Machine Learning and Discrete Algorithms for Reconstructing the Tree of Life

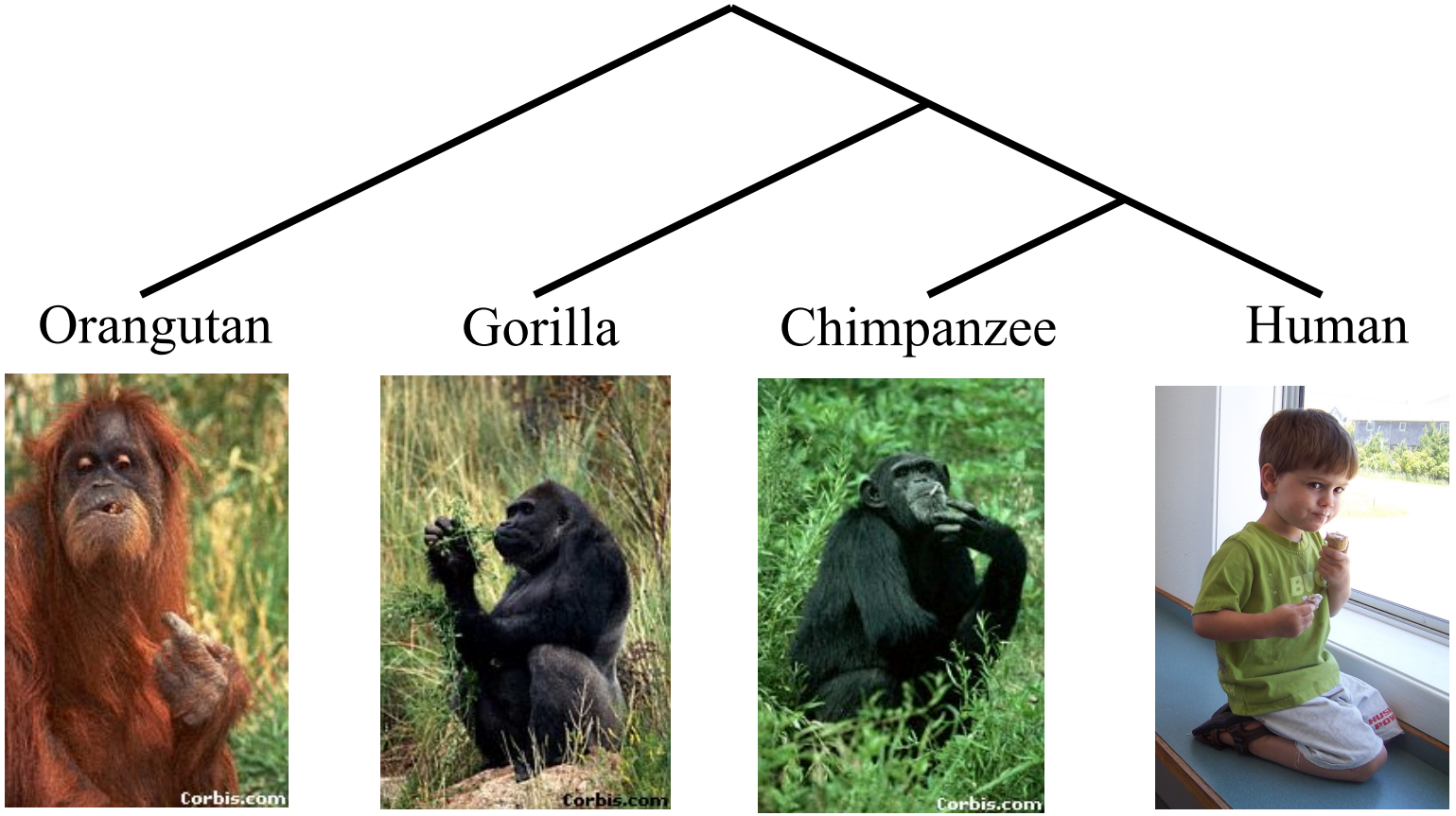
Tandy Warnow

The University of Illinois

My research (overview)

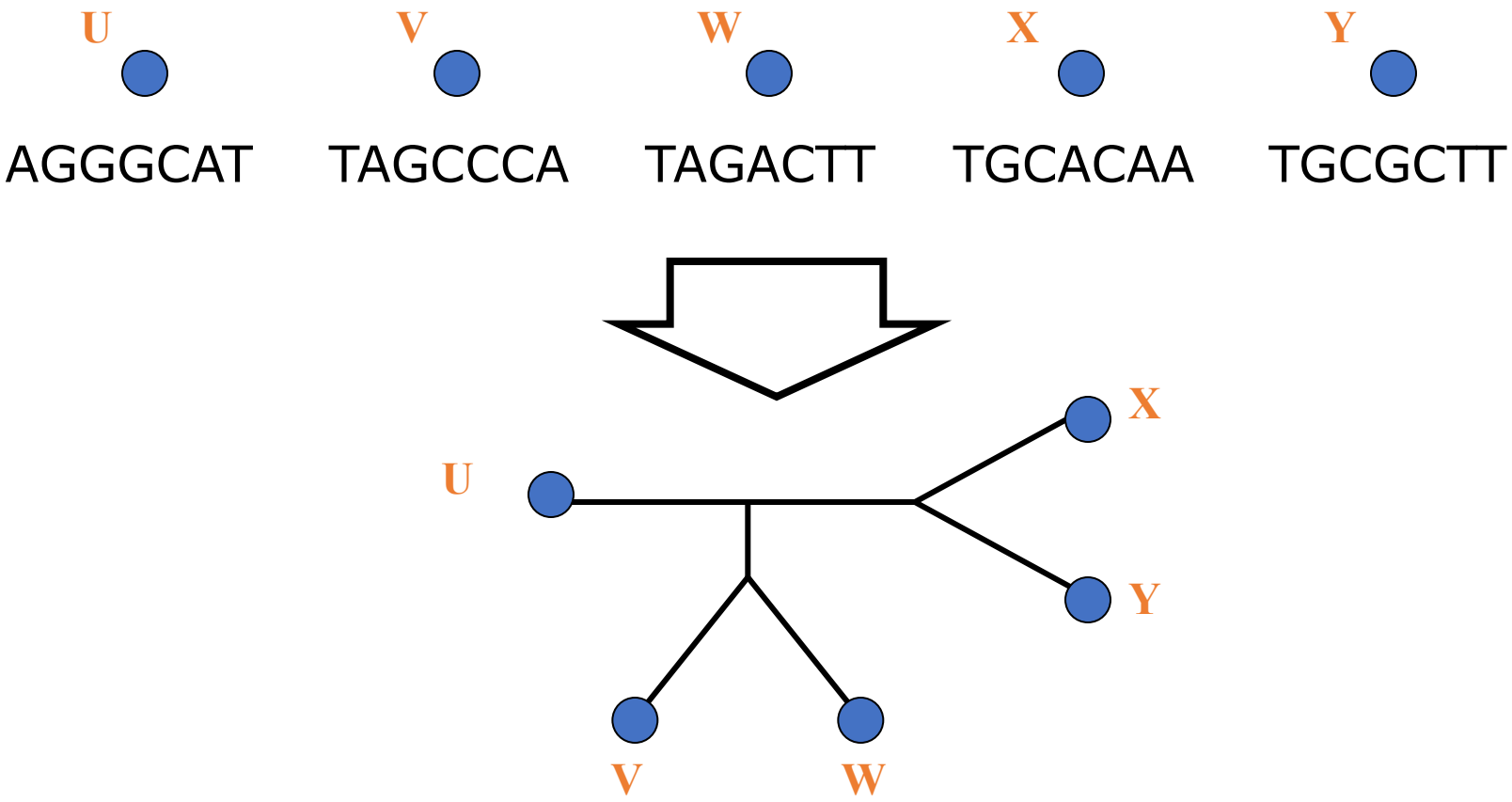
- Machine learning in bioinformatics (e.g., ensembles of profile Hidden Markov Models)
- Novel heuristics for NP-hard optimization problems
- Discrete and graph-theoretic algorithms for phylogeny estimation
- Collaborations with biologists and historical linguists

Phylogeny (evolutionary tree)



From the Tree of the Life Website,
University of Arizona

Phylogeny Problem



Phylogeny estimation as a statistical problem

- Assume DNA sequences are generated on an unknown model tree, and try to infer the tree from the observed sequences seen at the leaves

NP-hard optimization problems

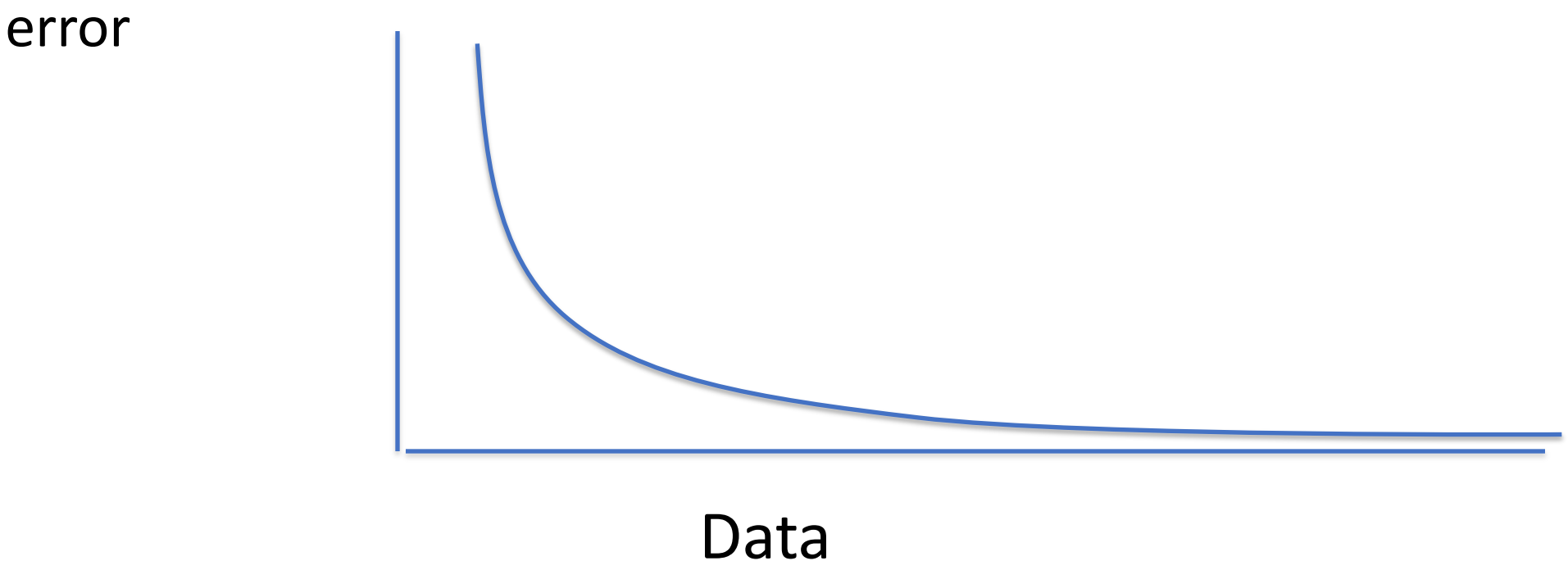
Large datasets

Years of CPU time for standard methods

This research combines many types of computer science:

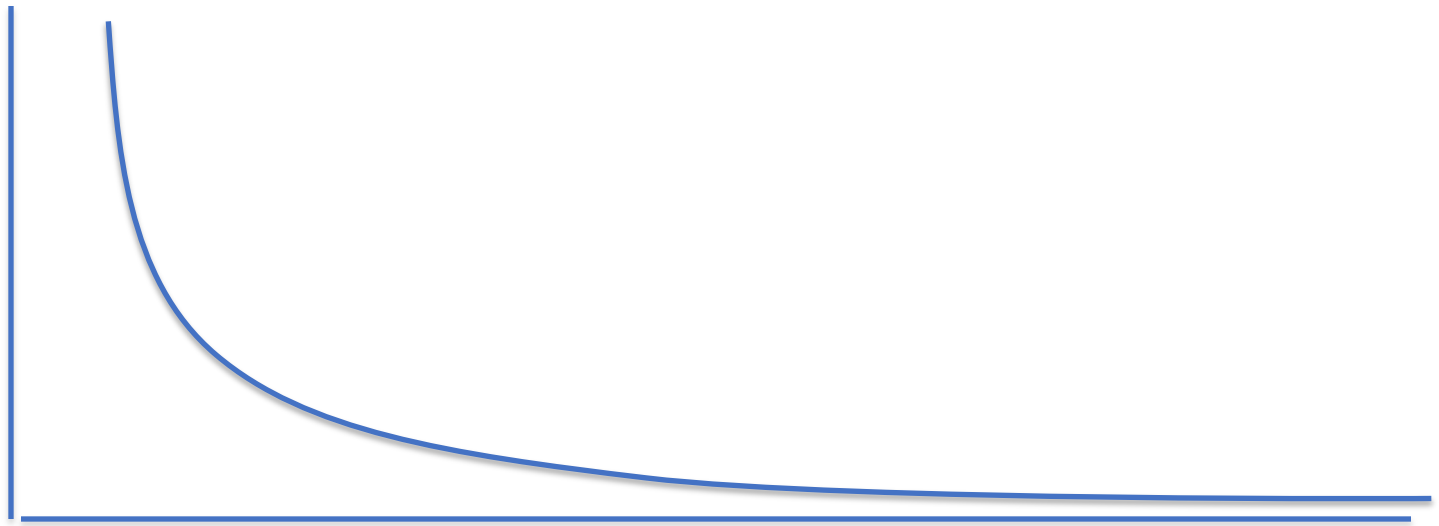
Algorithm design, **proofs**, implementation, simulations and testing

Statistical Consistency/Identifiability



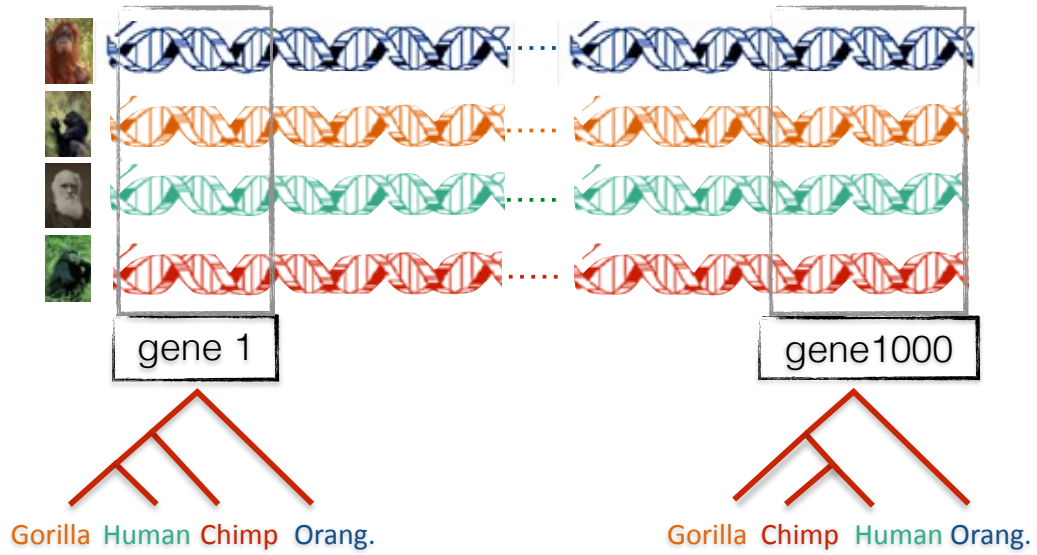
Genome-scale data?

error



Length of the genome

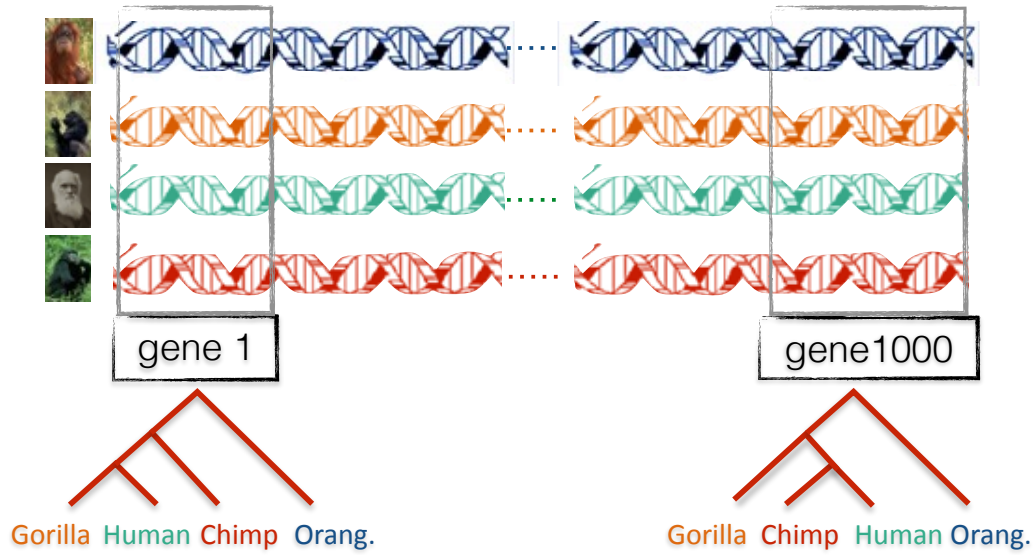
Gene tree discordance



Multiple causes for discord, including

- Incomplete Lineage Sorting (ILS),
- Gene Duplication and Loss (GDL), and
- Horizontal Gene Transfer (HGT)

Gene tree discordance



Multiple causes for discord, including

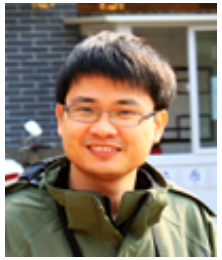
- Incomplete Lineage Sorting (ILS),
- Gene Duplication and Loss (GDL), and
- Horizontal Gene Transfer (HGT)

Avian Phylogenomics Project

Erich Jarvis, HHMI

MTP Gilbert, Copenhagen

Guojie Zhang, BGI

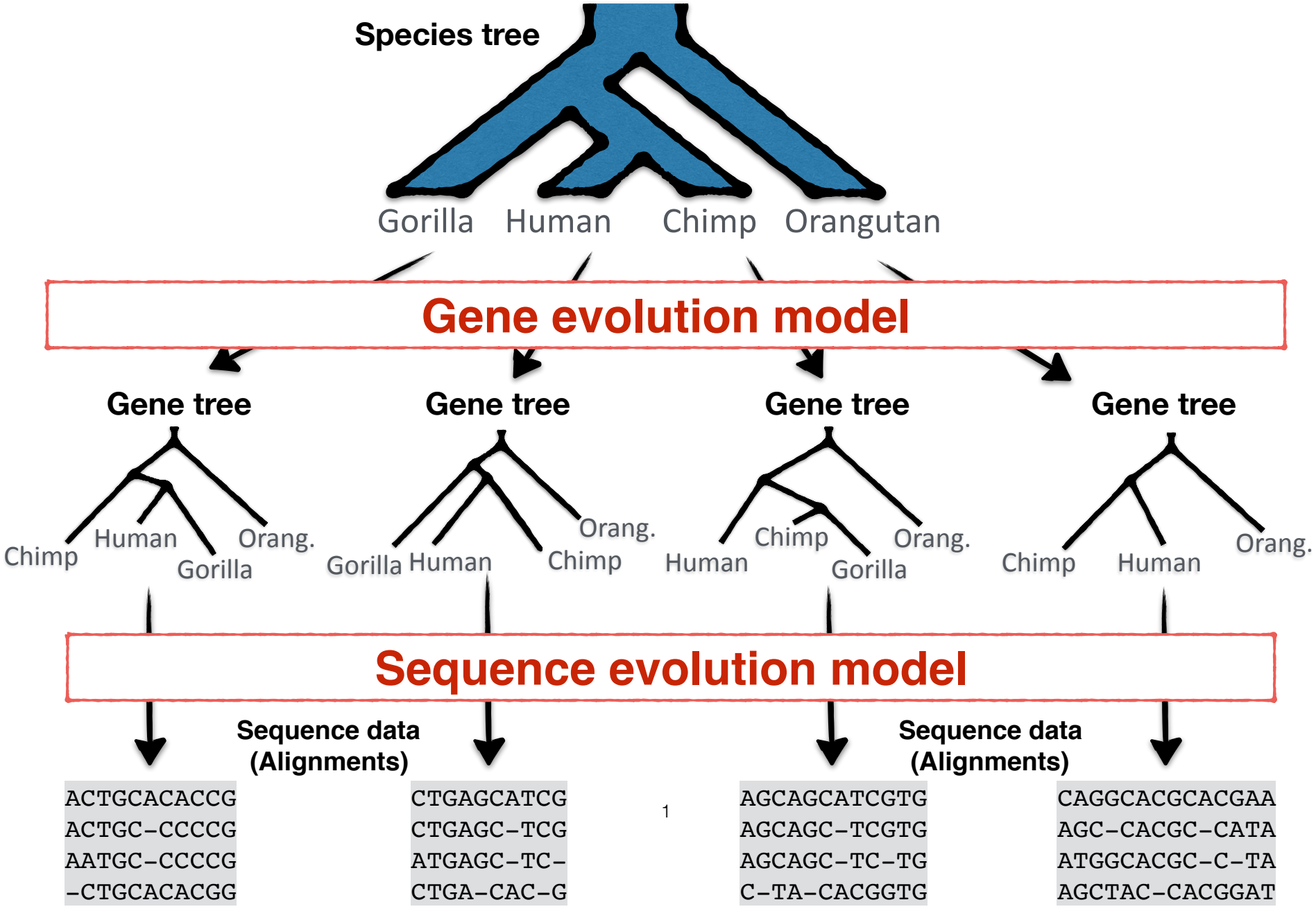


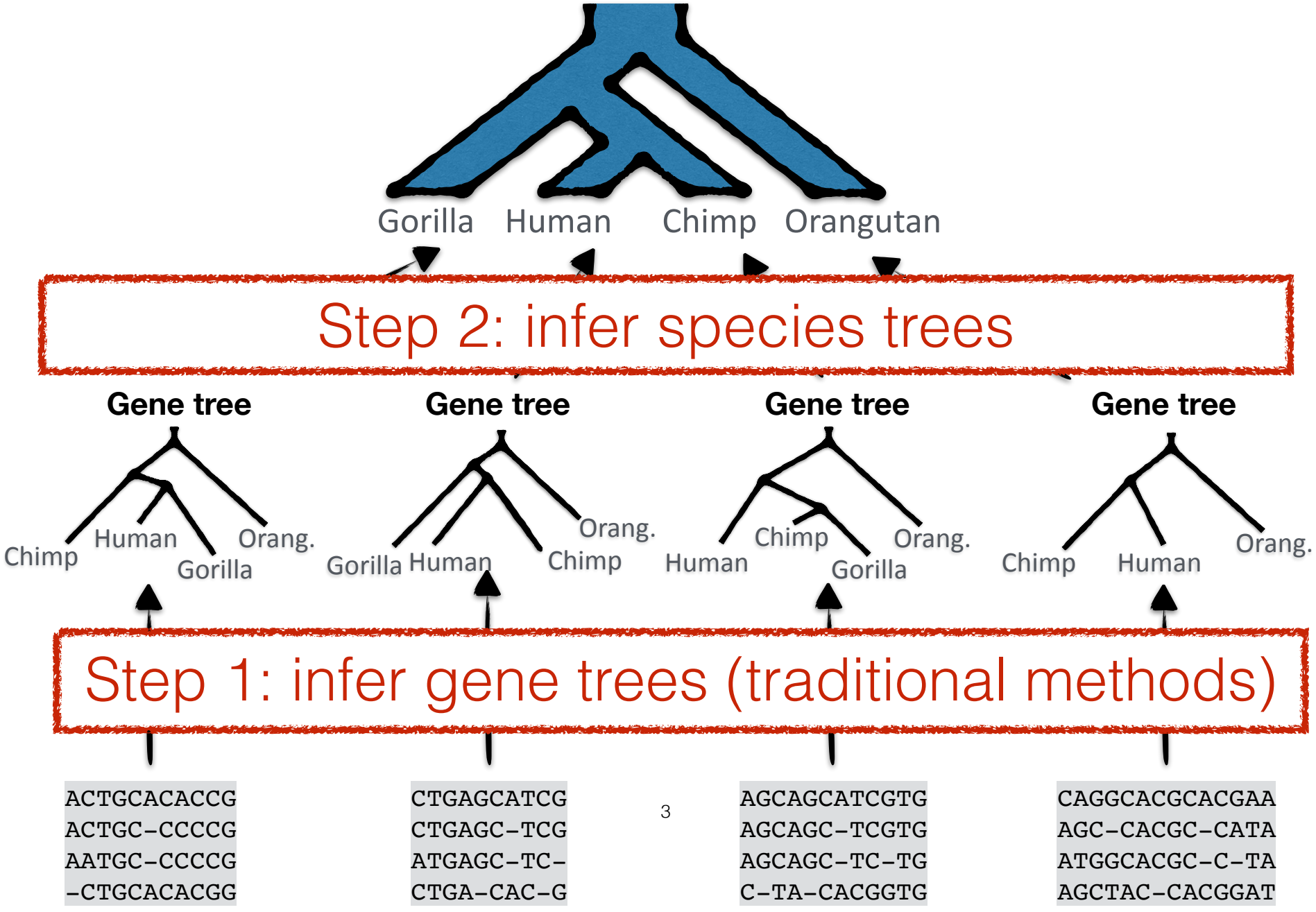
Siavash Mirarab, Texas

Tandy Warnow, Texas and UIUC

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

Major challenge: Massive gene tree heterogeneity


Species tree
Gorilla
Human
Chimp
Orangutan
Gene evolution model
Gene tree
Gene tree
Gene tree
Gene tree
Chimp
Human
Orang.
Gorilla
Gorilla
Human
Orang.
Chimp
Human
Chimp
Orang.
Gorilla
Chimp
Human
Orang.
Sequence evolution model
Sequence data
(Alignments)
Sequence data
(Alignments)
ACTGCACACCG
ACTGC-CCCCG
AATGC-CCCCG
-CTGCACACGG
CTGAGCATCG
CTGAGC-TCG
ATGAGC-TC-
CTGA-CAC-G
AGCAGCATCGTG
AGCAGC-TCGTG
AGCAGC-TC-TG
C-TA-CACGGTG
CAGGCACGCACGAA
AGC-CACGC-CATA
ATGGCACGC-C-TA
AGCTAC-CACGGAT


Gorilla Human Chimp Orangutan
Step 2: infer species trees
Gene tree
Gene tree
Gene tree
Gene tree
Chimp Human Orang. Gorilla
Gorilla Human Orang. Chimp
Human Chimp Orang. Gorilla
Chimp Human Orang.
Step 1: infer gene trees (traditional methods)
ACTGCACACCG
ACTGC-CCCCG
AATGC-CCCCG
-CTGCACACGG
CTGAGCATCG
CTGAGC-TCG
ATGAGC-TC-
CTGA-CAC-G
AGCAGCATCGTG
AGCAGC-TCGTG
AGCAGC-TC-TG
C-TA-CACGGTG
CAGGCACGCACGAA
AGC-CACGC-CATA
ATGGCACGC-C-TA
AGCTAC-CACGGAT

ASTRAL

[Mirarab, et al., ECCB/Bioinformatics, 2014]

- Optimization Problem (NP-Hard):

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

Set of quartet trees induced by T

$$Score(T) = \sum_{t \in \mathcal{T}} |Q(T) \cap Q(t)|$$

a gene tree

all input gene trees

- **Theorem**: Statistically consistent under the multi-species coalescent model when solved exactly

ASTRAL

[Mirarab, et al., ECCB/Bioinformatics, 2014]

- Optimization Problem (NP-Hard):

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

$$Score(T) = \sum_{t \in \mathcal{T}} |Q(T) \cap Q(t)|$$

Set of quartet trees induced by T

a gene tree

all input gene trees

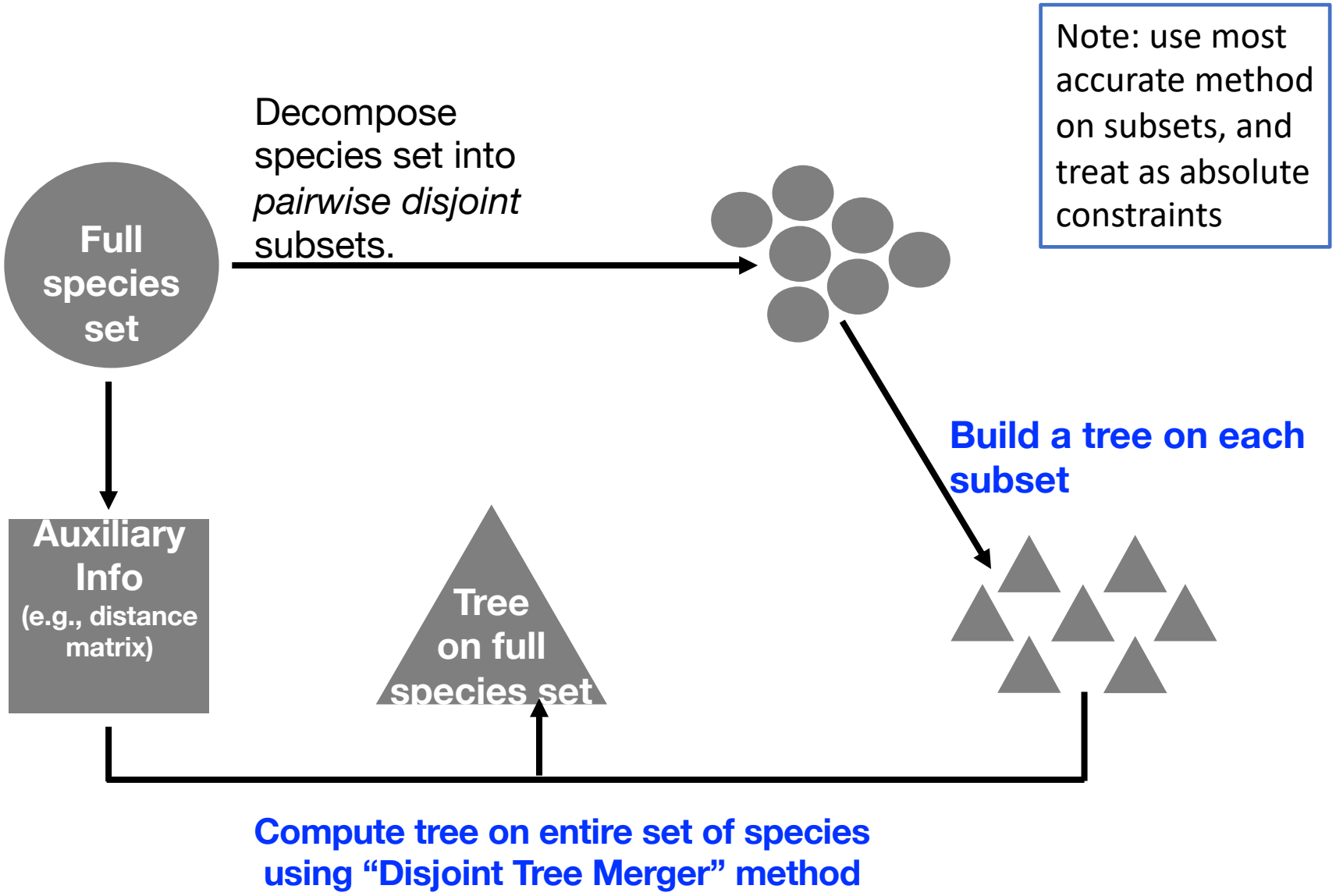
ASTRAL uses dynamic programming to solve a constrained version of this problem, and is provably statistically consistent

- **Theorem**: Statistically consistent under the multi-species coalescent model when solved exactly

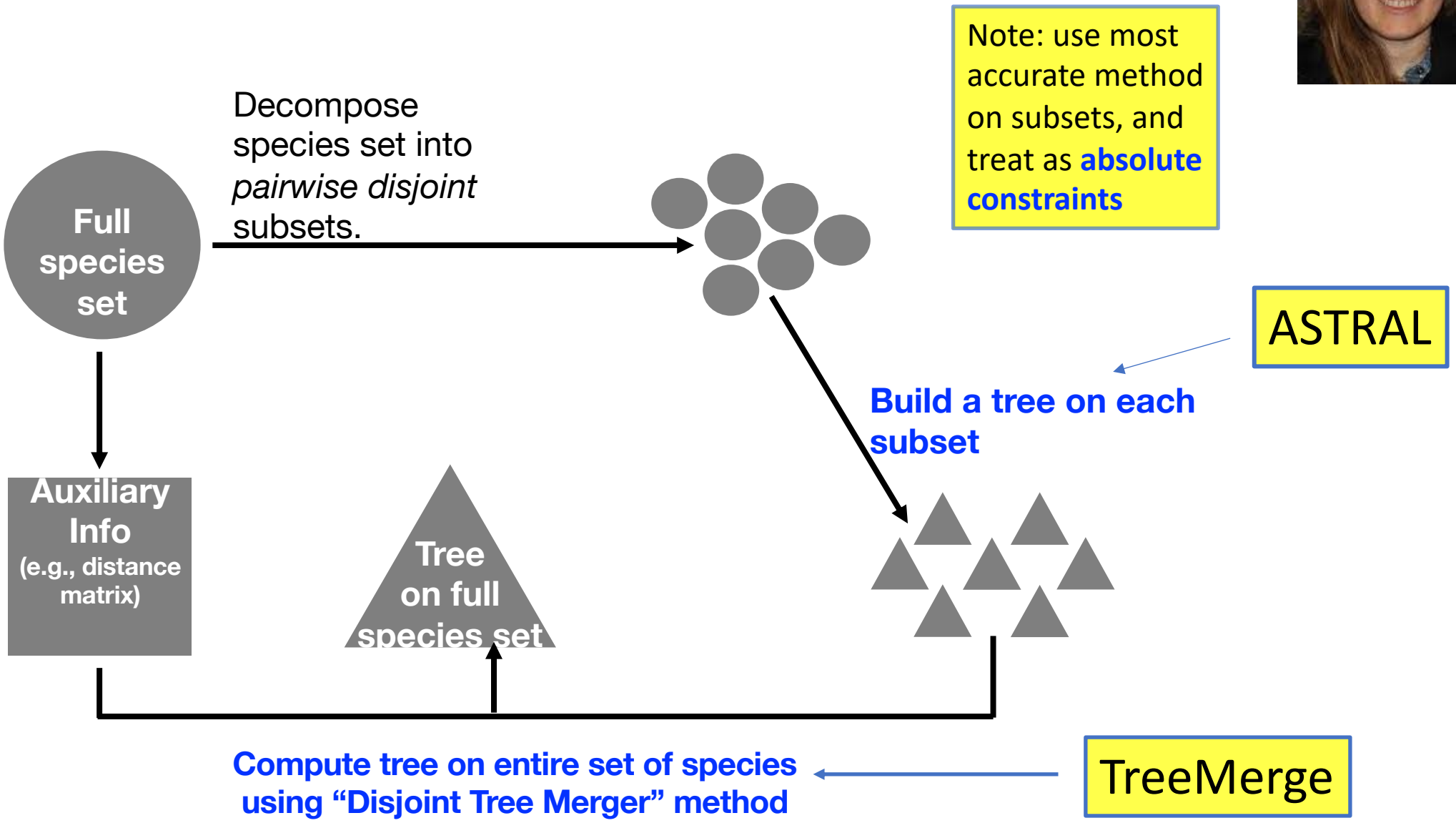
ASTRAL – pros and cons

- The good: ASTRAL is
  - Most popular statistically consistent method for species tree estimation among biologists
  - Very fast for many datasets (much faster than concatenation)
- The mixed:
  - Concatenation can be more accurate under some conditions
- The bad:
  - ASTRAL can fail to complete on large enough datasets within reasonable time frames (days of computation)

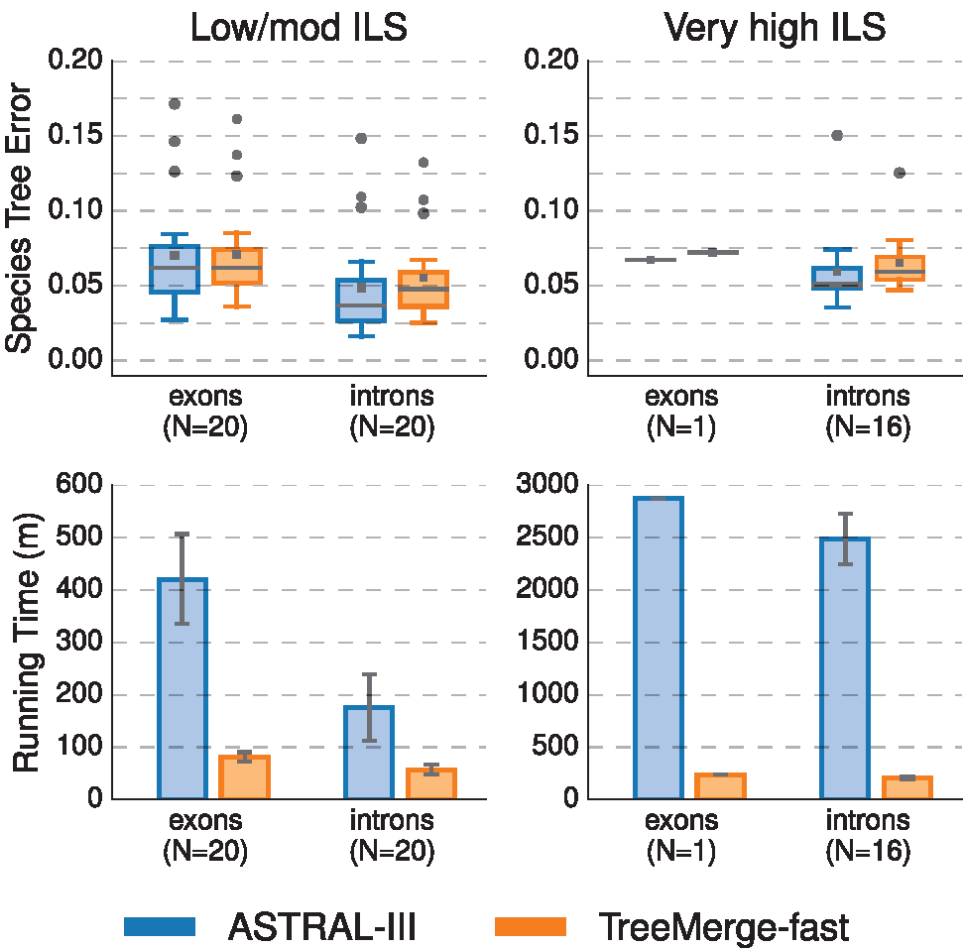
Divide-and-Conquer using Disjoint Tree Mergers



Divide-and-Conquer using Disjoint Tree Mergers



Impact of using TreeMerge with ASTRAL-III on 1000 species and 1000 genes



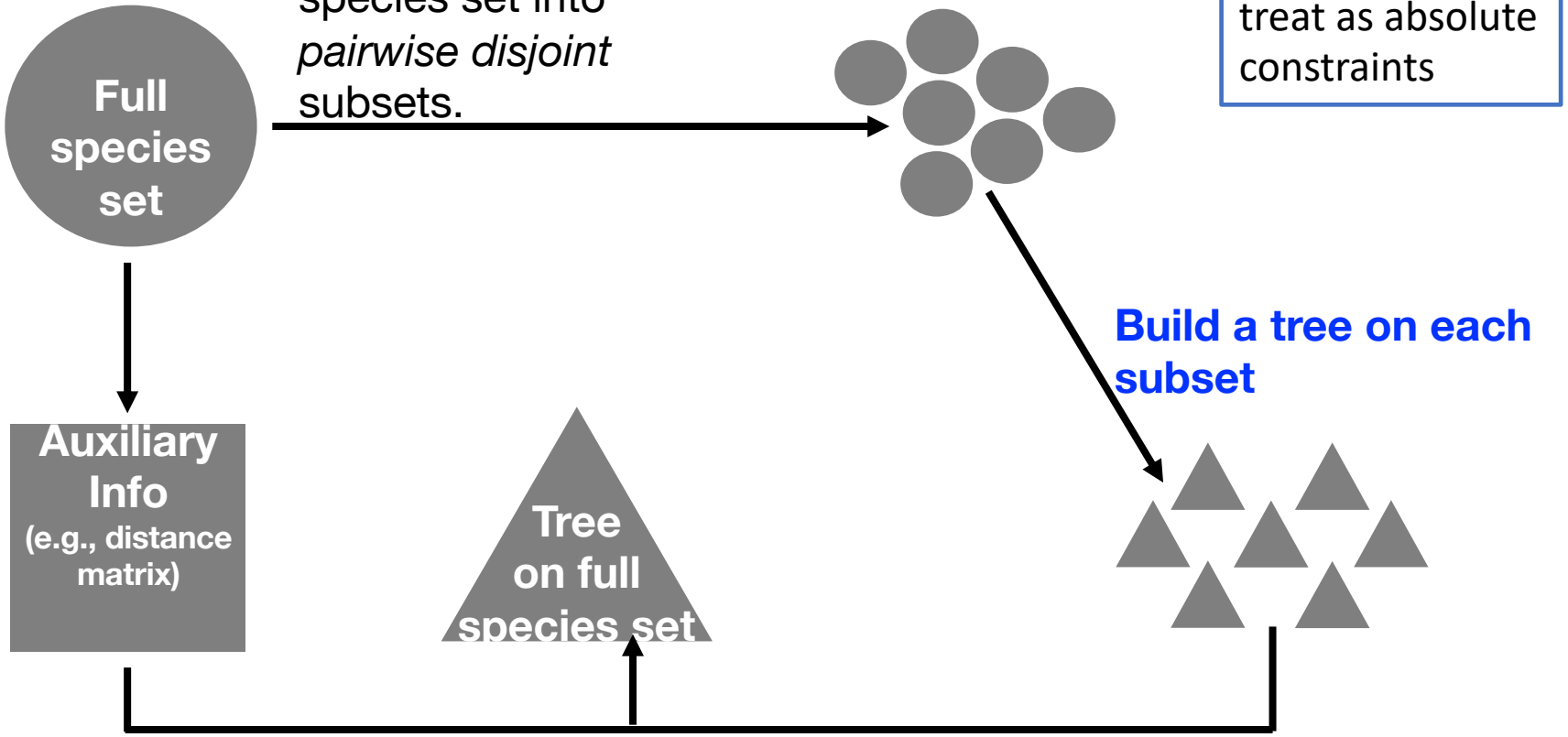
Theorem:
TreeMerge+ASTRAL
is **statistically consistent**
and **polynomial time**

Empirical: TreeMerge
maintains accuracy,
reduces running time,
and improves scalability

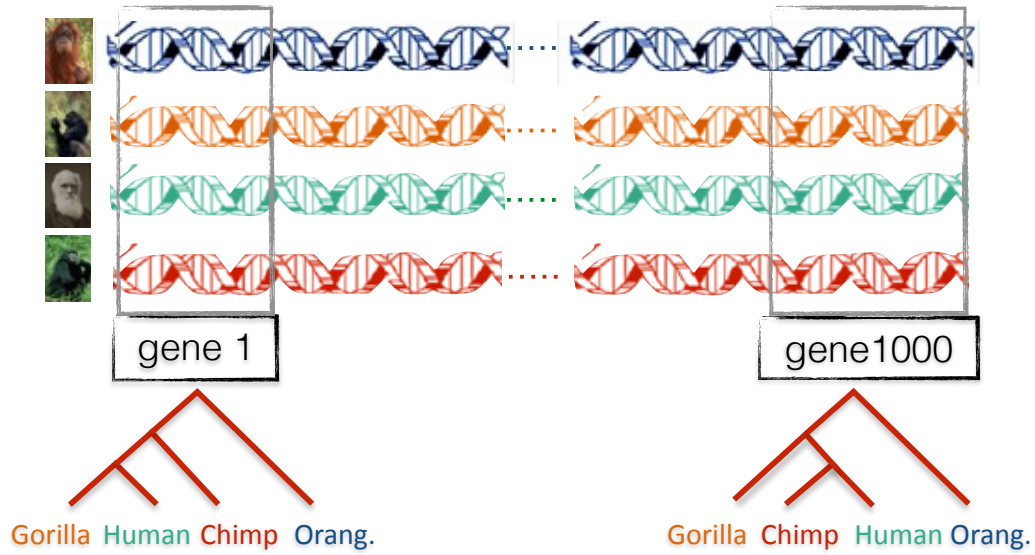
Bioinformatics, Volume 35, Issue 14, July 2019, Pages i417–i426, https://doi.org/10.1093/bioinformatics/btz344

OXFORD
UNIVERSITY PRESS

DTMs can be used for any tree estimation problem



Gene tree discordance



Multiple causes for discord, including

- Incomplete Lineage Sorting (ILS),
- Gene Duplication and Loss (GDL), and
- Horizontal Gene Transfer (HGT)

Gene Family Trees

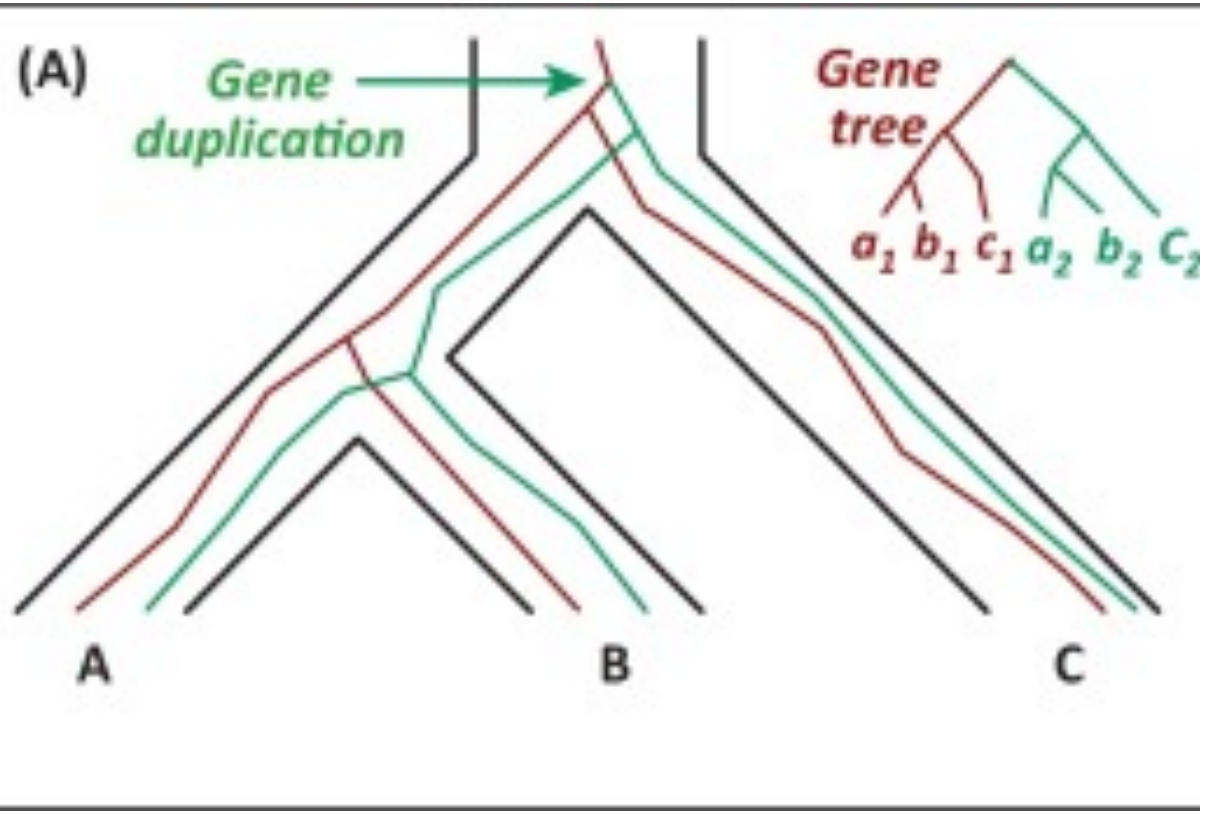
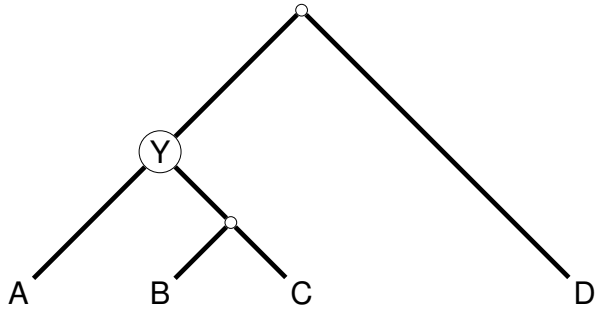


Figure by Luay Nakhleh, TREE 2013

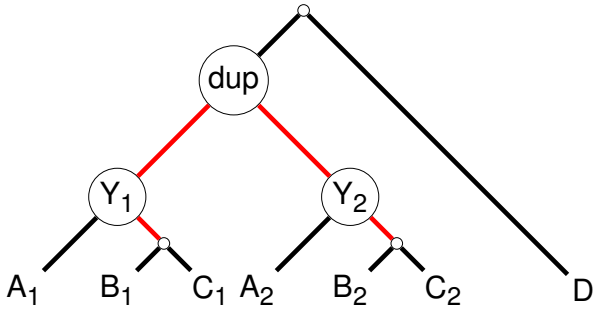
The species tree has one duplication (at the root), which produces a gene family tree that has two copies of the species tree!

Multi-copy trees: MUL-trees

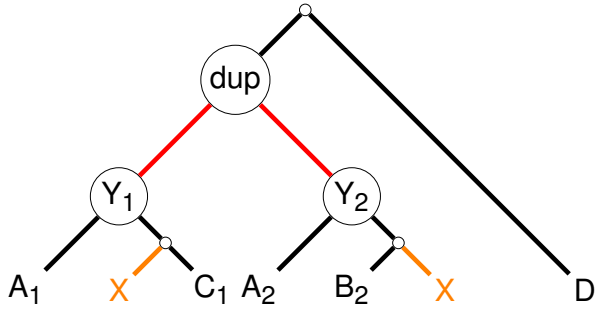
Problem: Given set of MUL-trees, infer the species tree



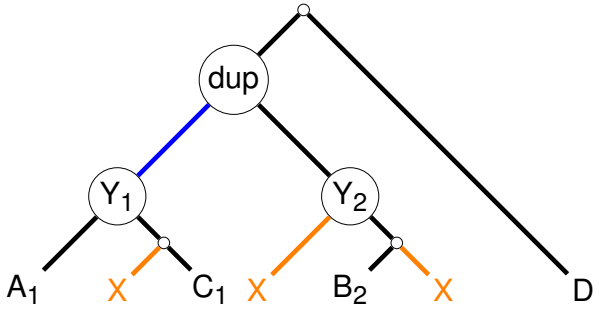
(a) Species tree T^*



(b) Gene tree M_1 with one duplication.



(c) Gene tree M_2 with one duplication and two losses.



(d) Gene tree with one duplication and three losses.

Many methods, but until Fall 2019, none proven statistically consistent under GDL

Theorem (Legried, Molloy, Warnow, and Roch, 2019): **ASTRAL-multi** is statistically consistent under GDL and runs in polynomial time.



Papers on bioRxiv (under review)

Theorem (Legried, Molloy, Warnow, and Roch, 2019): **ASTRAL-multi** is statistically consistent under GDL and runs in polynomial time.

Theorem (Molloy and Warnow, 2019): **FastMulRFS** is statistically consistent under a generic duplication-only or loss-only model, and runs in polynomial time.

Papers on bioRxiv (under review)

Theorem (Legried, Molloy, Warnow, and Roch, 2019): **ASTRAL-multi** is statistically consistent under GDL and runs in polynomial time.

Theorem (Molloy and Warnow, 2019): **FastMulRFS** is statistically consistent under a generic duplication-only or loss-only model, and runs in polynomial time.

Note: Both methods use dynamic programming to solve NP-hard **discrete optimization problems** within constrained search space in polynomial time.



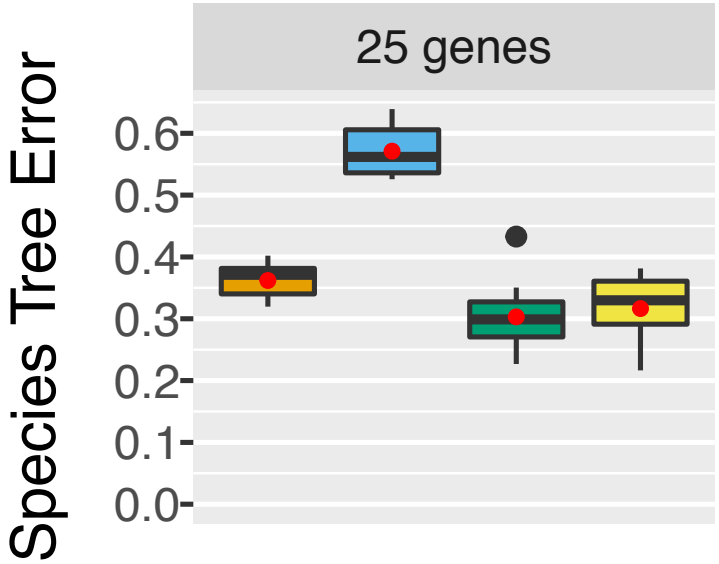
Papers on bioRxiv (under review)

Theorem (Legried, Molloy, Warnow, and Roch, 2019): **ASTRAL-multi** is statistically consistent under GDL and runs in polynomial time.

Theorem (Molloy and Warnow, 2019): **FastMulRFS** is statistically consistent under a generic duplication-only or loss-only model, and runs in polynomial time.

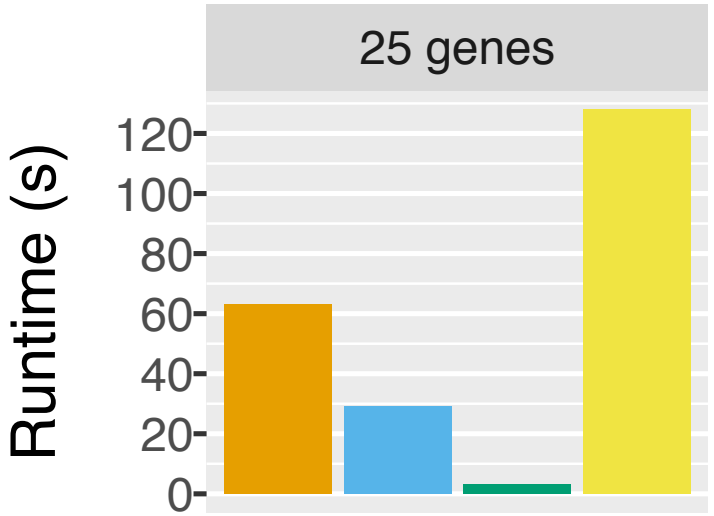
Note: Both methods use dynamic programming to solve NP-hard **discrete optimization problems** within constrained search space in polynomial time.

Papers on bioRxiv (under review)



Results on data (100 species):

- FastMulRFS and MulRF tied for best in terms of accuracy
- FastMulRFS is by far the fastest



Data: 100 species, moderate GDL, moderately high ILS, high gene tree estimation error

ASTRAL DupTree FastMulRFS MulRF

Molloy and Warnow, bioRxiv 2019
doi: https://doi.org/10.1101/835553

Opportunities for PhD students:

- Large impact on biology through innovative algorithm design
- Interesting mathematical problems, including discrete algorithms and machine learning
- Not necessary to understand biology (seriously!)
- Most important skills: enjoying coding, testing, looking at data, and collaborating with other people.
- Many types of research: high performance computing, parallel algorithms, graph algorithms, combinatorial optimization, machine learning, etc.

My students go on to successful careers in academia (UCSD, Rice, etc.) and industry (Apple, Google, Amazon)

Acknowledgments

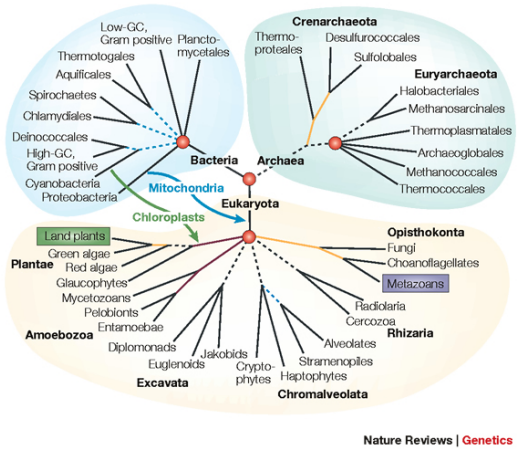
Papers available at http://tandy.cs.illinois.edu/papers.html

Presentations available at http://tandy.cs.illinois.edu/talks.html

Funding: NSF (CCF 1535977 and also NSF Graduate Fellowship to Erin Molloy)

Supercomputers: Blue Waters and Campus Cluster, both supported by NCSA

Phylogenetic Inference



Nature Reviews | Genetics

“Big Data”:

- Heterogeneous
- Large
- Noisy
- Error-ridden
- Streaming
- Model-misspecification

Approaches:

- NP-hard optimization problems and large datasets
- Statistical estimation under stochastic models of evolution
- Probabilistic analysis of algorithms
- Graph-theoretic divide-and-conquer
- Chordal graph theory
- Combinatorial optimization