

394C, Fall 2013

Tandy Warnow
tandy@cs.utexas.edu

Course Title?

- Computational Algorithmic Biology
- Computational Biology
- Big Data in Biology
- Applications of Algorithms and Mathematical Modelling to Biology

Course Title?

- Computational Algorithmic Biology (and Linguistics)
- Computational Biology (and Linguistics)
- Big Data in Biology (and Linguistics)
- Applications of Algorithms and Mathematical Modelling to Biology (and Linguistics)

Basics

- Course topic: algorithmic problems in biology and historical linguistics.
- Objective: give overview of some hot topics in computational biology and computational historical linguistics, and get started on research problems.

Today

- Describe some important problems in computational biology and computational historical linguistics, for which students in this course could develop improved methods.
- Explain how the course will be run.
- Answer questions.

Basics

- Prerequisites: Computer Science (algorithm design and programming), mathematical maturity (ability to understand proofs). No background in biology or linguistics is needed!
- Note: if you are not a CS, ECE, or Math major, you can still take the course – but will do slightly different homework problems. Please see me.

Some Problems

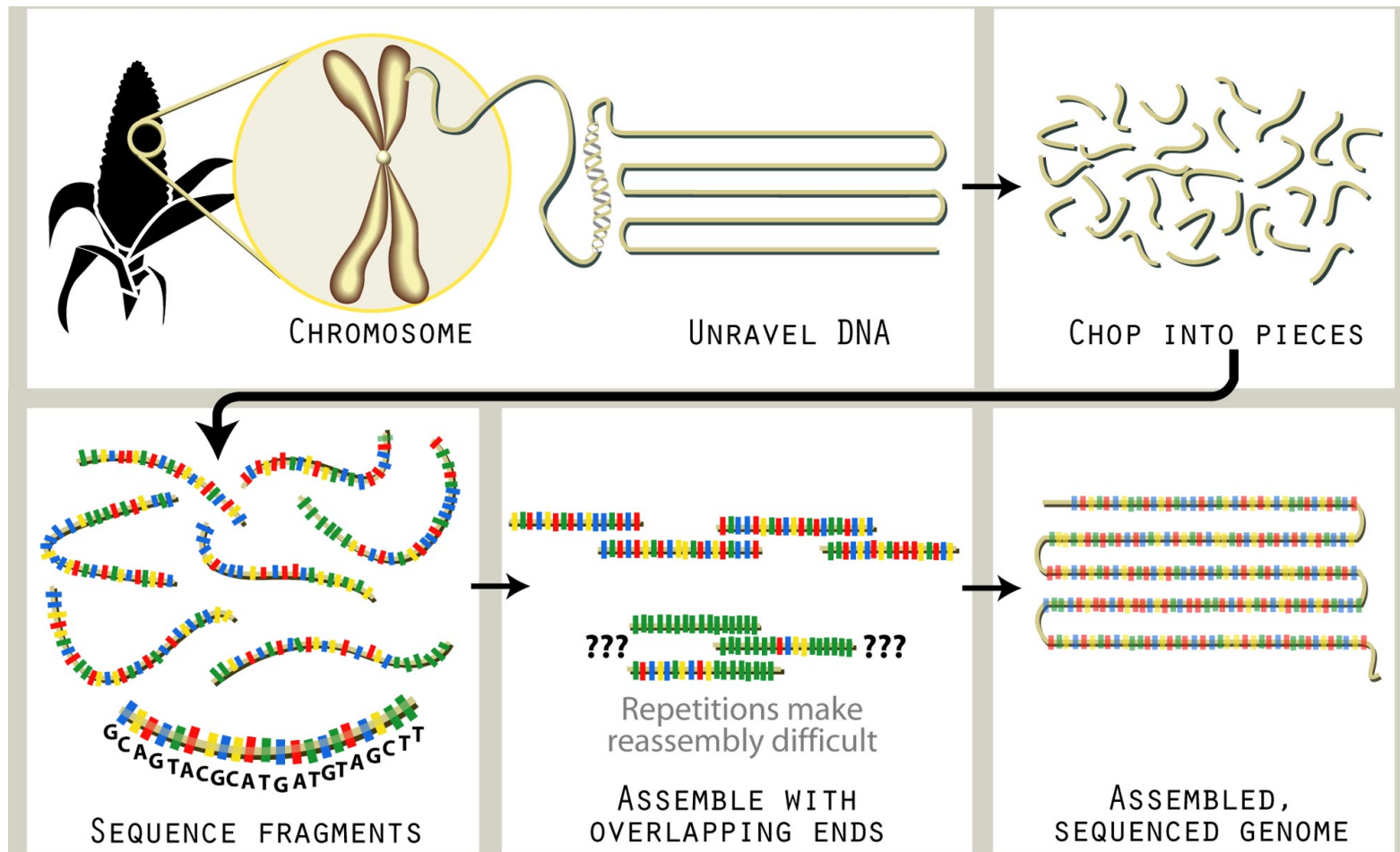
1. Genome Assembly
2. Phylogeny Estimation
3. Multiple Sequence Alignment
4. Metagenomics
5. Reconstructing Language
Phylogenies

TACC



Courtesy of NSF

Genome Assembly



Courtesy of NSF

Genome Assembly Algorithms

- Beautiful Graph Theory and Algorithms:
 - DeBruijn Graphs
 - Hamiltonian Cycles
 - Eulerian Graphs

Genome Assembly Challenges

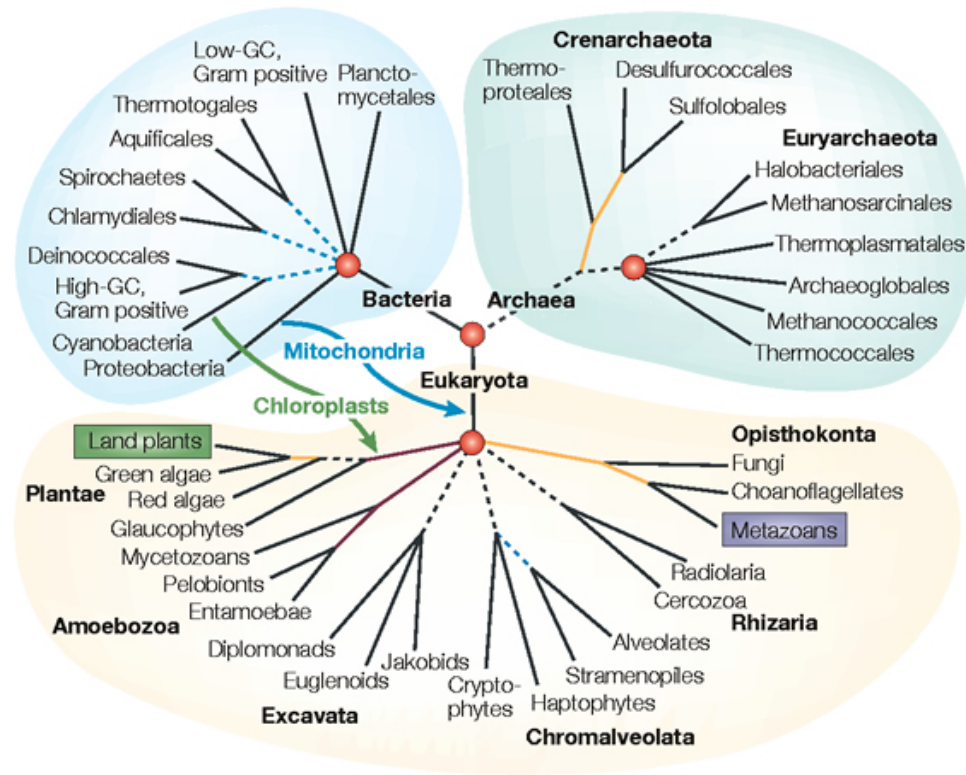
- Sanger sequencing (old approach) produced long reads with low error rates
- Next Generation Sequencing platforms make shorter reads, but at a greatly reduced cost
- Illumina sequencing has few “indels” but very short reads, while 454 sequencing has longer reads with more indels
- Assembly of genomes from NGS data is much harder than from Sanger sequencing data.

Genome Assembly Challenges

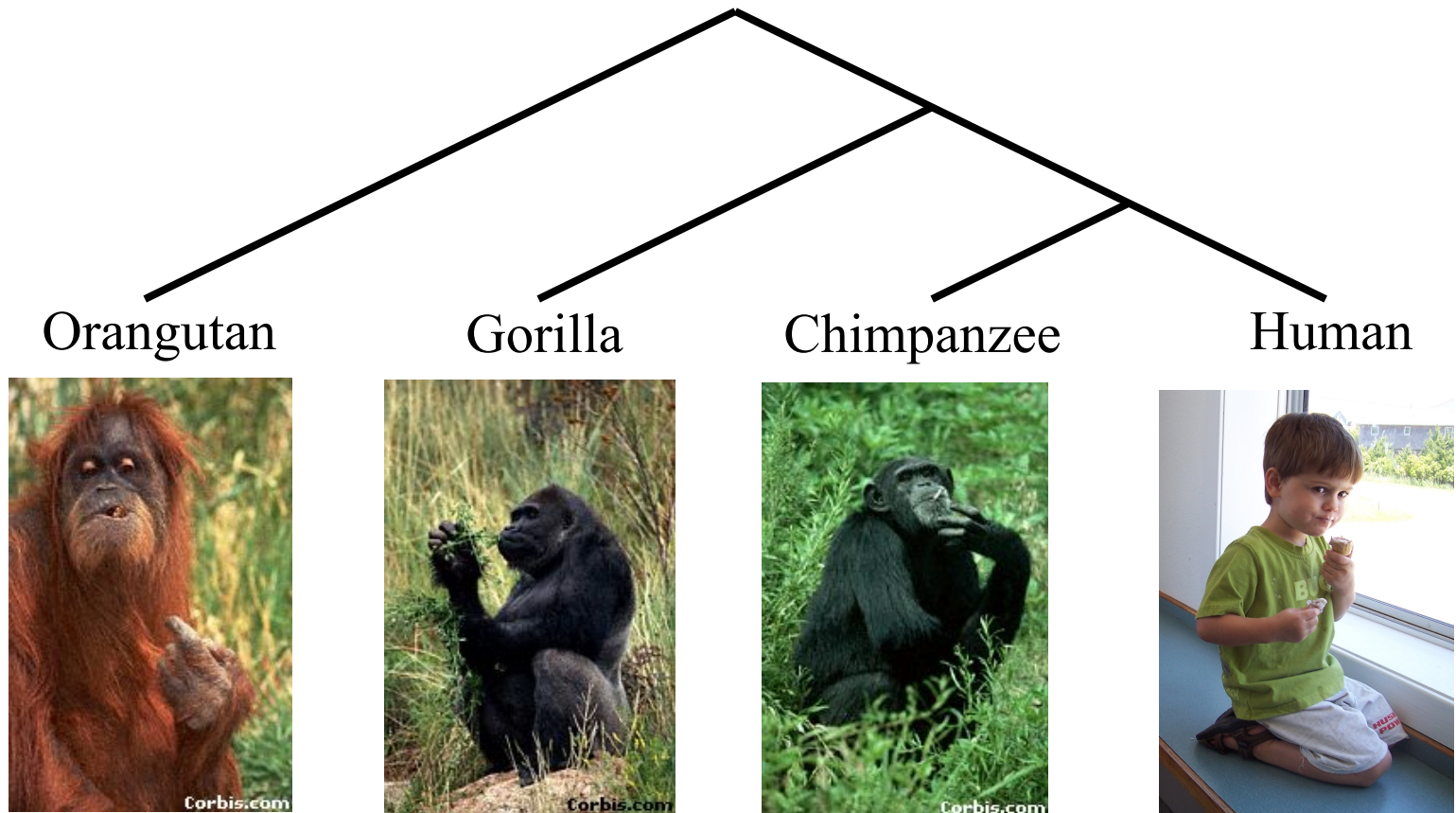
- Sanger sequencing (old approach) produced long reads with low error rates
- Next Generation Sequencing platforms make shorter reads, but at a greatly reduced cost
- Illumina sequencing has few “indels” but very short reads, while 454 sequencing has longer reads with more indels
- Assembly of genomes from NGS data is much harder than from Sanger sequencing data.

New data are not like traditional data, and are often harder to analyze!

Assembling the Tree of Life

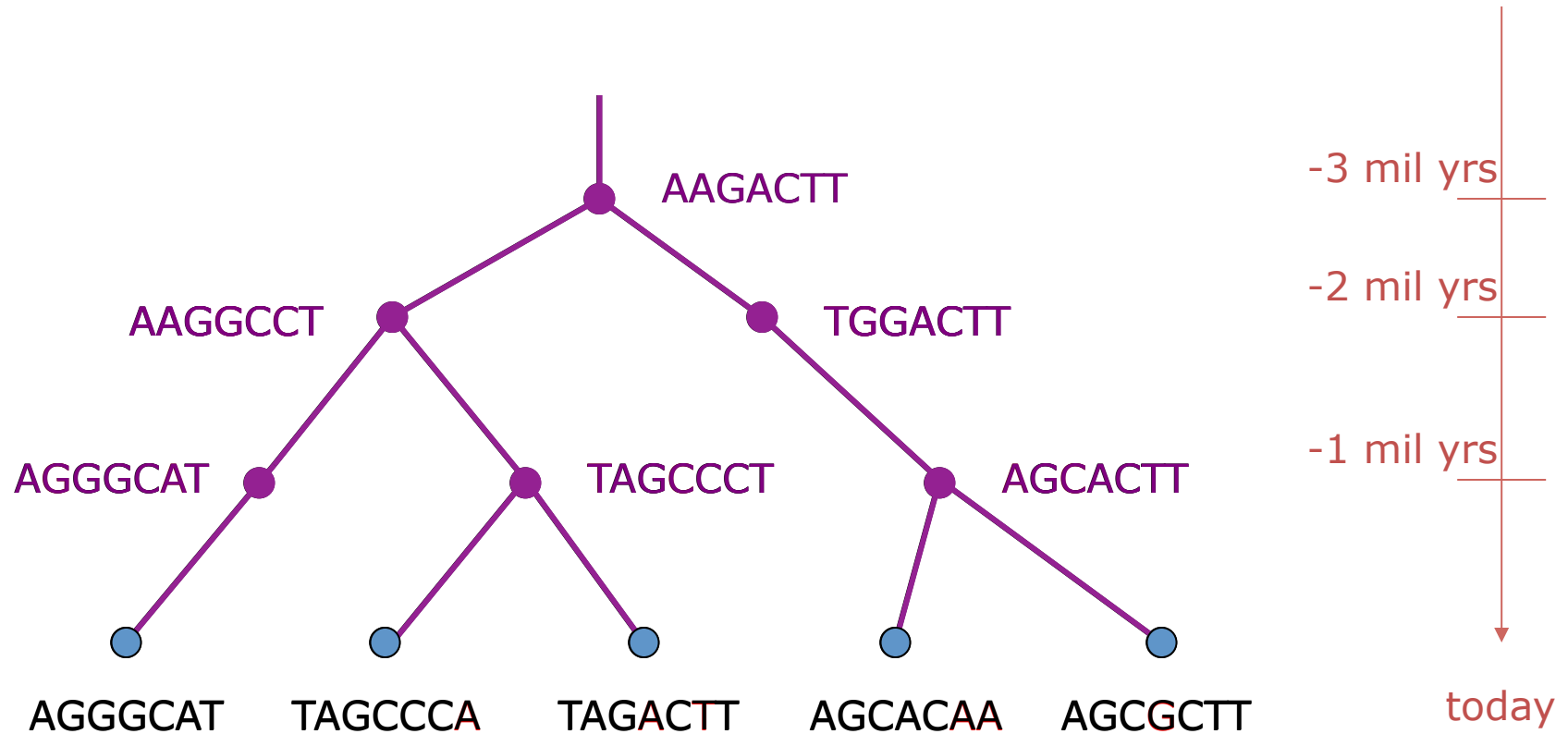


Species Tree

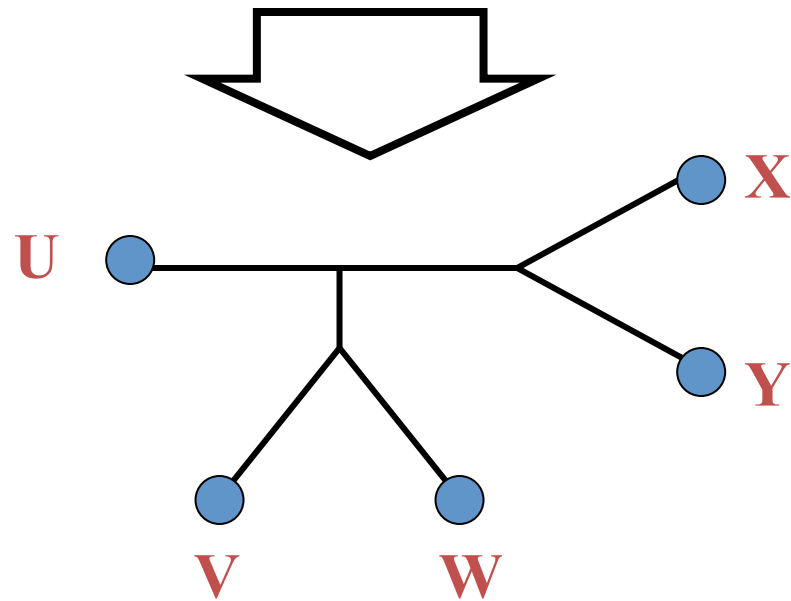


*From the Tree of the Life Website,
University of Arizona*

DNA Sequence Evolution



U AGGGCATGA V AGAT W TAGACTT X TGCACAA Y TGCGCTT





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

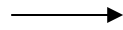
S2 = TAGCTATCACGACCGC

S3 = TAGCTGACCGC

S4 = TCACGACCGACA

Phase 1: Alignment

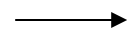
S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC
S3 = TAGCTGACCGC
S4 = TCACGACCGACA



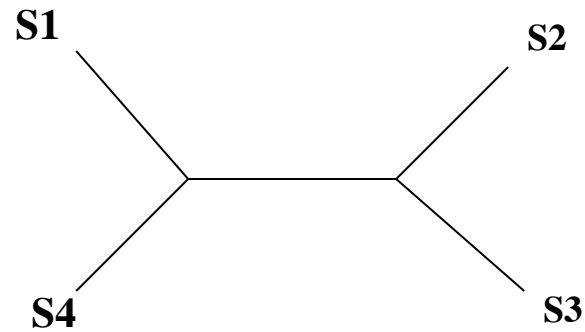
S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAC--GACCGC--
S3 = TAG-CT-----GACCGC--
S4 = -----TCAC--GACCGACA

Phase 2: Construct tree

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



S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAC--GACCGC--
S3 = TAG-CT-----GACCGC--
S4 = -----TCAC--GACCGACA



Two-phase estimation

Alignment methods

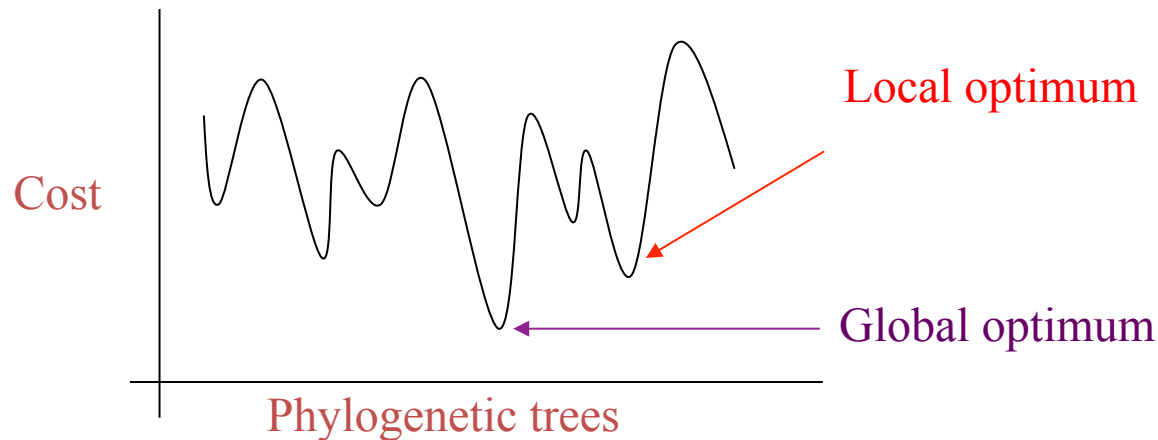
- Clustal
- POY (and POY*)
- Probcons (and Probtree)
- Probalign
- MAFFT
- Muscle
- Di-align
- T-Coffee
- Prank (PNAS 2005, Science 2008)
- Opal (ISMB and Bioinf. 2007)
- *FSA (PLoS Comp. Bio. 2009)*
- *Infernal (Bioinf. 2009)*
- Etc.

Phylogeny methods

- Bayesian MCMC
- Maximum parsimony
- Maximum likelihood
- Neighbor joining
- FastME
- UPGMA
- Quartet puzzling
- Etc.

Phylogenetic reconstruction methods

1. Hill-climbing heuristics for hard optimization criteria (Maximum Parsimony and Maximum Likelihood)



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

Solving maximum likelihood (and other hard optimization problems) is... unlikely

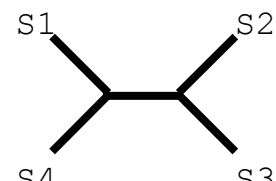
# of Taxa	# of Unrooted Trees
4	3
5	15
6	105
7	945
8	10395
9	135135
10	2027025
20	2.2×10^{20}
100	4.5×10^{190}
1000	2.7×10^{2900}

Simulation Studies

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

Unaligned
Sequences

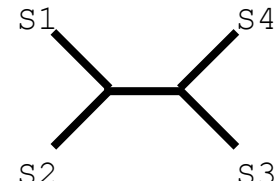
```
S1 = -AGGCTATCACCTGACCTCCA  
S2 = TAG-CTATCAC--GACCGC--  
S3 = TAG-CT-----GACCGC--  
S4 = -----TCAC--GACCGACA
```



A phylogenetic tree diagram showing the relationships between four sequences. The root is at the bottom center. Two branches lead to nodes S1 and S2 at the top. From the node leading to S1, a branch leads to S4 at the bottom. From the node leading to S2, a branch leads to S3 at the bottom.

True tree and
alignment

```
S1 = -AGGCTATCACCTGACCTCCA  
S2 = TAG-CTATCAC--GACCGC--  
S3 = TAG-C--T-----GACCGC--  
S4 = T---C-A-CGACCGA-----CA
```

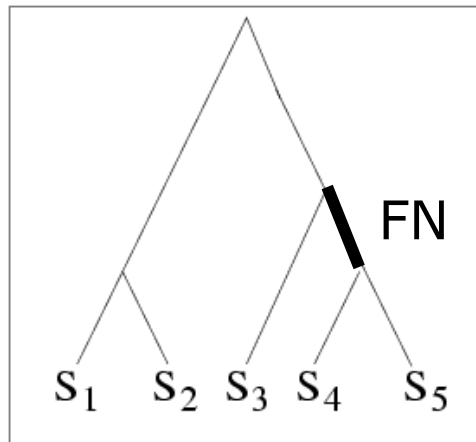


A phylogenetic tree diagram showing the estimated relationships between four sequences. The root is at the bottom center. Two branches lead to nodes S1 and S4 at the top. From the node leading to S1, a branch leads to S2 at the bottom. From the node leading to S4, a branch leads to S3 at the bottom.

Estimated tree and
alignment

Compare

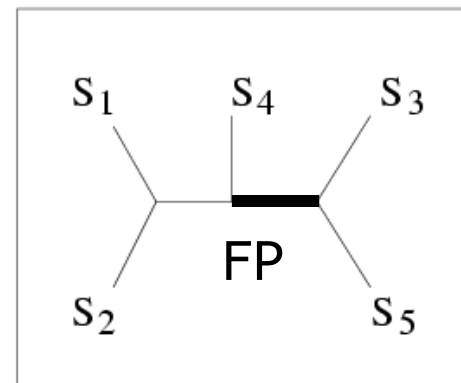
Quantifying Error



TRUE TREE

S ₁	ACAATTAGAAC
S ₂	ACCCTTAGAAC
S ₃	ACCATTCCAAC
S ₄	ACCAGACCAAC
S ₅	ACCAGACCGGA

DNA SEQUENCES

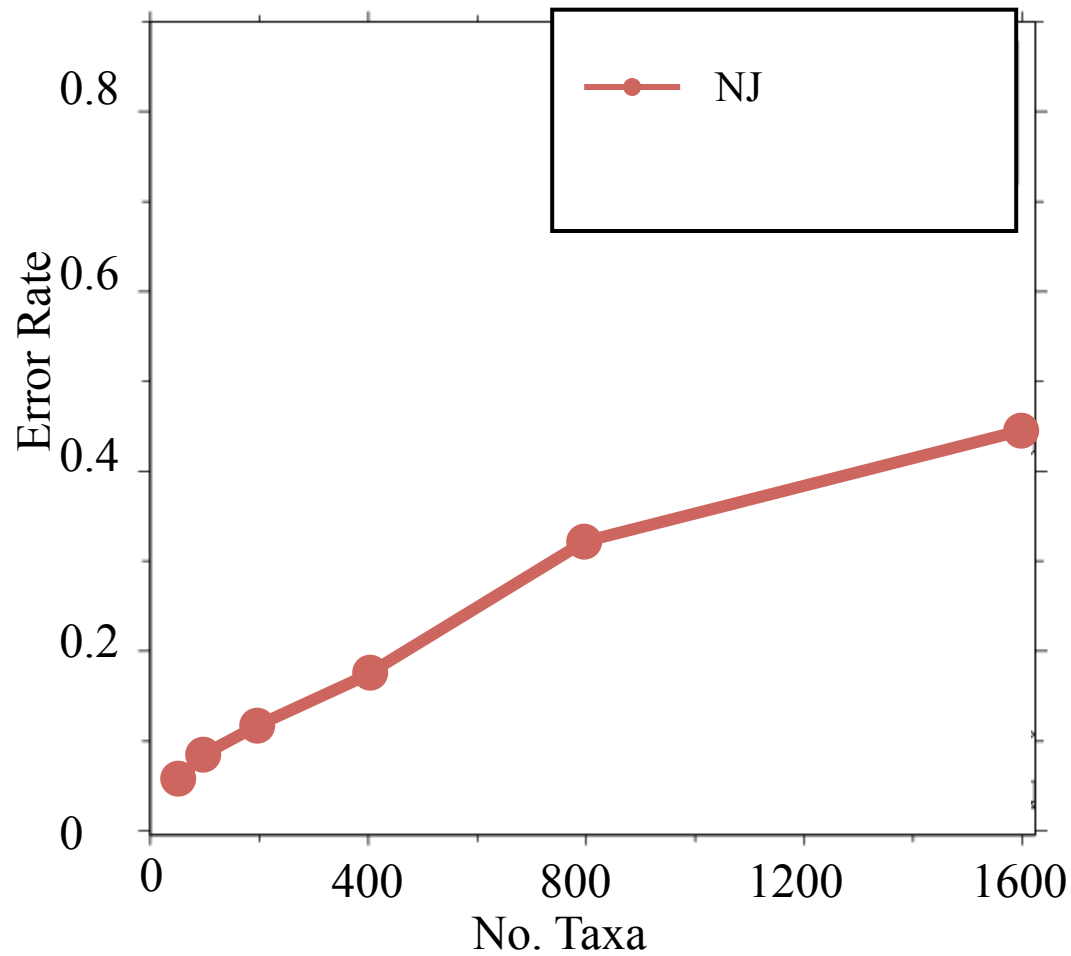


INFERRED TREE

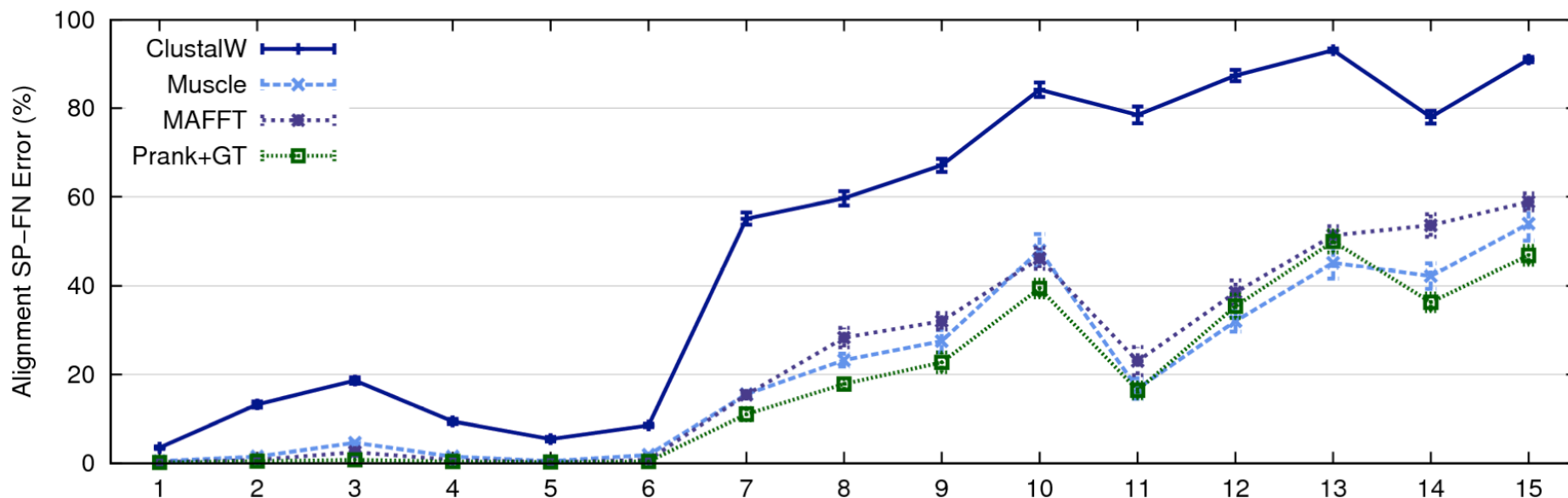
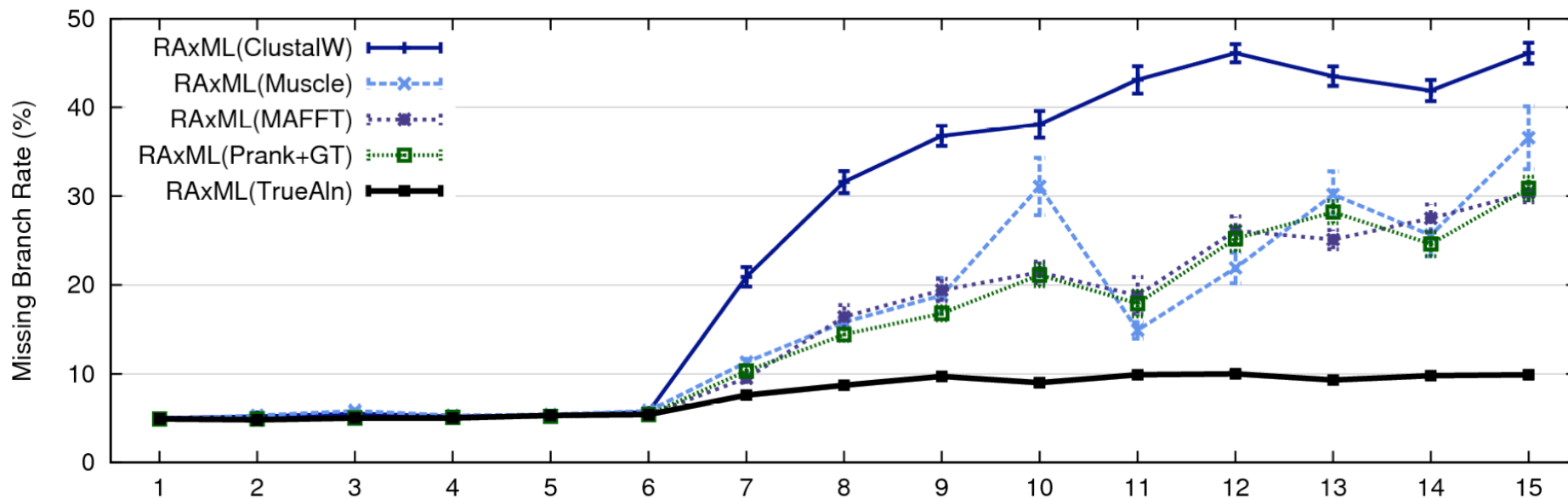
FN: false negative
(missing edge)
FP: false positive
(incorrect edge)

50% error rate

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



Theorem (Atteson):
Exponential sequence
length requirement for
Neighbor Joining!

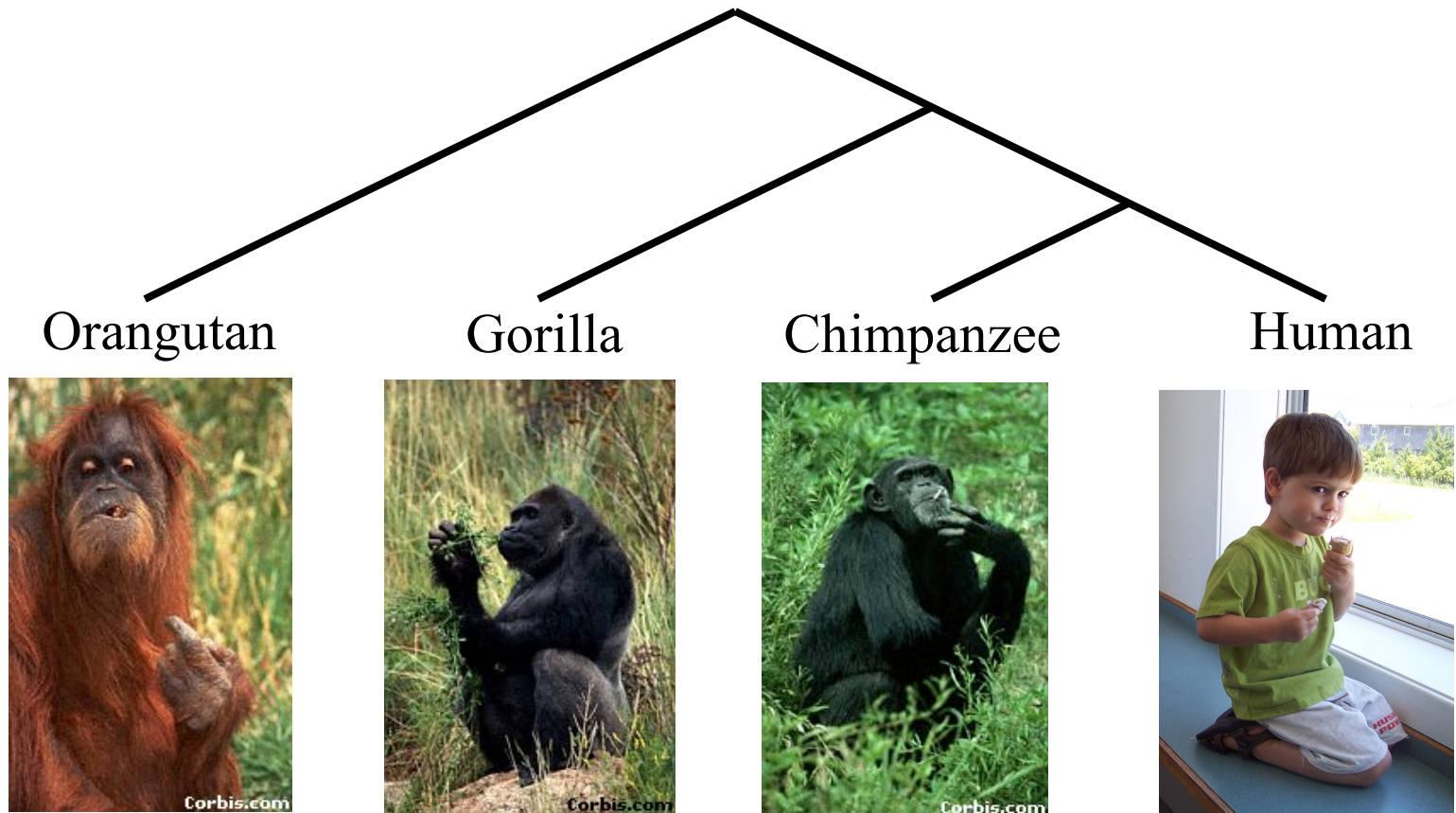


1000 taxon models, ordered by difficulty (Liu et al., 2009)

Major Challenges

- **Phylogenetic analyses:** 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)
- **Multiple sequence alignment:** key step for many biological questions (protein structure and function, phylogenetic estimation), but few methods can run on large datasets. Alignment accuracy is generally poor for large datasets with high rates of evolution.

Species Tree Estimation requires multiple genes!



*From the Tree of the Life Website,
University of Arizona*

Two basic approaches for species tree estimation

- Concatenate (“combine”) sequence alignments for different genes, and run phylogeny estimation methods
- Compute trees on individual genes and combine gene trees

Not all genes present in all species

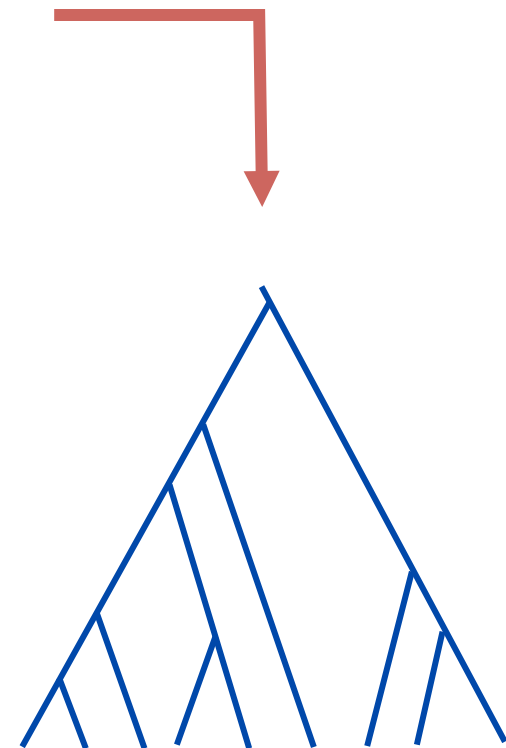
	gene 1
S ₁	TCTAATGGAA
S ₂	GCTAAGGGAA
S ₃	TCTAAGGGAA
S ₄	TCTAACGGAA
S ₇	TCTAATGGAC
S ₈	TATAACGGAA

	gene 2
S ₄	GGTAACCCTC
S ₅	GCTAAACCTC
S ₆	GGTGACCATC
S ₇	GCTAAACCTC

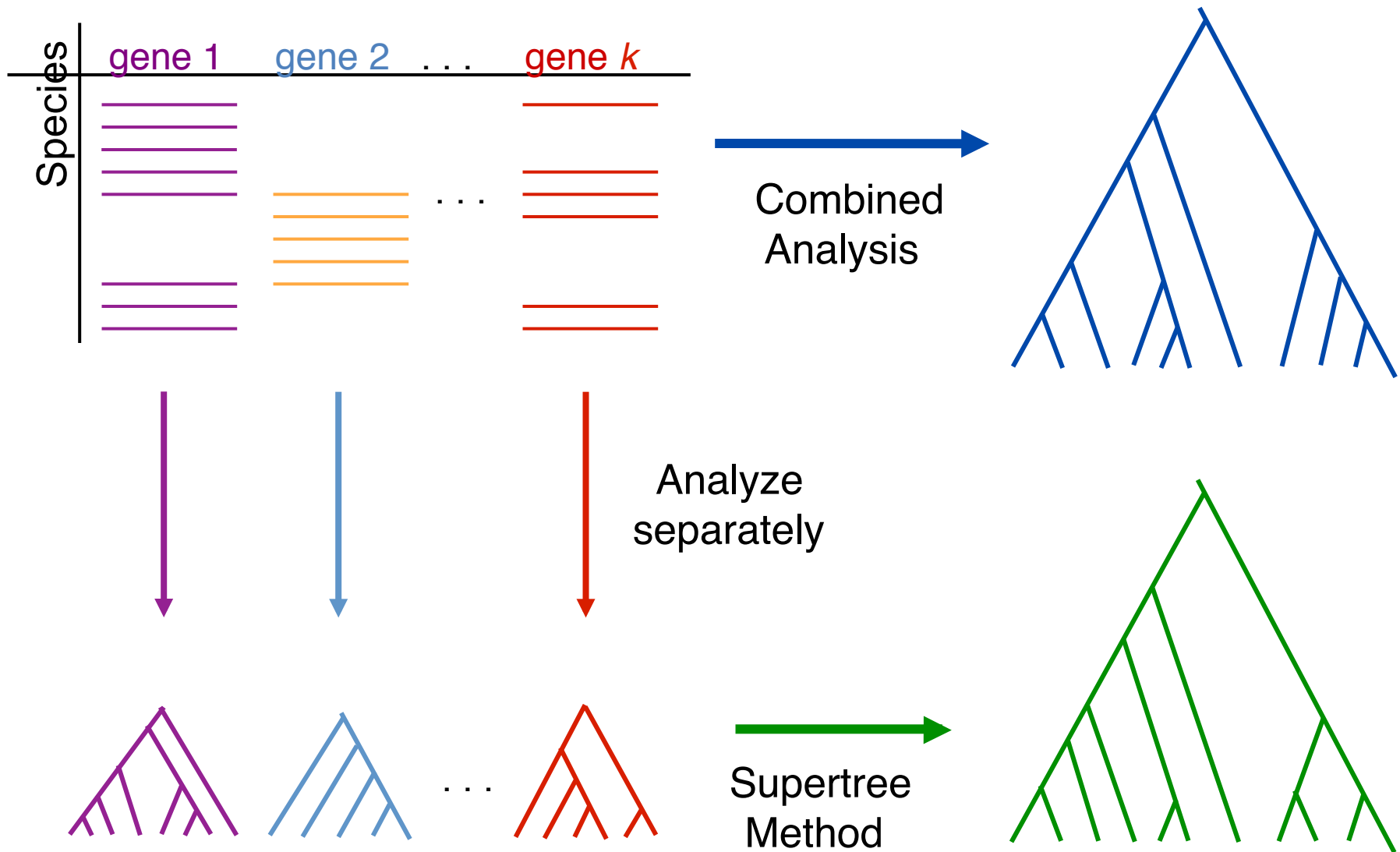
	gene 3
S ₁	TATTGATACA
S ₃	TCTTGATACC
S ₄	TAGTGATGCA
S ₇	TAGTGATGCA
S ₈	CATTCATACC

Combined analysis

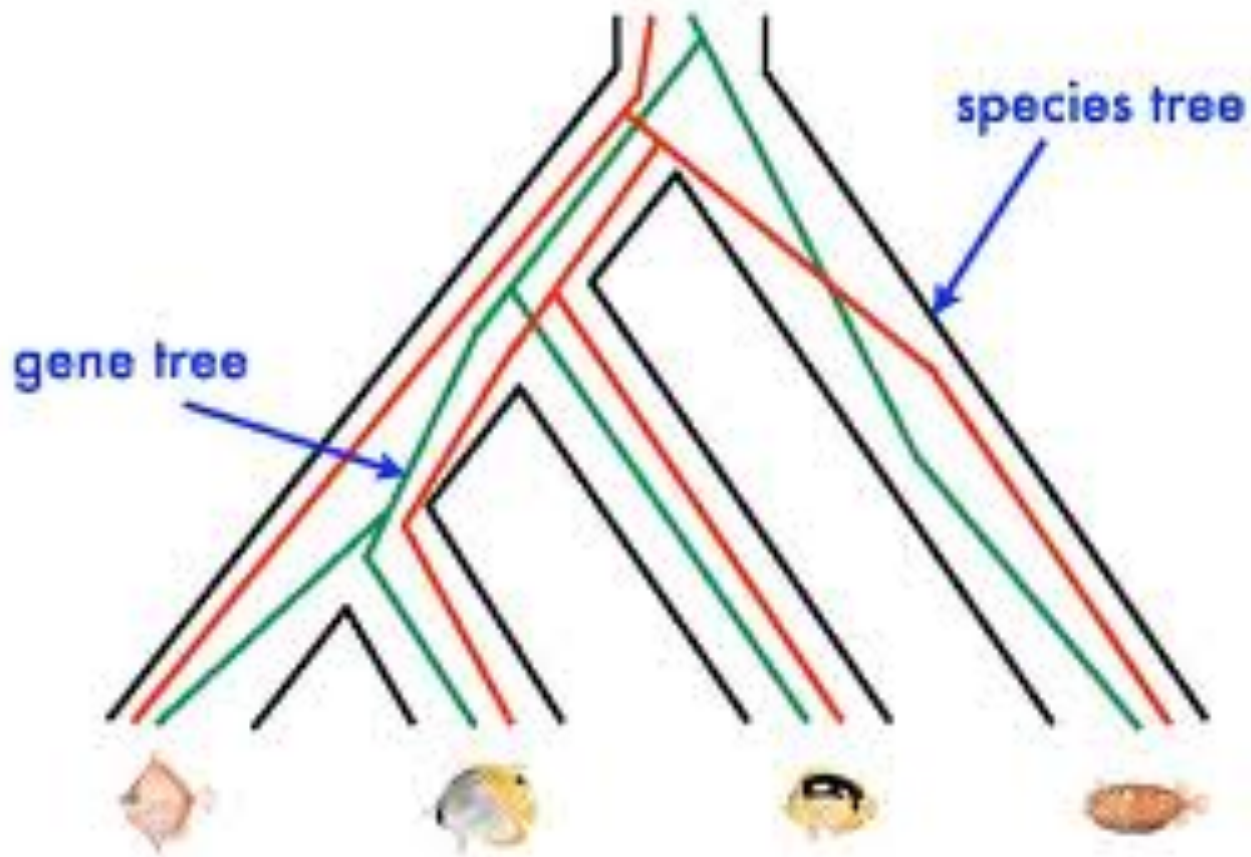
	gene 1	gene 2	gene 3
S ₁	TCTAATGGAA	??????????	TATTGATACA
S ₂	GCTAAGGGAA	??????????	??????????
S ₃	TCTAAGGGAA	??????????	TCTTGATACC
S ₄	TCTAACGGAA	GGTAACCCTC	TAGTGATGCA
S ₅	??????????	GCTAAACCTC	??????????
S ₆	??????????	GGTGACCATC	??????????
S ₇	TCTAATGGAC	GCTAAACCTC	TAGTGATGCA
S ₈	TATAACGGAA	??????????	CATTCATACC



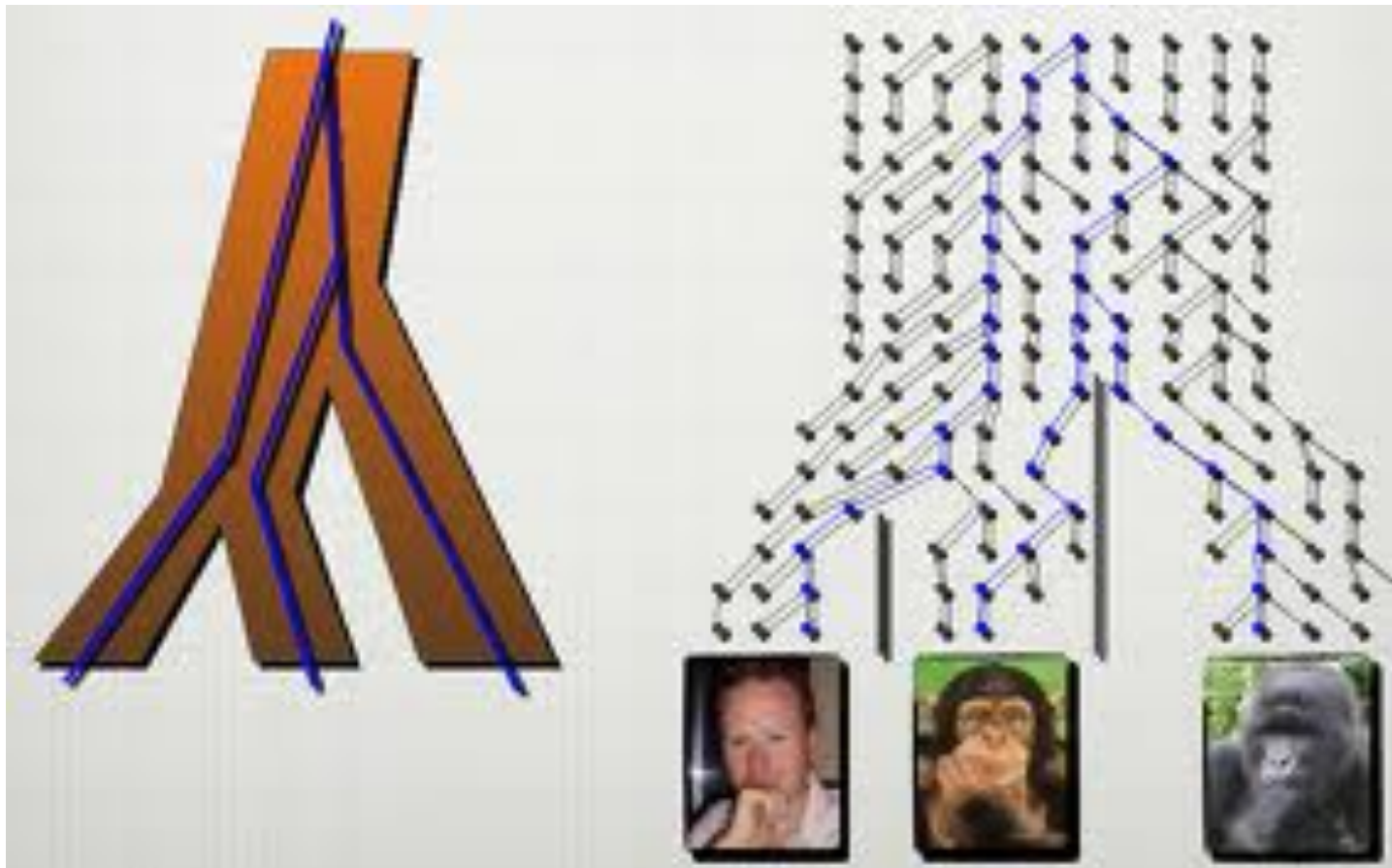
Two competing approaches



**Red gene tree \neq species tree
(green gene tree okay)**



Deep Coalescence



1kp (<http://www.onekp.com/>)



Gane Ka-Shu
Wong
U Alberta



Jim
Leebens-Mack
U Georgia



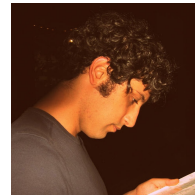
Norm
Wickett
Northwestern



Naim Matasci
iPlant – U Arizona



Tandy Warnow,



Siavash Mirarab,
UT-Austin



Nam Nguyen, and



Md. S. Bayzid

- Transcriptomes of approx. 1200 species
- More than 13,000 gene families (most not single copy)
- Multi-institutional project (10+ universities)

Challenges:

- Estimating very large gene alignments and trees (100,000+ sequences)
- Estimating species trees from incongruent gene trees

Avian Phylogenomics Project

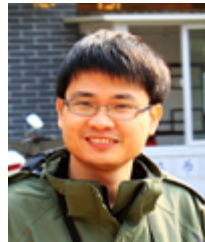
E. Jarvis,
HHMI



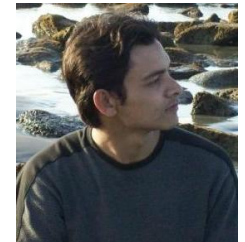
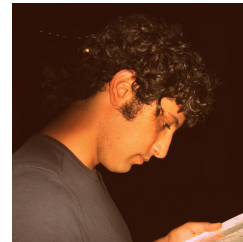
MTP Gilbert,
Copenhagen



G. Zhang,
BGI



S. Mirarab, T. Warnow, and Md. S. Bayzid,
UT-Austin



- Approx. 50 species, whole genomes
- 8000+ genes, UCEs
- Gene trees and sequence alignments computed using SATé
- Species tree estimated using maximum likelihood (RAxML)
- Multi-national team (20+ investigators)

Biggest challenges:

Estimating species tree from incongruent gene trees,
Poor phylogenetic signal in most genes

“Big” phylogenetic datasets

- Large numbers of genes
 - “Concatenation” can become computationally infeasible
 - Gene tree incongruence can make accurate species tree estimation challenging

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*.

BigData in Phylogenetics

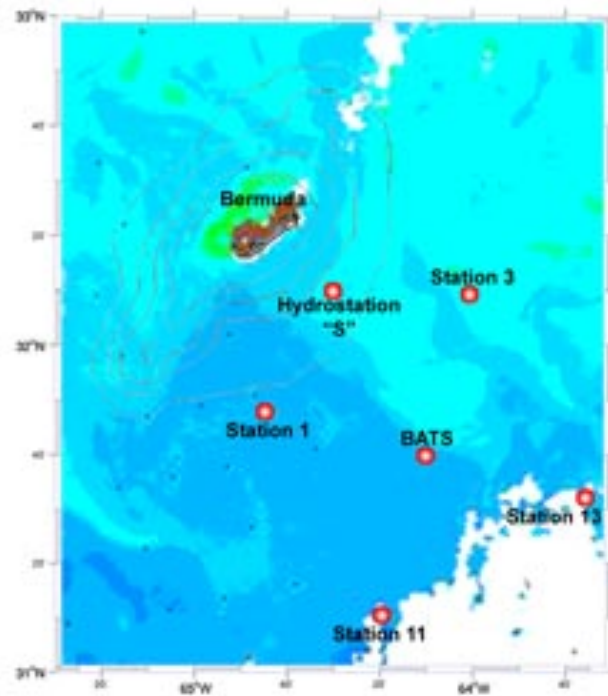
- Many phylogenetic datasets contain hundreds to thousands of species, some with thousands of genes.
- Future datasets will be substantially larger (e.g., iPlant plans to construct a tree on 500,000 plant species)

Our research group is working on datasets with more than **100,000 species**, and some datasets with **thousands of genes**.

Metagenomics:

Venter et al., Exploring the Sargasso Sea:

Scientists Discover One Million New Genes in Ocean Microbes



Metagenomic data analysis

NGS data produce fragmentary sequence data
Metagenomic analyses include unknown species

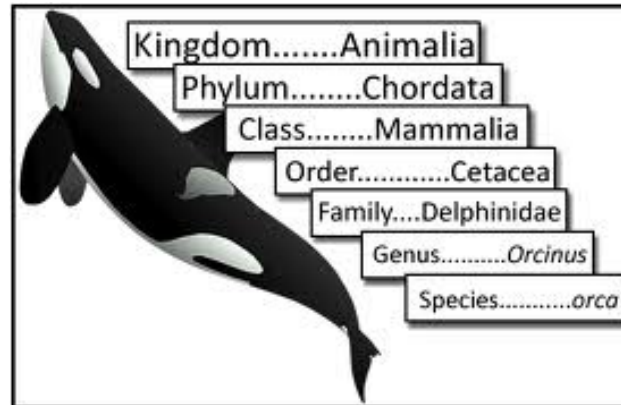
Taxon identification: given short sequences, identify the species for each fragment

Applications: Human Microbiome

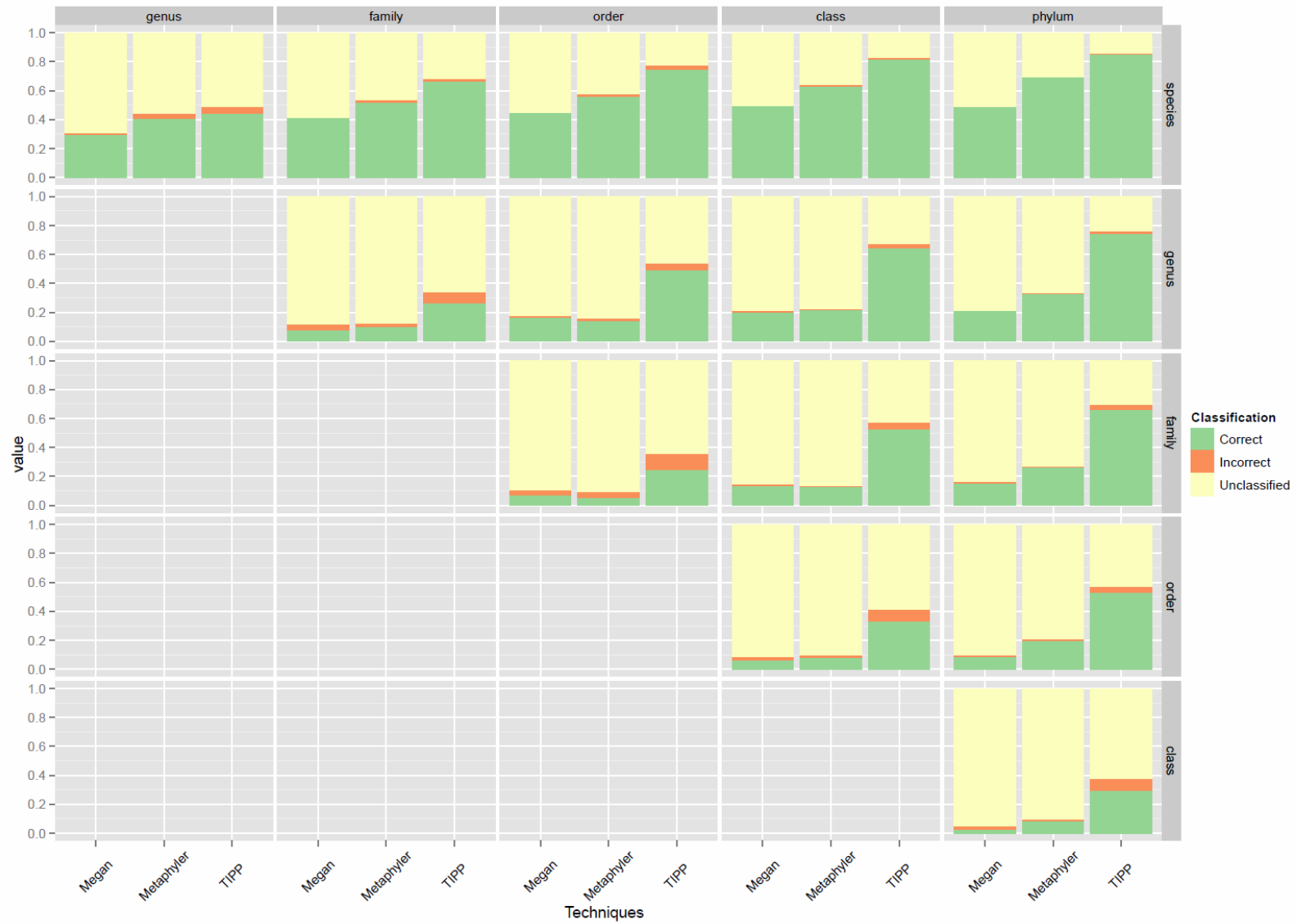
Issues: accuracy and speed

Taxon Identification

Objective: classify short reads in a metagenomic sample



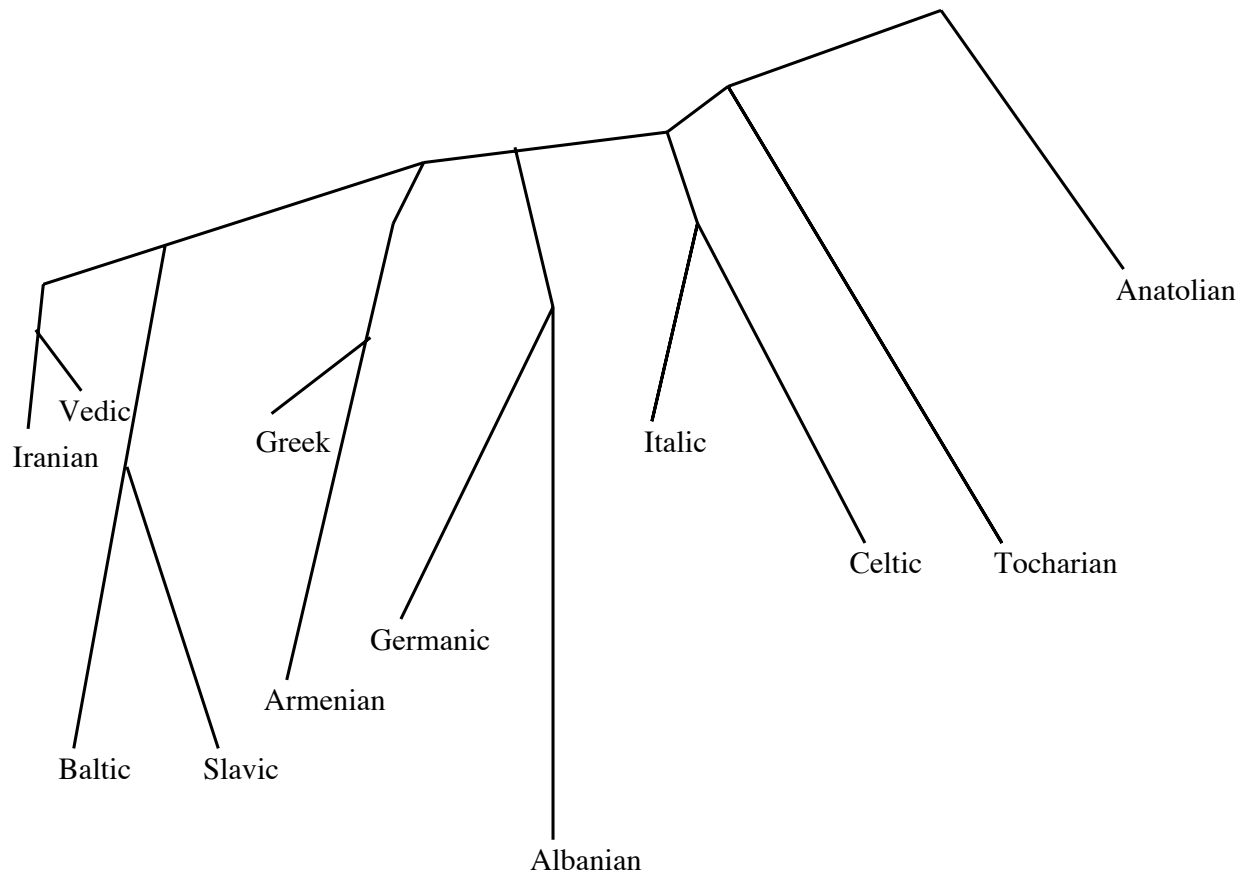
60bp error-free reads on rpsB marker gene



Evolution informs about everything in biology

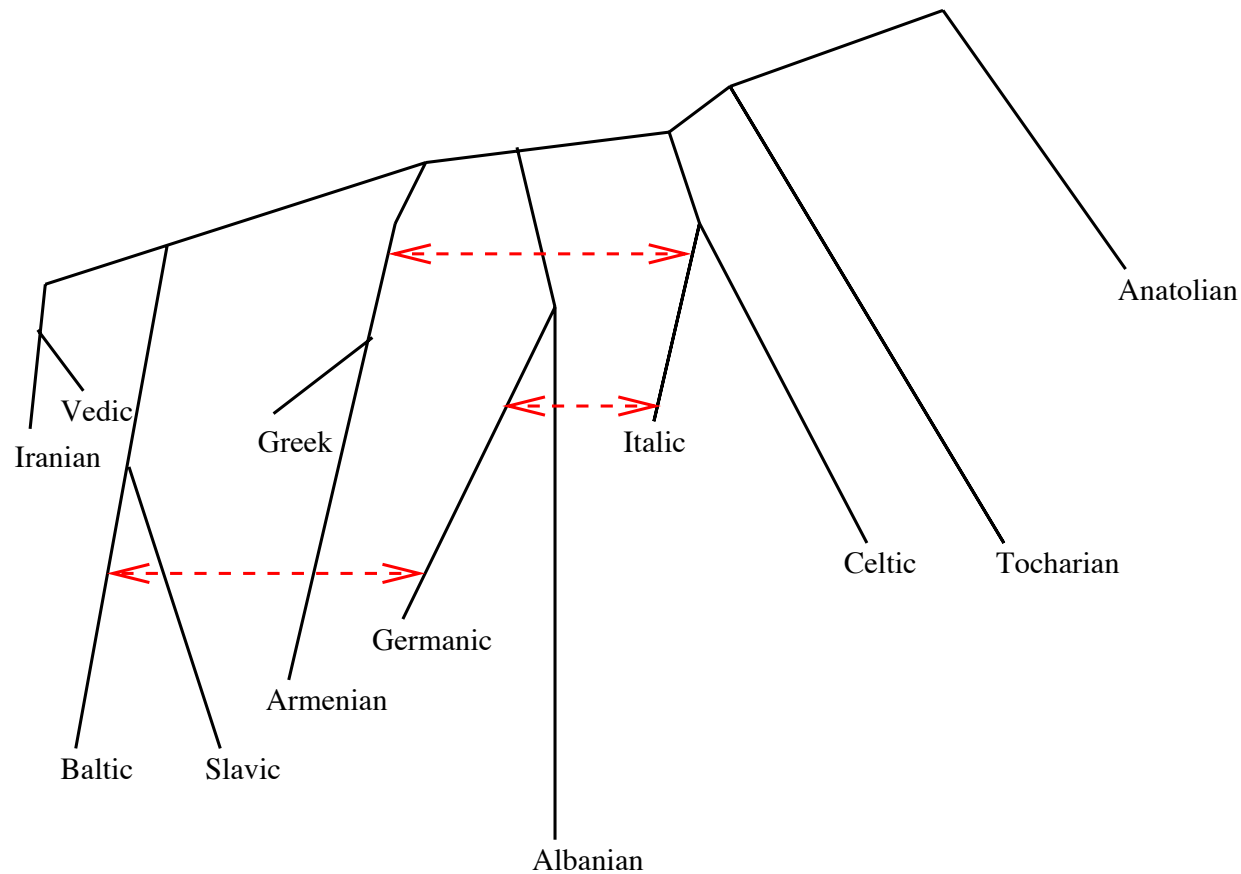
- Big genome sequencing projects just produce data -- so what?
- Evolutionary history relates all organisms and genes, and helps us understand and predict
 - interactions between genes (genetic networks)
 - drug design
 - predicting functions of genes
 - influenza vaccine development
 - origins and spread of disease
 - origins and migrations of humans

Possible Indo-European tree (Ringe, Warnow and Taylor 2000)



“Perfect Phylogenetic Network” for IE

Nakhleh et al., Language 2005



Research Opportunities

- There is lots of “low hanging fruit” in computational biology.
- Many of these problems have very clean formulations, and you don’t need to know any biology (or linguistics) to work on them – especially in computational phylogenetics and multiple sequence alignment.
- It is possible that your course project would be publishable. Seriously!
- Our research group can help get you started.
- TACC is an amazing resource!

Some Research Problems

- Quartet-based species-tree estimation
- Multiple sequence alignment of fragmentary sequences
- Inferring language phylogenies using statistical models
- Evaluating impact of multiple sequence alignment error on biological inference

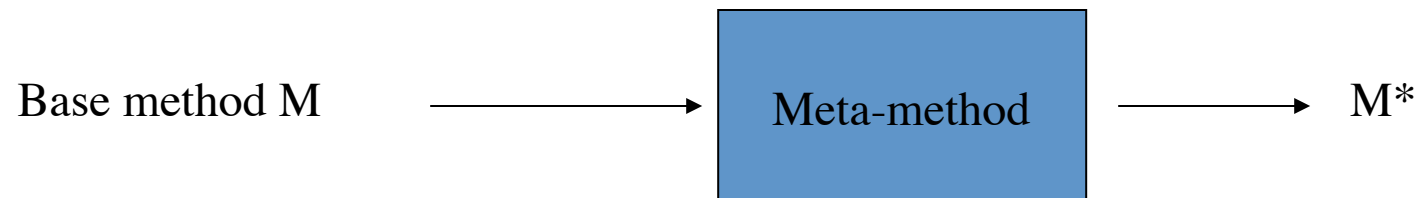
Evaluating impact of alignment error on biological inference

MSAs (multiple sequence alignments) are used to:

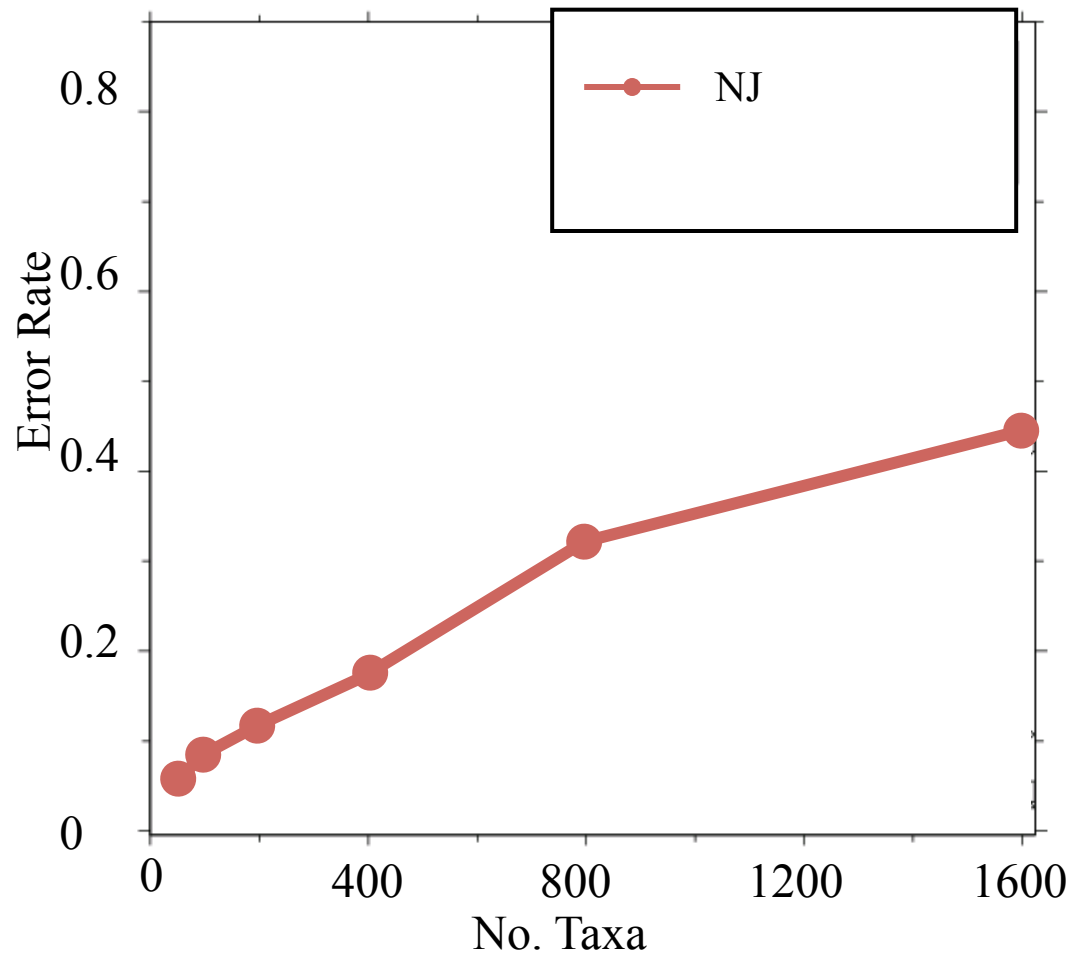
- Construct phylogenies
- Detect selection
- Estimate branch lengths
- Infer protein structure and function
- Estimate dates at internal nodes

Meta-Methods

- Meta-methods “boost” the performance of base methods (phylogeny reconstruction, alignment estimation, etc).

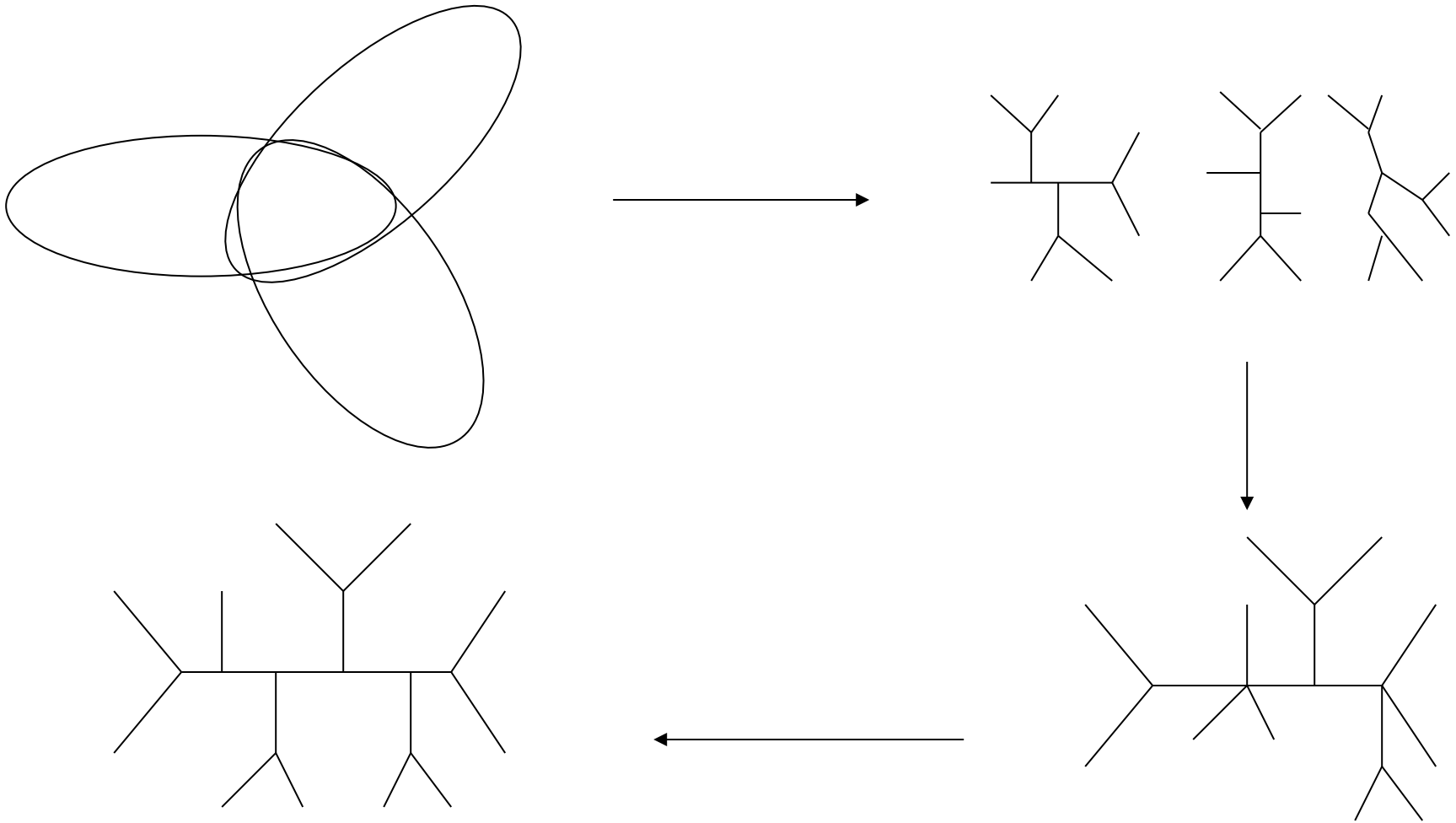


Neighbor joining has poor performance on large diameter trees [Nakhleh et al. ISMB 2001]



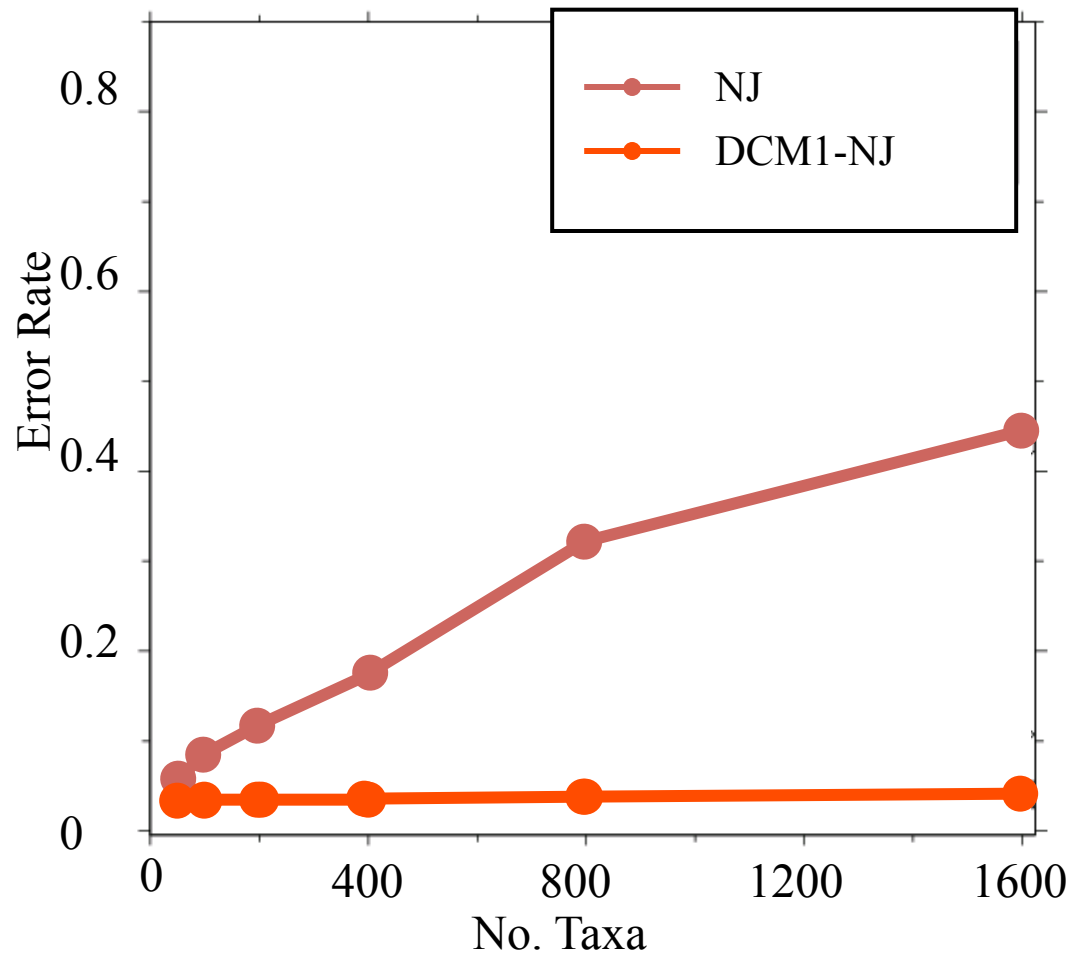
Theorem (Atteson):
Exponential sequence
length requirement for
Neighbor Joining!

Disk-Covering Methods (DCMs) (starting in 1998)



DCM1-boosting distance-based methods

[Nakhleh et al. ISMB 2001]

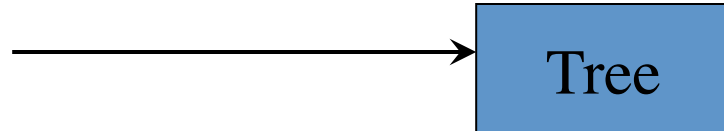


DCM1-boosting makes distance-based methods more accurate

Theoretical guarantees that DCM1-NJ converges to the true tree from **polynomial length** sequences

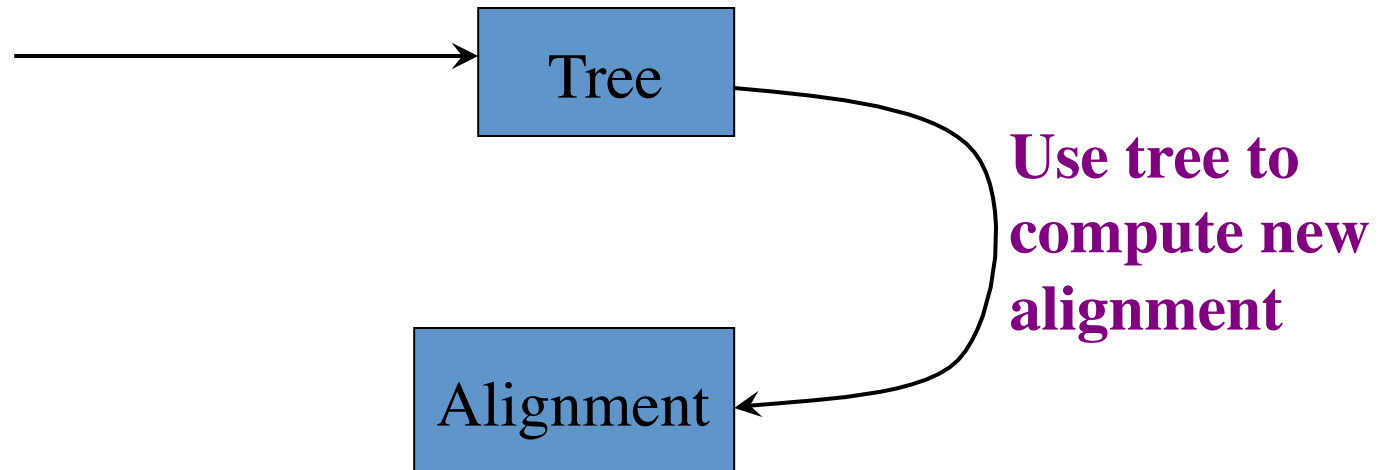
SATé Algorithm

Obtain initial alignment
and estimated ML tree



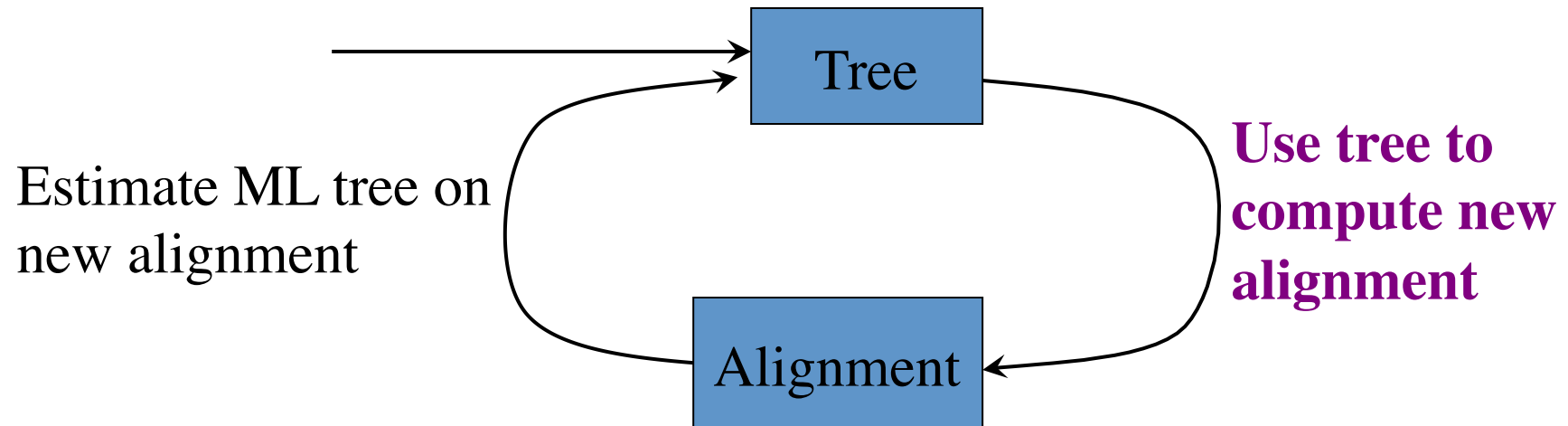
SATé Algorithm

Obtain initial alignment
and estimated ML tree



SATé Algorithm

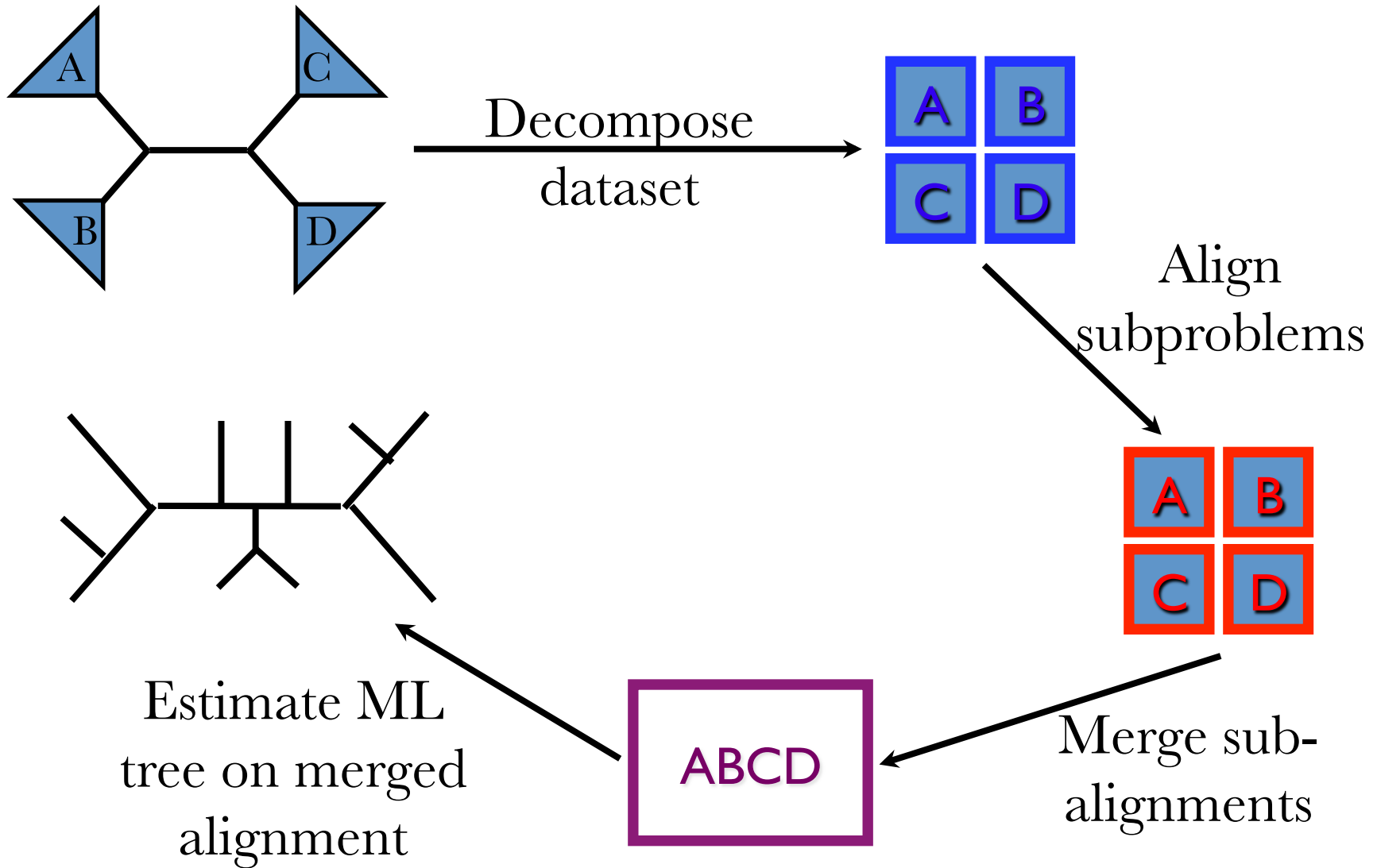
Obtain initial alignment
and estimated ML tree

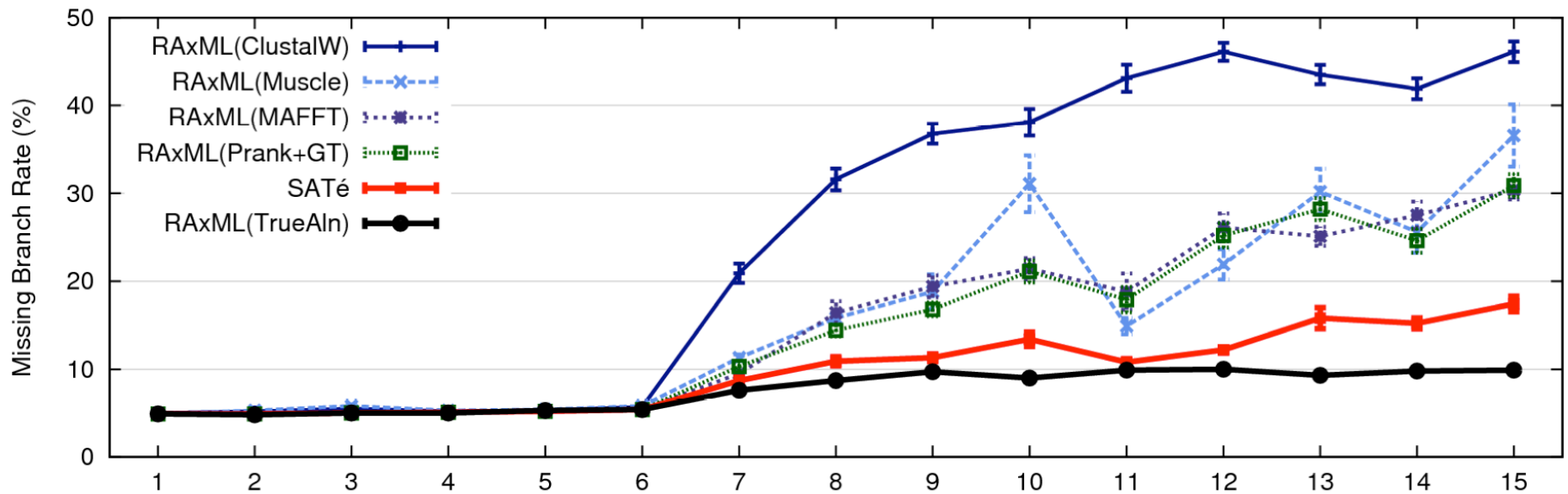


Estimate ML tree on
new alignment

**Use tree to
compute new
alignment**

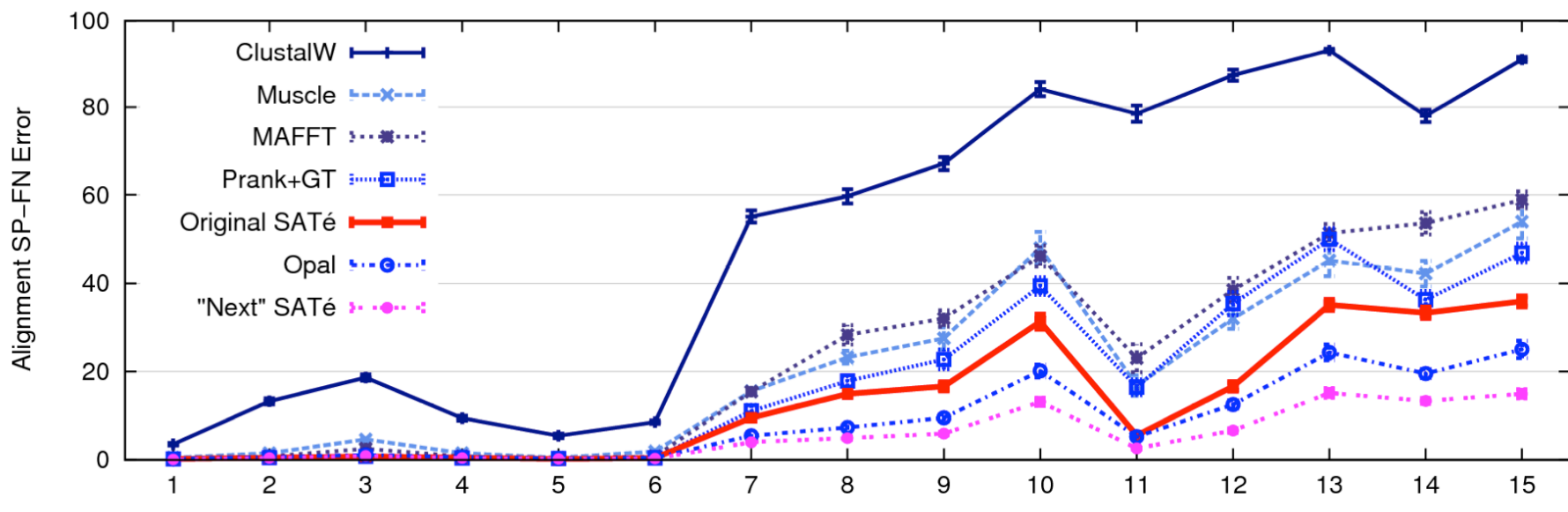
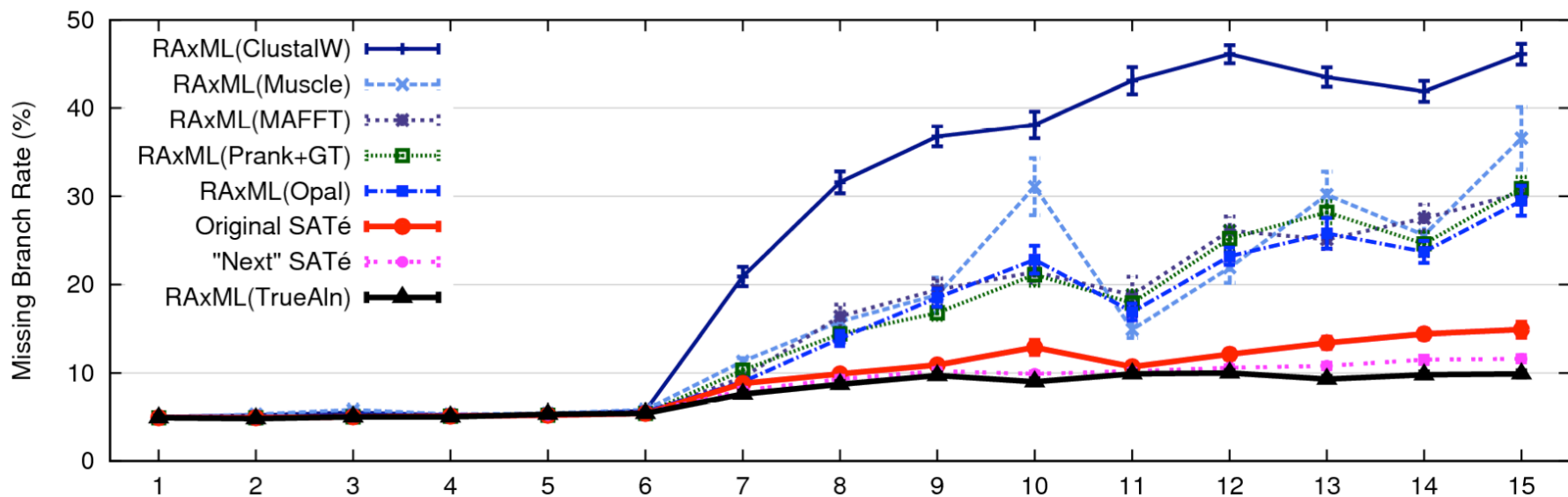
Re-aligning on a tree





1000 taxon models, ordered by difficulty

24 hour SATé analysis, on desktop machines
 (Similar improvements for biological datasets)



1000 taxon models ranked by difficulty

Algorithmic Strategies

- Divide-and-conquer
- “Bin-and-conquer”
- Iteration
- Hidden Markov Models
- Graph-theory

Course stuff

- How the course will be run
- Homeworks
- Grading Scheme
- Final Project
- Final Exam
- My office hours and location

Grading Scheme

- HW: 40%
- Class participation (including presentation of a research paper): 10%
- Project: 30% (research paper or survey article)
- Final: 20%

Course Structure

- Basics: Hidden Markov models, statistical inference, and computational complexity: 1 week
- Phylogeny estimation methods and models: 3 weeks
- Multiple sequence alignment methods and models: 2 weeks
- Phylogenomics: 2 weeks
- Genome Assembly: 1 week
- Metagenomics: 1 week
- Historical Linguistics: 1.5 weeks

Homework

Homework assignments will be of three types:

- pen and paper (doing calculations, proving theorems, etc.),
- programming (developing, implementing, and/or testing methods for computational biology or computational historical linguistics problems), and
- discussing published papers.

Final Project

- You are strongly encouraged to do a research project, but you can also do a survey paper on some topic relevant to the course material.
- In both cases, your project should be a paper (of about 15 pages) in a format and style appropriate for submission to a journal.
- Research projects can involve two students, but survey papers must be done by yourself.

Final Project Schedule

- Oct 2: One page proposal for final project due
- Oct 14: 2-3 page detailed proposal due
- Nov 13: First draft of final project (results and discussion)
- Nov 25: Final version (complete) due

I will consider one week extensions to the due date for the final version of the final project only for research projects – not for survey papers.

Final Exam

- Comprehensive, open book
- Could be take home, if the entire class agrees to this

Other Stuff

- My office is in Gates 4.510
- My office hours are Mondays, 12:30-1:30 PM (except for rare occasions)
- My email address is tandy@cs.utexas.edu