict Consensus Merger (SCM)
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
The Tree of Life: *Multiple Challenges*

Scientific challenges:

- Ultra-large multiple-sequence alignment
- 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

Applications:

- metagenomics
- protein structure and function prediction
- trait evolution
- detection of co-evolution
- systems biology

Techniques:

- Graph theory (especially chordal graphs)
- Probability theory and statistics
- Hidden Markov models
- Combinatorial optimization
- Heuristics
- Supercomputing
1kp: Thousand Transcriptome Project

- Plant Tree of Life based on transcriptomes of ~1200 species
- More than 13,000 gene families (most not single copy)
- First paper: PNAS 2014 (~100 species and ~800 loci)

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 – Mirarab et al. 2014, Mirarab & Warnow 2015)
- Multiple sequence alignment of >100,000 sequences (with lots of fragments!) – we will use UPP (Nguyen et al., 2015)
phylogenomics

I’ll use the term “gene” to refer to “c-genes”: recombination-free orthologous stretches of the genome