New Machine Learning Techniques for Alignment, Phylogeny, and Metagenomics

Tandy Warnow
Department of Computer Science
University of Texas

Big Data: Hype, or will we learn something new and important?

Is it just a matter of data management and sufficient parallelism? (or, Are TACC and a good programmer enough?)

NSF/NIH Joint Solicitation for Big Data 12-499

Big Data is not just about “size”, but about “complexity”:

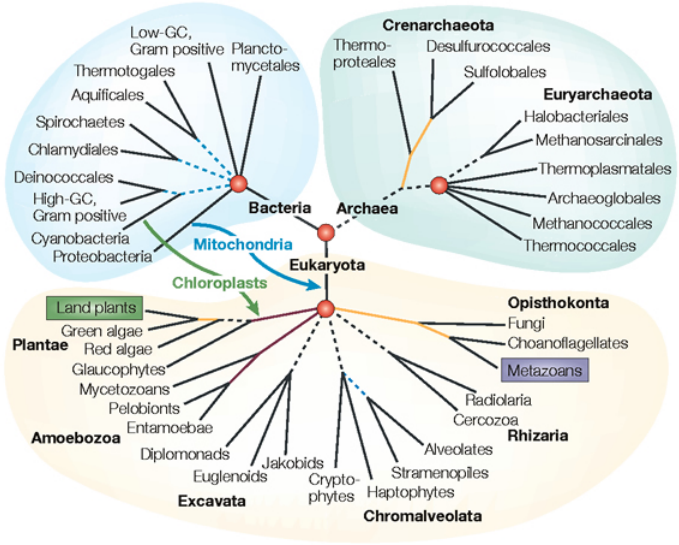
Error rates and types, heterogeneous data, missing data

Methods that have high accuracy on small datasets may have poor accuracy on “bigdata”.

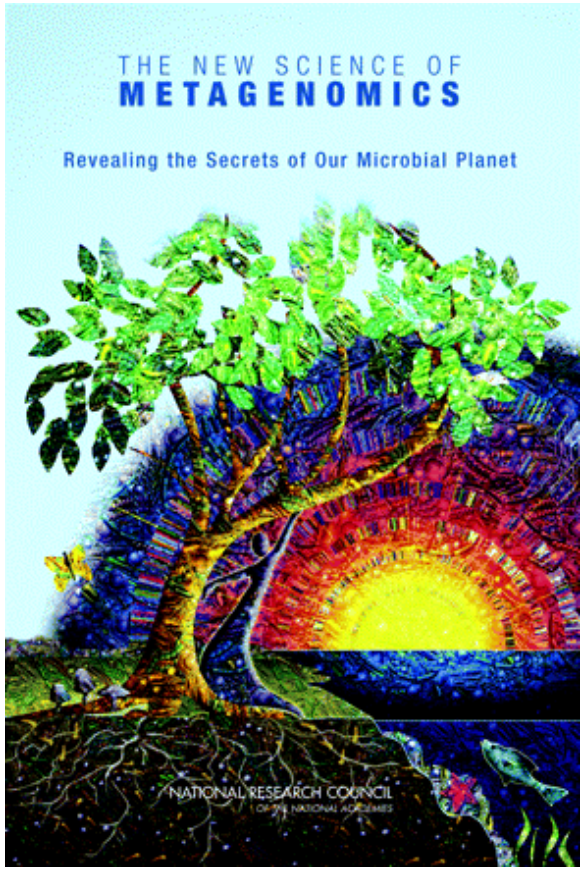
“Scaling up” of existing methods may not be not enough!

Requires fundamentally new methods from Computer Science, Mathematics, and Statistics.

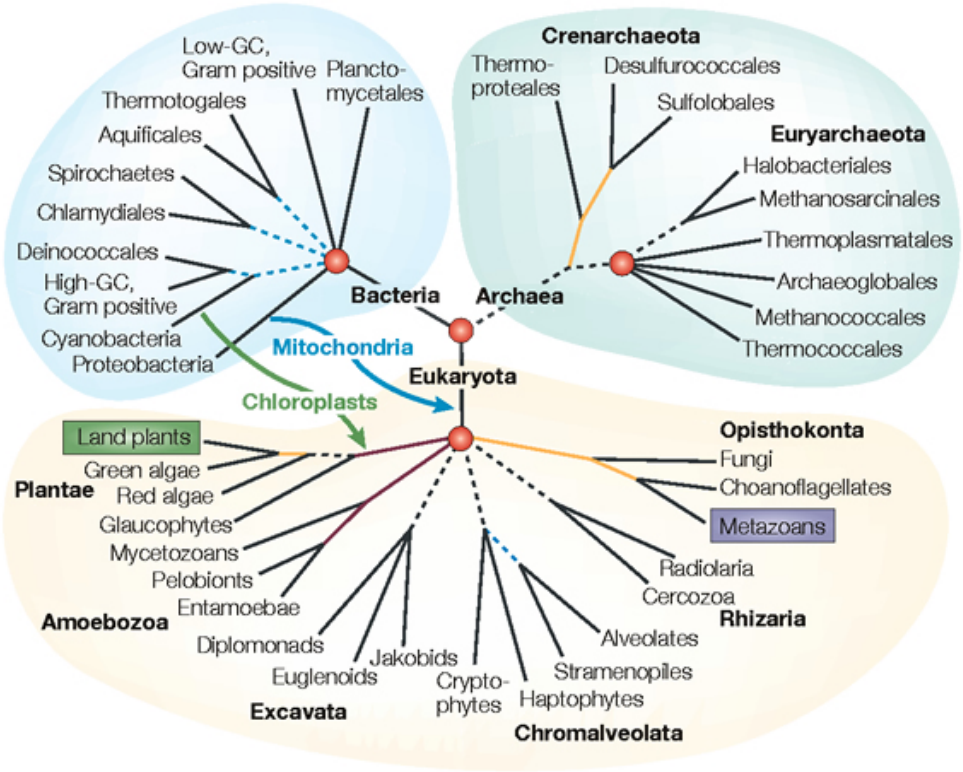
Computational Phylogenetics and Metagenomics



Nature Reviews | Genetics



NSF Program: **Assembling the Tree of Life**



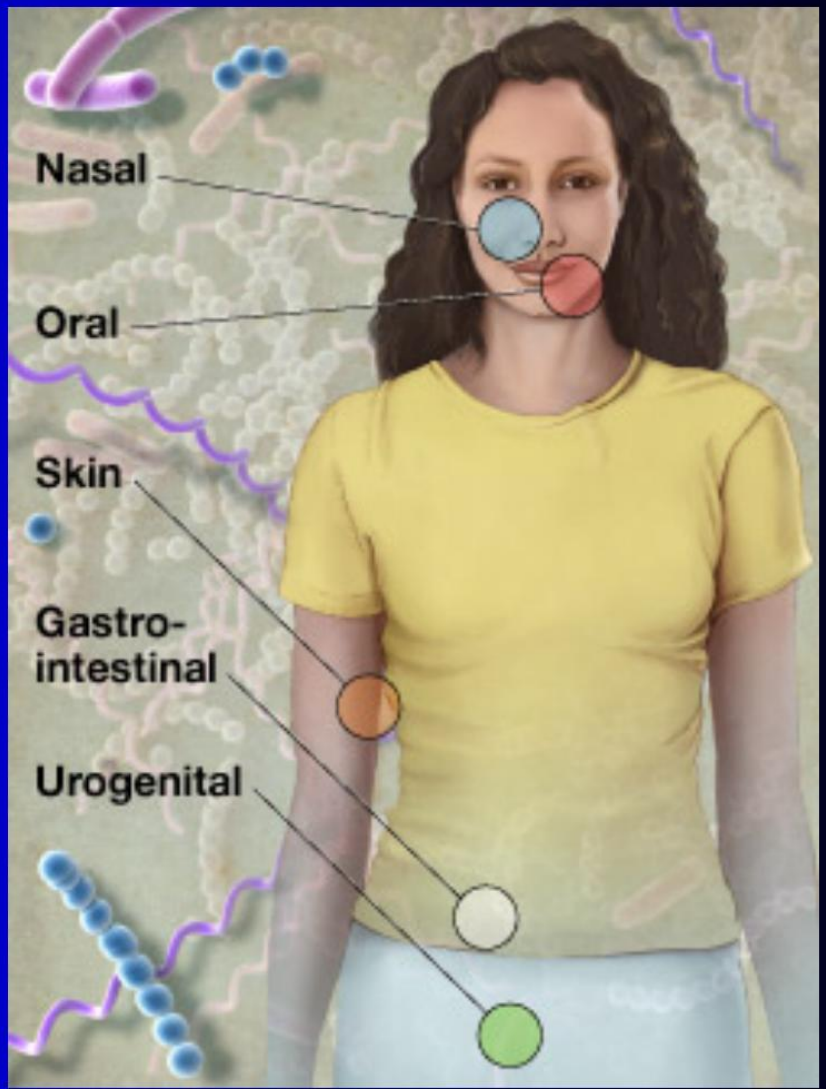
Nature Reviews | Genetics

Enormously hard computational challenges

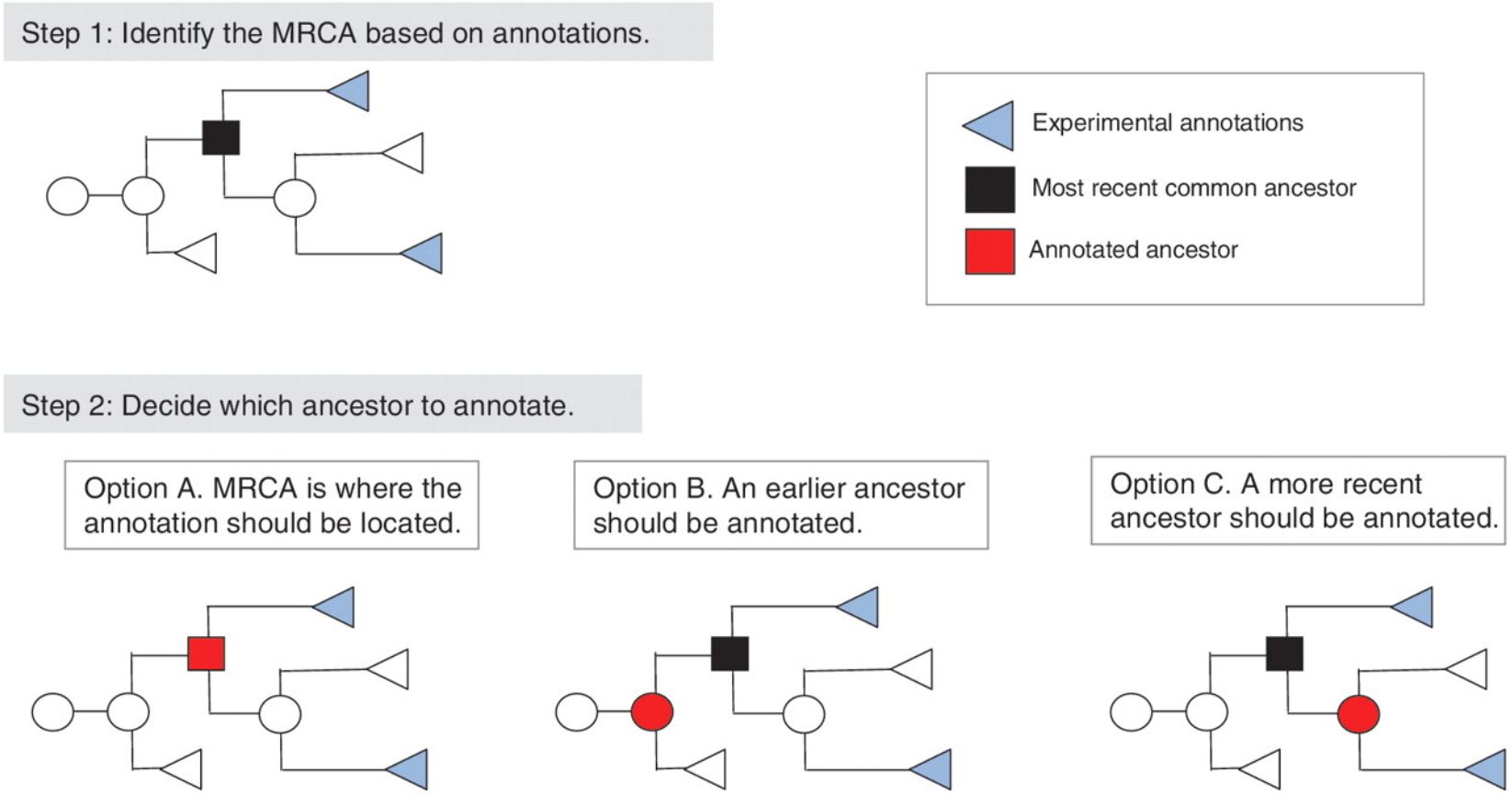
Current methods do not provide good accuracy on large datasets

The NIH Human Microbiome Project

25,000 human genes,
1,000,000 bacterial genes



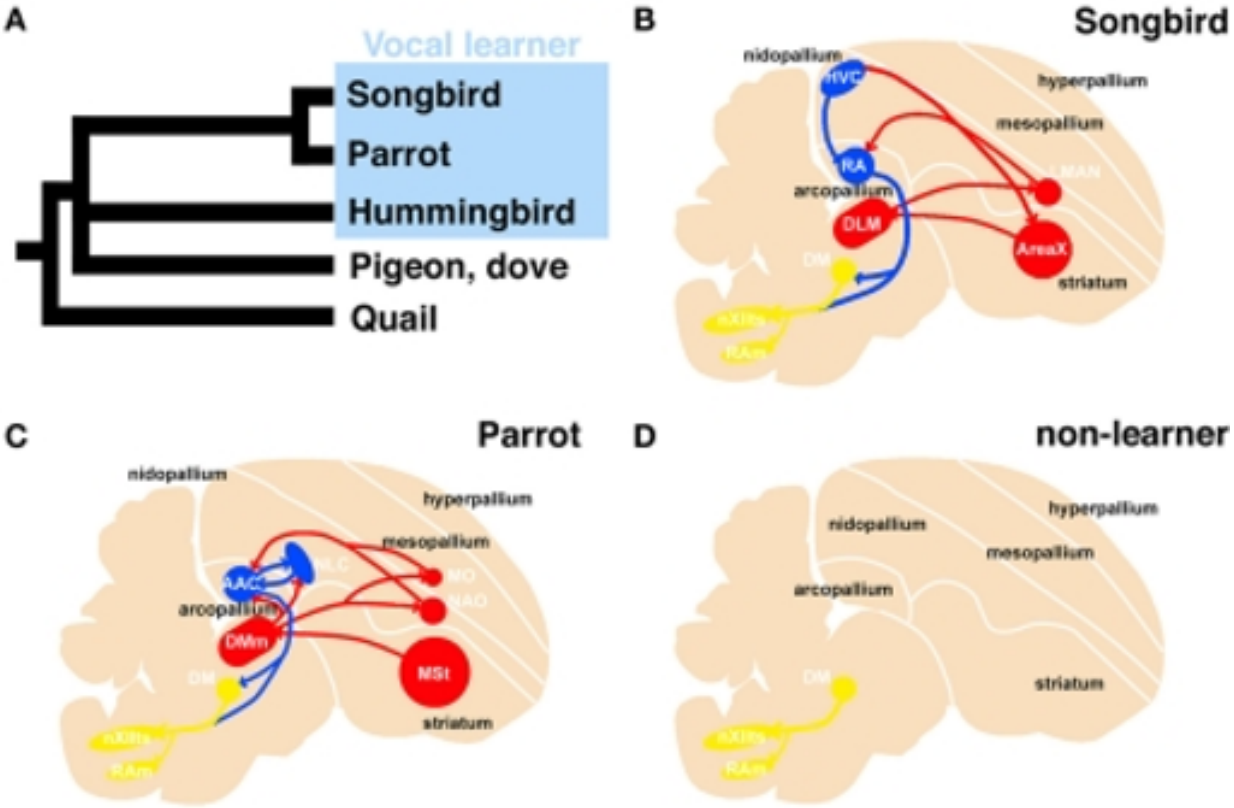
Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium



Gaudet P et al. Brief Bioinform 2011;bib.bbr042, Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium

Briefings in Bioinformatics

Evolution of Vocal Learning



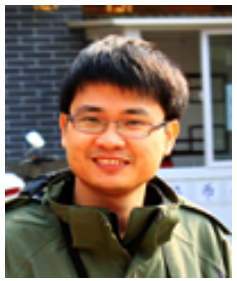
Erich Jarvis, HHMI

Avian Phylogenomics Project

Erich Jarvis, HHMI

MTP Gilbert, Copenhagen

Guojie Zhang, BGI



Siavash Mirarab, Tandy 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

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

Gane Ka-Shu Wong
U Alberta

Jim Leebens-Mack
U Georgia

Norm Wickett
Northwestern

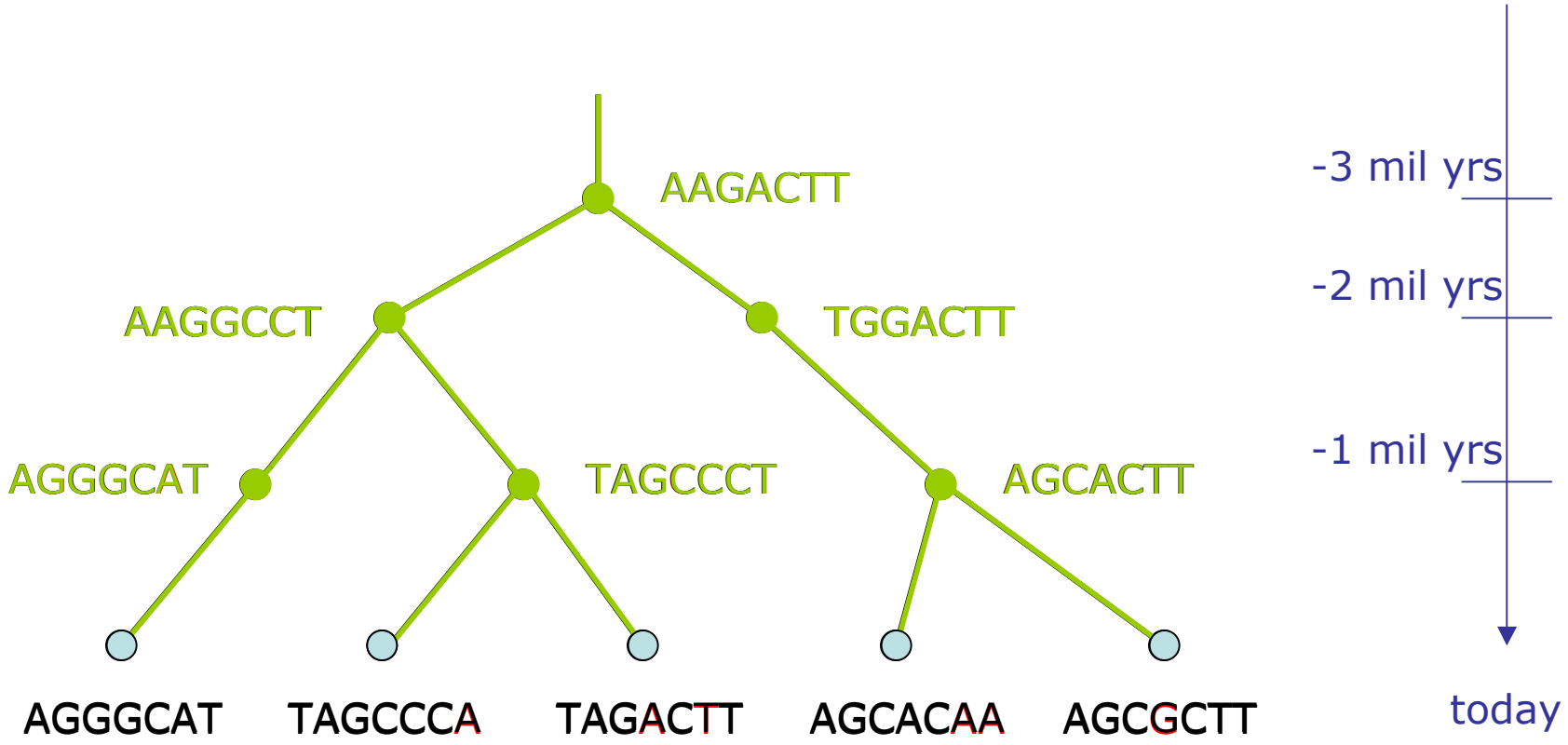
Naim Matasci
iPlant – U Arizona

Siavash Mirarab, Tandy Warnow, and Md. S. Bayzid at UT-Austin

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

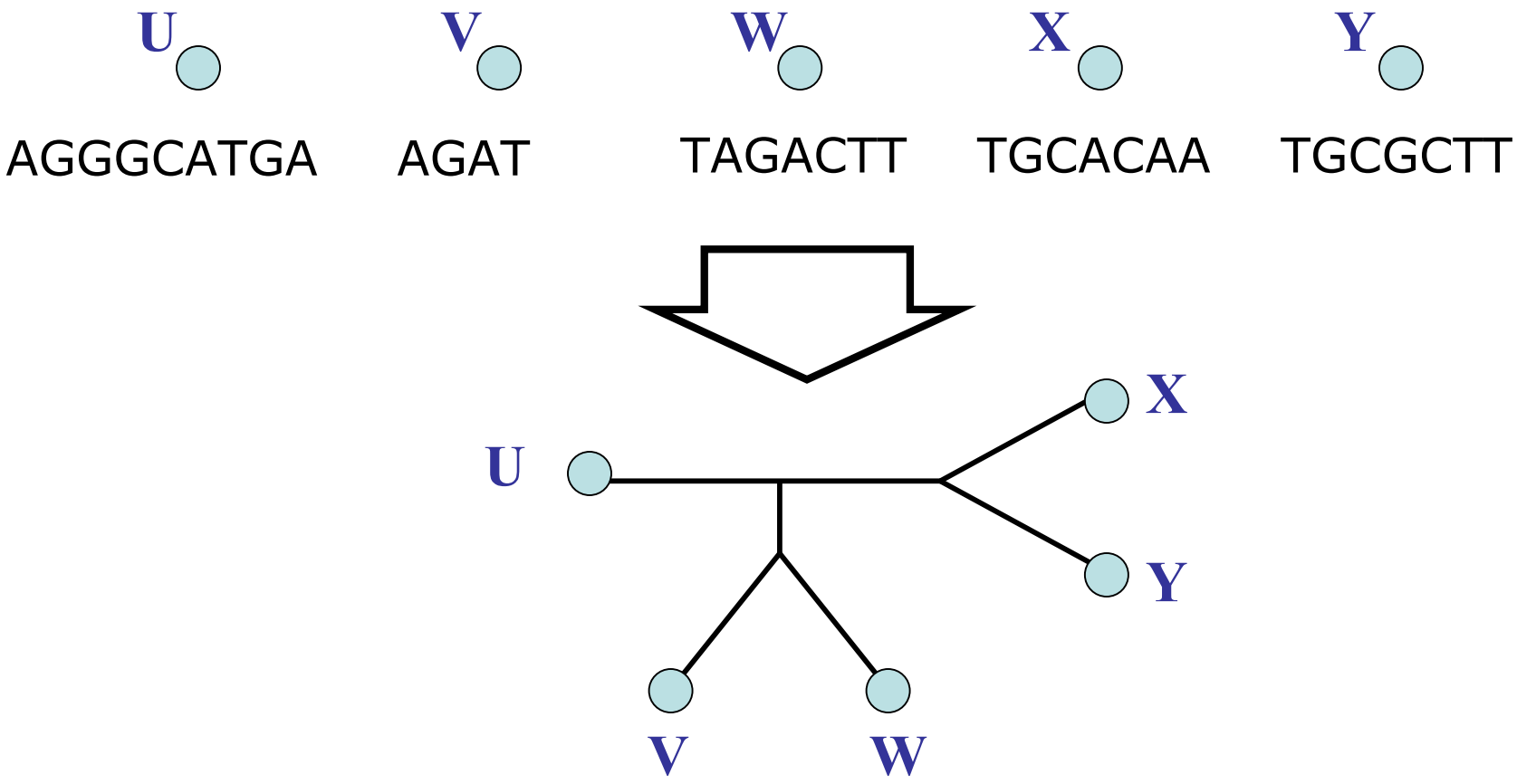
Challenge: estimating very large gene alignments and trees (100,000+ sequences)

DNA Sequence Evolution



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**


U
V
W
X
Y
AGGGCATGA
AGAT
TAGACTT
TGCACAA
TGCGCTT
X
U
Y
V
W

Input: unaligned sequences

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

Phase 1: Multiple Sequence Alignment

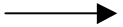
```
S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC
S3 = TAGCTGACCGC
S4 = TCACGACCGACA
```



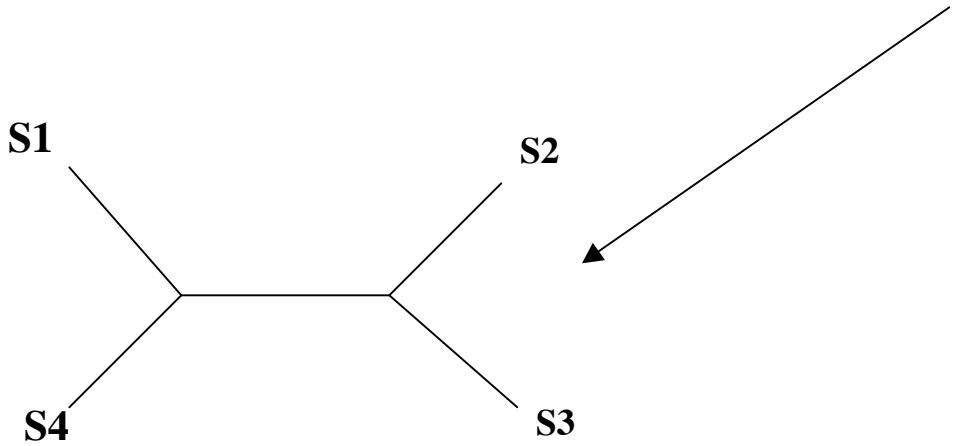
```
S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAC--GACCGC--
S3 = TAG-CT-------GACCGC--
S4 = --------TCAC--GACCGACA
```

Phase 2: Construct tree

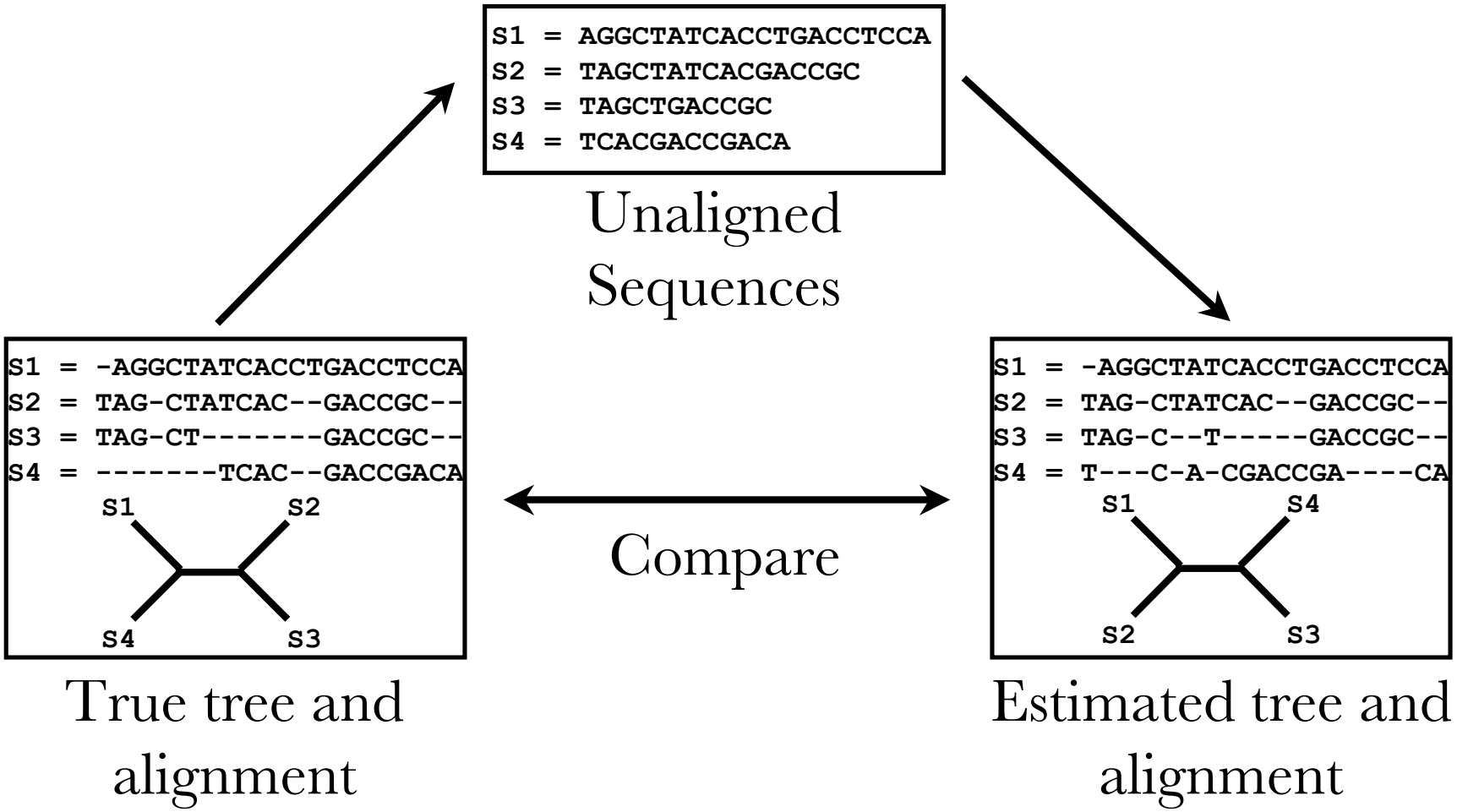
```
S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC
S3 = TAGCTGACCGC
S4 = TCACGACCGACA
```

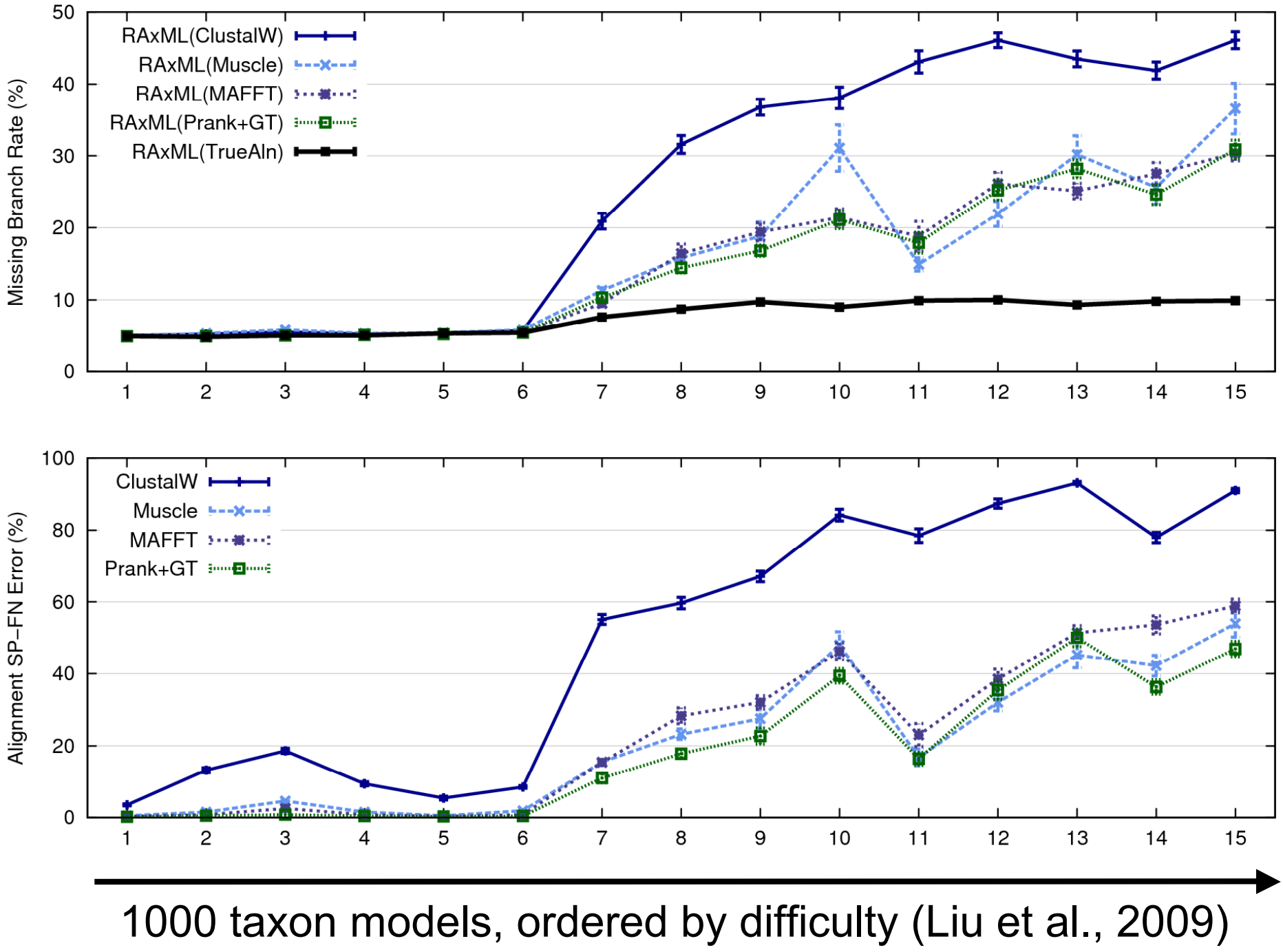


```
S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAC--GACCGC--
S3 = TAG-CT-------GACCGC--
S4 = -------TCAC--GACCGACA
```



Simulation Studies




RAxML(ClustalW)
RAxML(Muscle)
RAxML(MAFFT)
RAxML(Prank+GT)
RAxML(TrueAln)
Missing Branch Rate (%)
0
10
20
30
40
50
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
ClustalW
Muscle
MAFFT
Prank+GT
Alignment SP–FN Error (%)
0
20
40
60
80
100
1000 taxon models, ordered by difficulty (Liu et al., 2009)

Major Challenges:
large datasets, fragmentary sequences

- **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**: Few methods can run on large datasets, and alignment accuracy is generally poor for large datasets with high rates of evolution.
- **Metagenomic** analyses: methods for species classification of short reads have *poor sensitivity*. Efficient high throughput is necessary (millions of reads).

These methods are also impacted by ***fragmentary data.***

Today's Talk

- SATé: Simultaneous Alignment and Tree Estimation (Liu et al., Science 2009 and Systematic Biology, 2012)
- TIPP: Taxon Identification using SEPP (in preparation, Nguyen et al.)
- UPP: Ultra-large alignment using SEPP (in preparation, Nguyen et al.) - time permitting

Key algorithm design techniques:

- **Statistical estimation** methods (Hidden Markov Models, Maximum Likelihood)
- Data-driven **divide-and-conquer** to improve the accuracy and scalability of the "base" method
- **Iteration**

Part 1: SATé

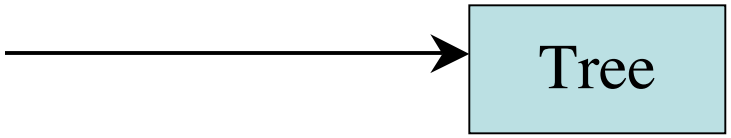
Liu, Nelesen, Raghavan, Linder, and Warnow, *Science* 2009, pp. 1561-1564.

Liu et al., *Syst Biol* 2012, 61(1): 90-106.

Public software distribution (open source) through the University of Kansas, in use, world-wide.

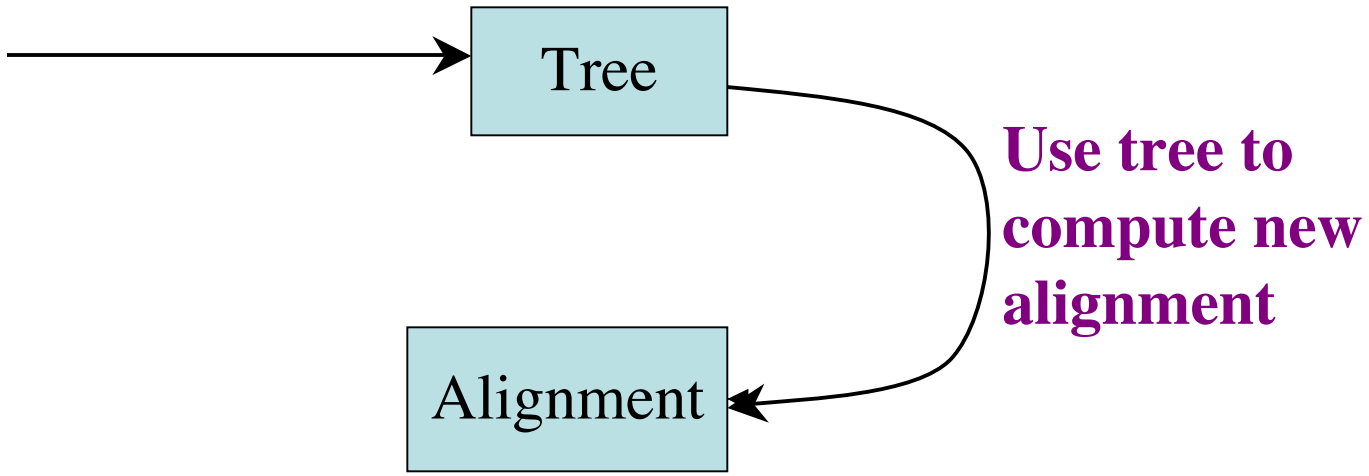
SATé Algorithm

Obtain initial alignment
and estimated ML tree



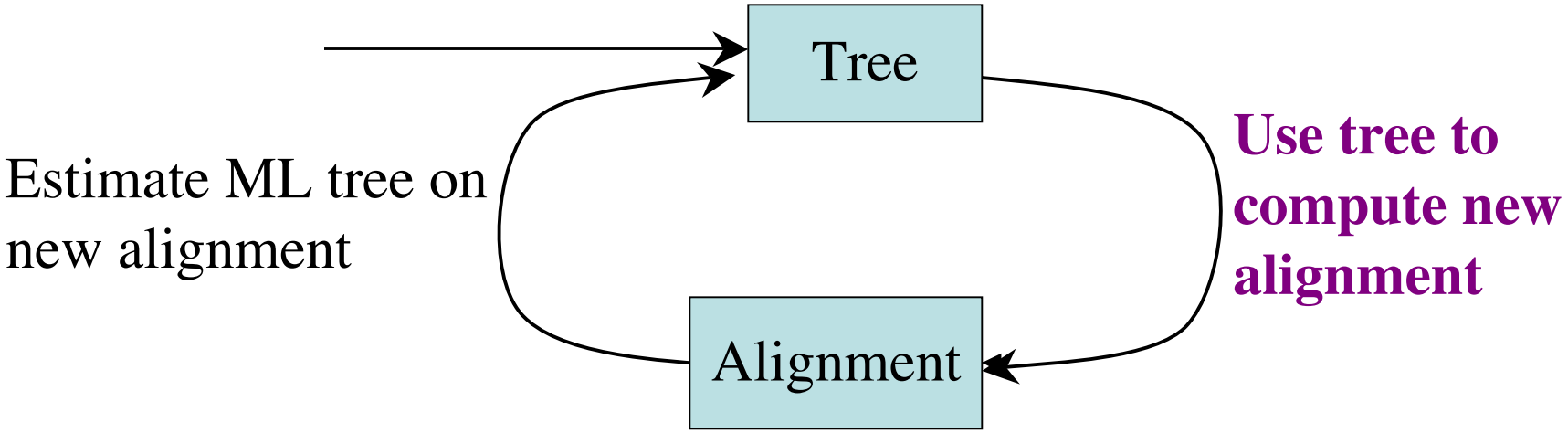
SATé Algorithm

Obtain initial alignment
and estimated ML tree

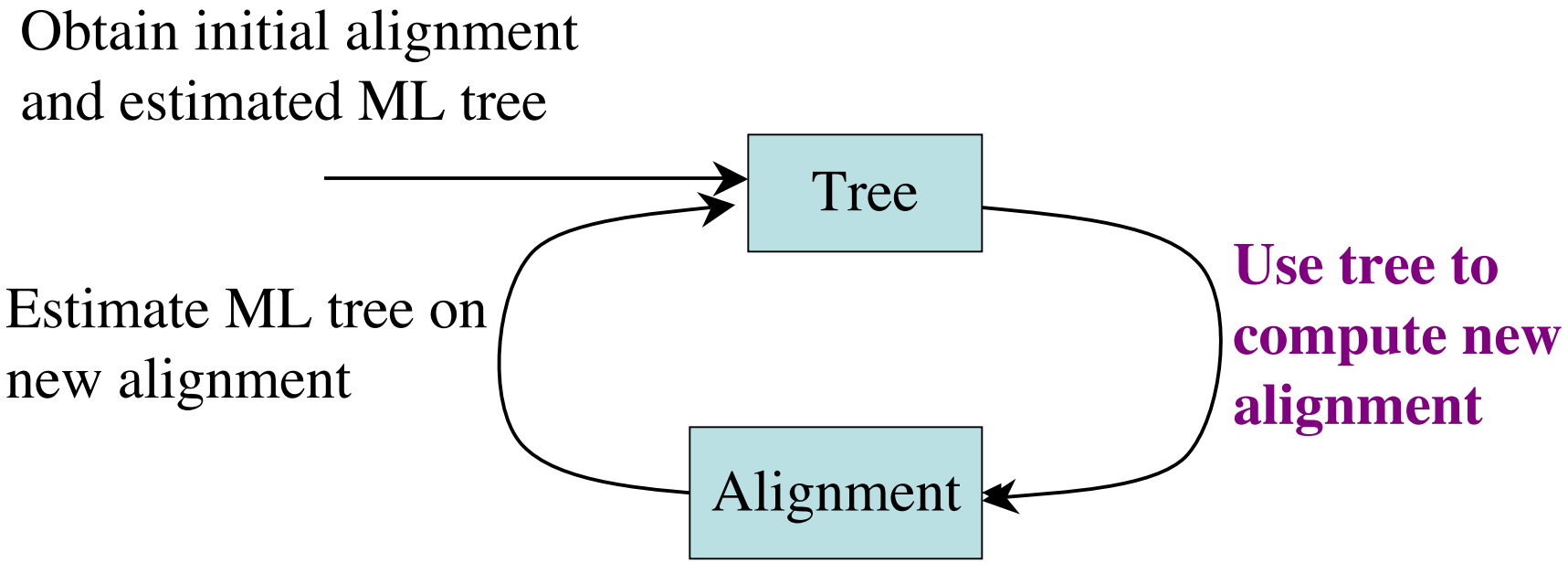


SATé Algorithm

Obtain initial alignment
and estimated ML tree



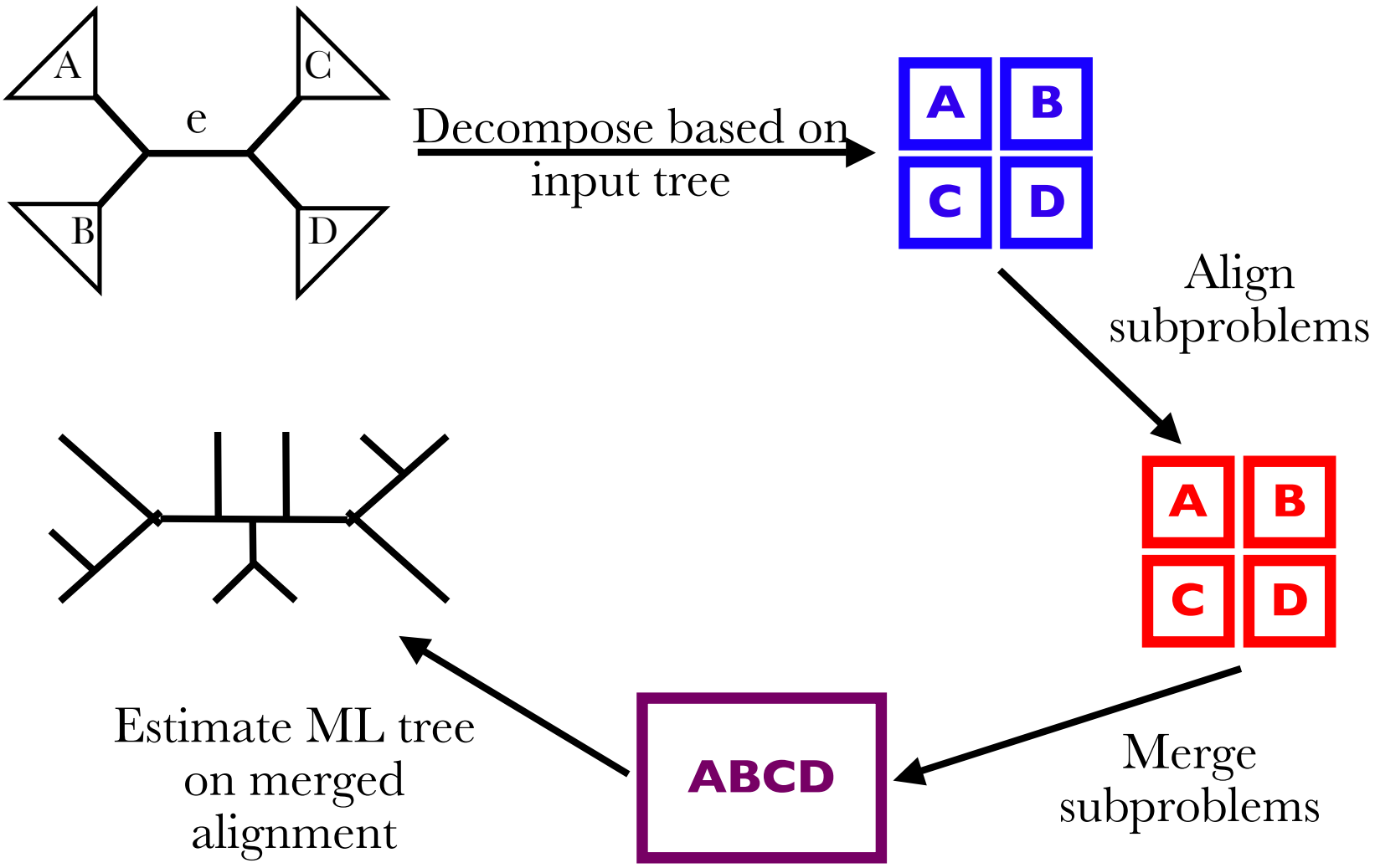
SATé Algorithm

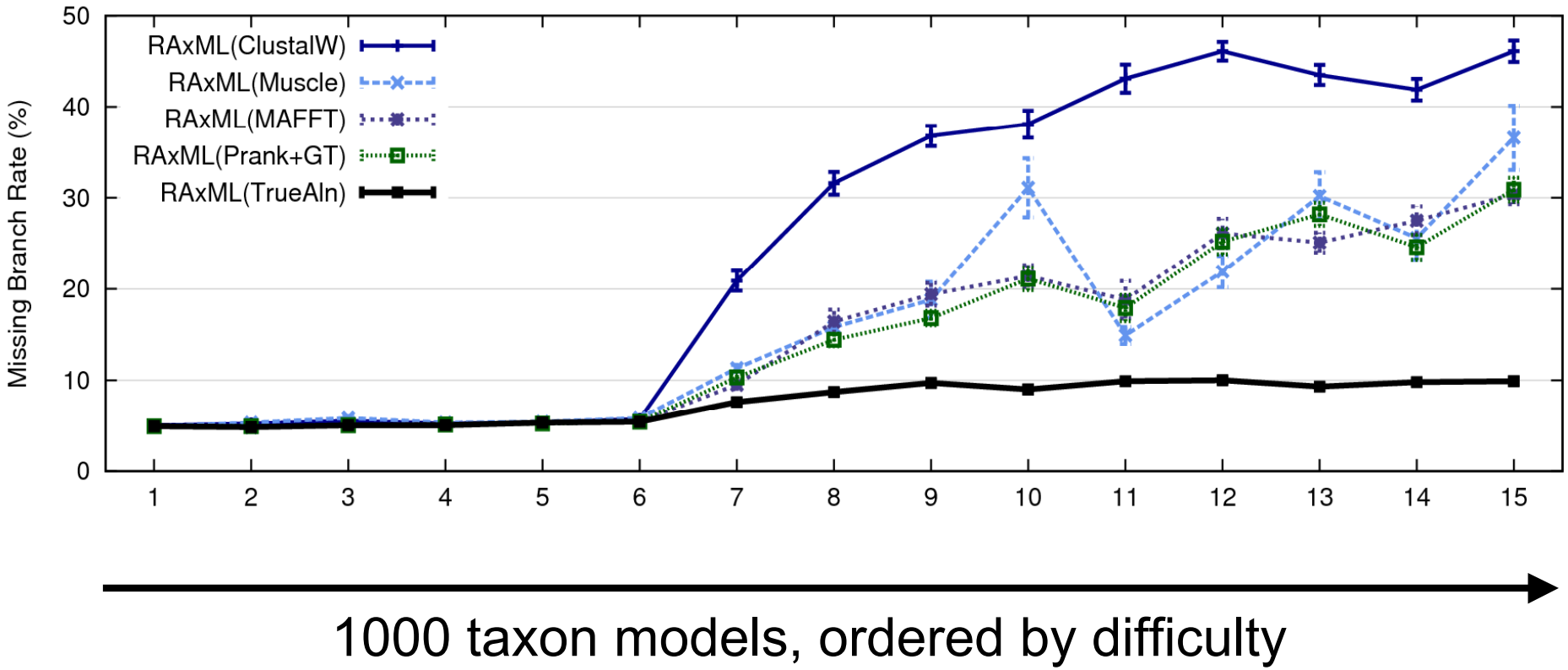


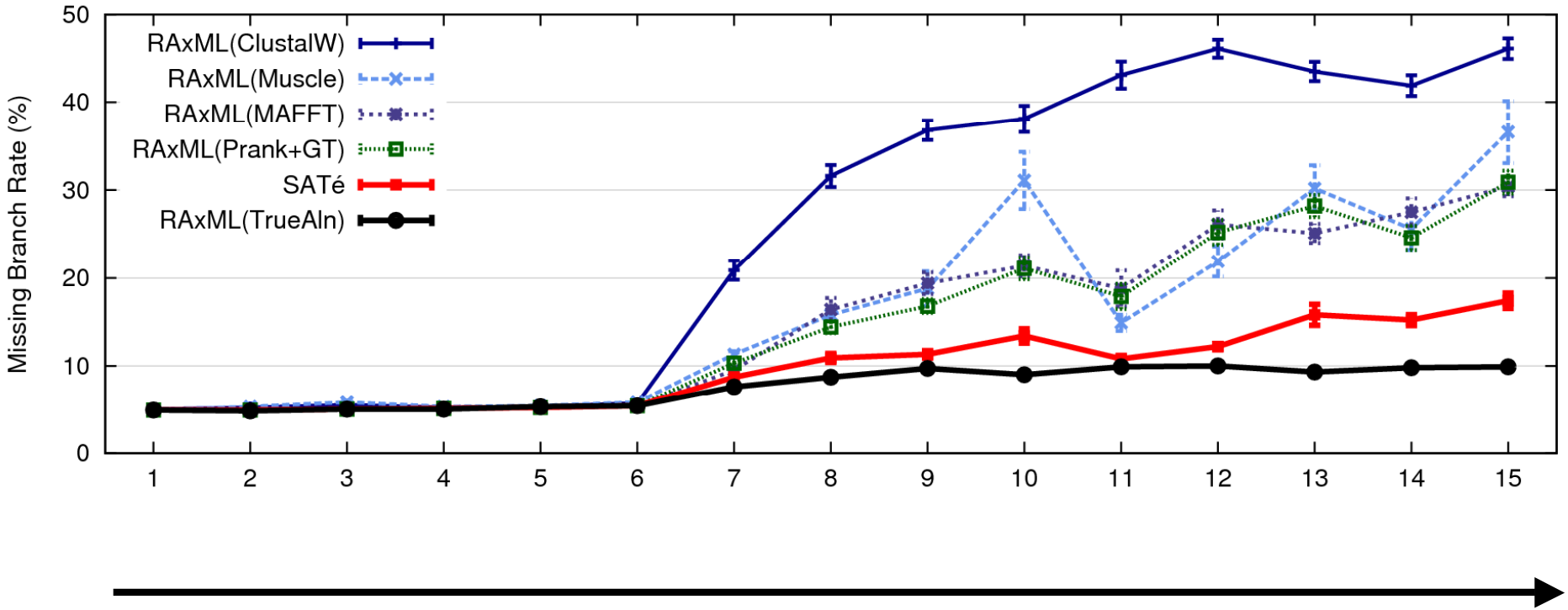
If new alignment/tree pair has worse ML score, realign using a different decomposition

Repeat until termination condition (typically, 24 hours)

One SATé iteration (really 32 subsets)



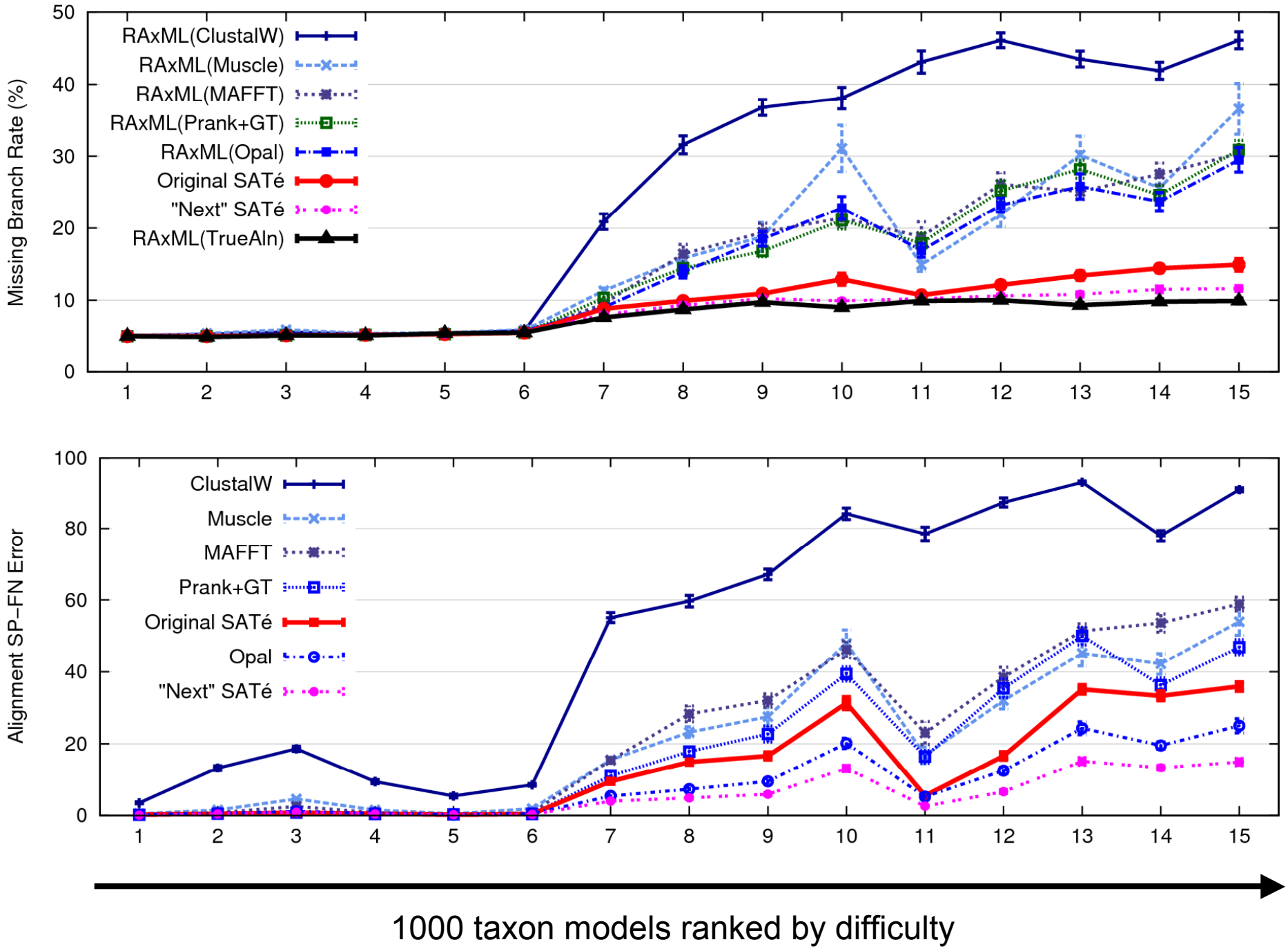

RAxML(ClustalW)
RAxML(Muscle)
RAxML(MAFFT)
RAxML(Prank+GT)
RAxML(TrueAln)
Missing Branch Rate (%)
0
10
20
30
40
50
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
1000 taxon models, ordered by difficulty

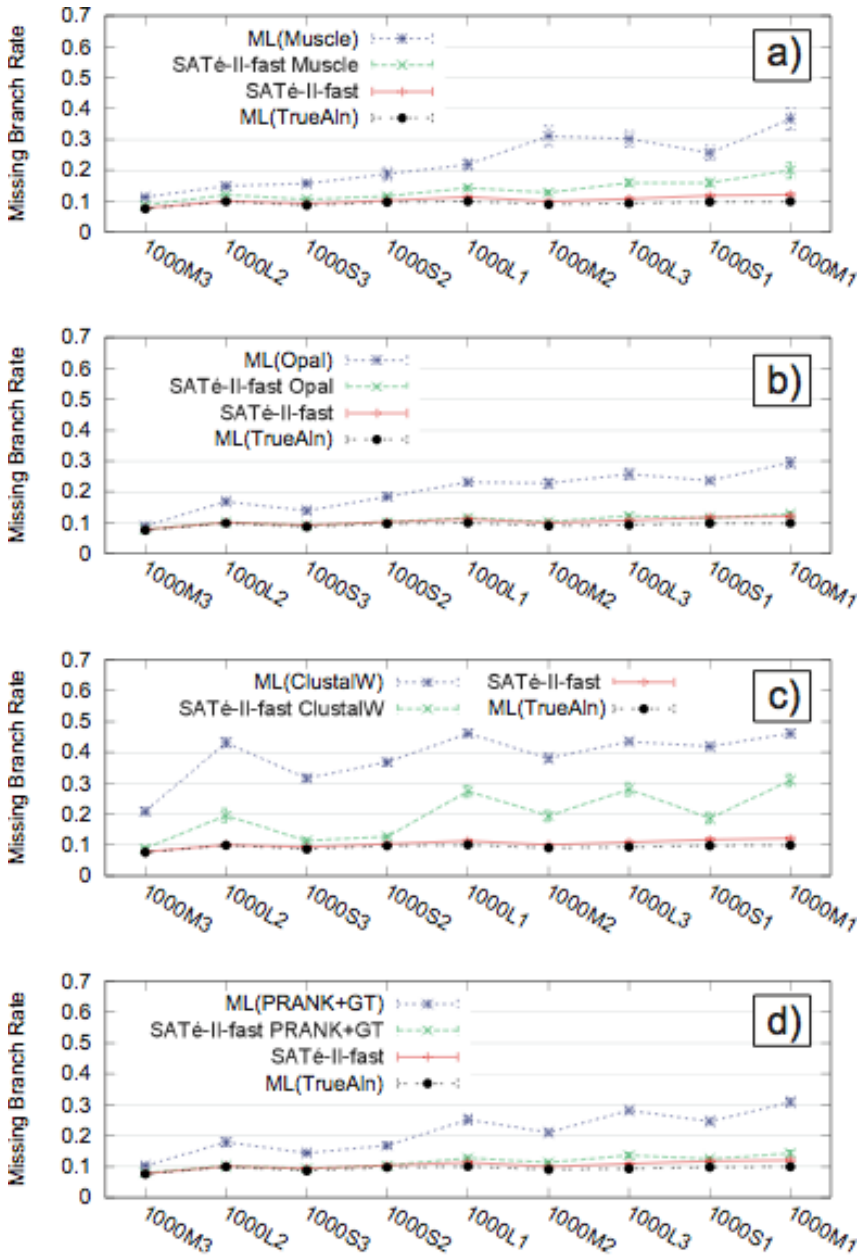


1000 taxon models, ordered by difficulty

24 hour SATé analysis, on desktop machines

(Similar improvements for biological datasets)


Missing Branch Rate (%)
RAxML(ClustalW)
RAxML(Muscle)
RAxML(MAFFT)
RAxML(Prank+GT)
RAxML(Opal)
Original SATé
"Next" SATé
RAxML(TrueAln)
Alignment SP–FN Error
ClustalW
Muscle
MAFFT
Prank+GT
Original SATé
Opal
"Next" SATé
1000 taxon models ranked by difficulty



SATe can be used to “boost” other alignment methods.

In the Avian project, we used SATe+PRANK to identify erroneously annotated portions of the sequences.

Final gene sequence alignments were computed using SATe+MAFFT (largely due to computational issues).

SATé

- SATé is more accurate than standard methods on large datasets with high rates of evolution (both biological and simulated data), and has been used on both proteins and nucleotides.

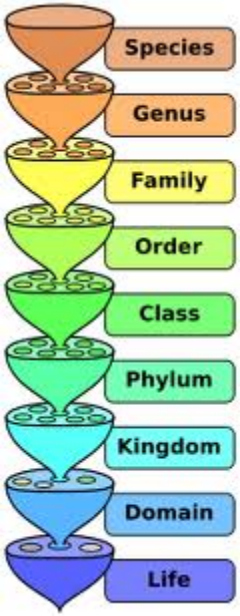
- The current implementation has been tested on datasets with up to 50,000 sequences.

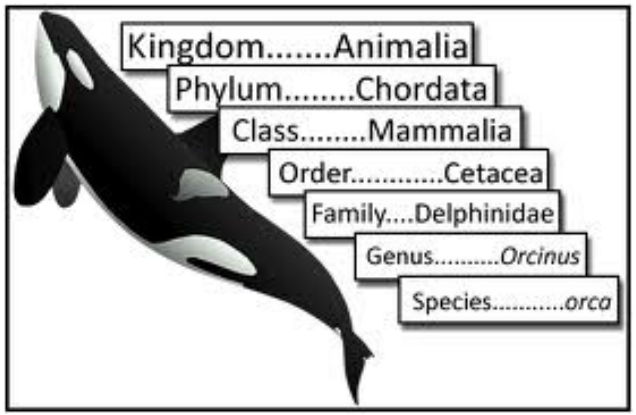
Open-source downloadable program at:

http://phylo.bio.ku.edu/software/sate/sate.html

Part II: Metagenomic Taxon Identification

Objective: classify short reads in a metagenomic sample





Key technique: SEPP

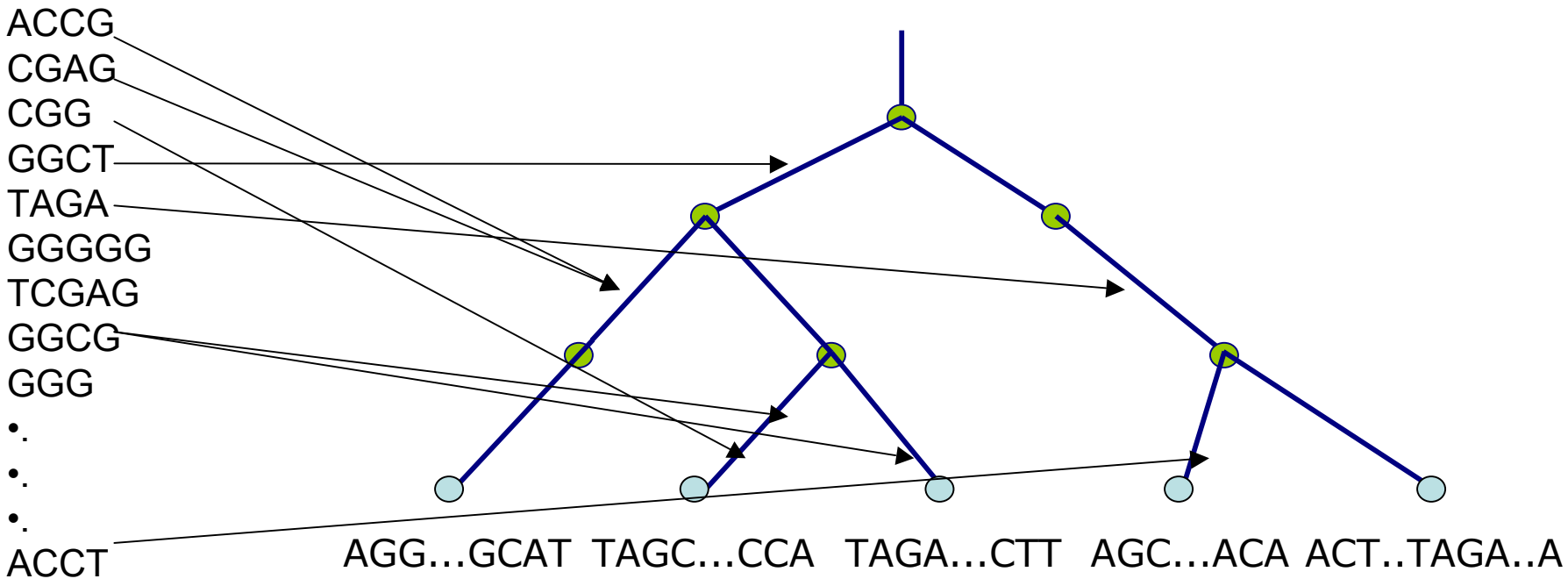
- SEPP: SATé-enabled Phylogenetic Placement, by Mirarab, Nguyen, and Warnow

- Pacific Symposium on Biocomputing, 2012 (special session on the Human Microbiome)

Phylogenetic Placement

Fragmentary sequences from some gene

Full-length sequences for same gene, and an alignment and a tree



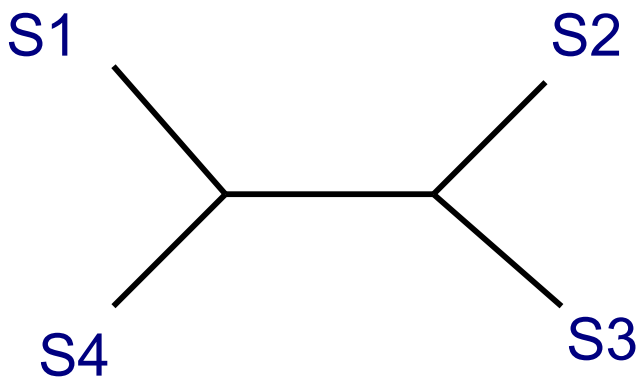
Phylogenetic Placement

Step 1: Align each query sequence to backbone alignment

Step 2: Place each query sequence into backbone tree, using extended alignment

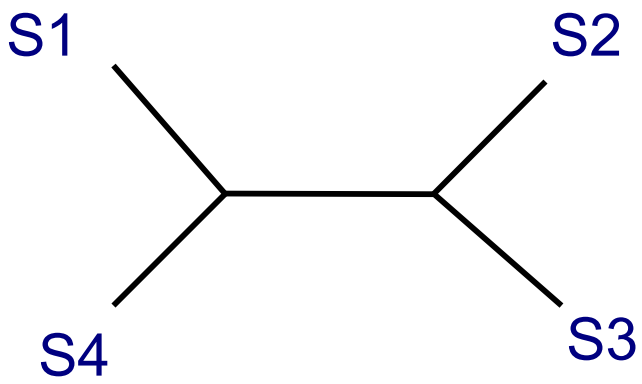
Align Sequence

```
S1   = -AGGCTATCACCTGACCTCCA-AA
S2   = TAG-CTATCAC--GACCGC--GCA
S3   = TAG-CT-------GACCGC--GCT
S4   = TAC----TCAC--GACCGACAGCT
Q1   = TAAAAC
```



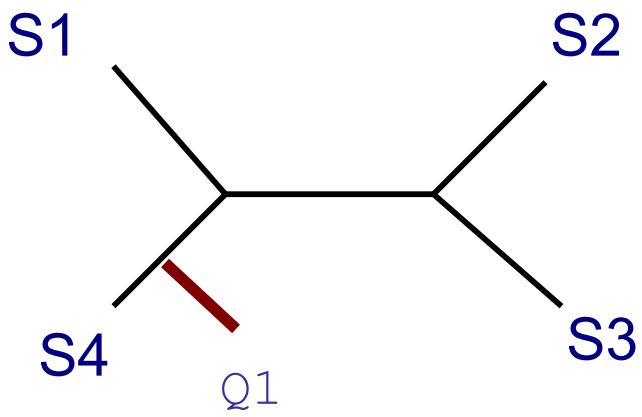
Align Sequence

```
S1   = -AGGCTATCACCTGACCTCCA-AA
S2   = TAG-CTATCAC--GACCGC--GCA
S3   = TAG-CT-------GACCGC--GCT
S4   = TAC----TCAC--GACCGACAGCT
Q1   = -------T-A--AAAC--------
```



Place Sequence

```
S1  = -AGGCTATCACCTGACCTCCA-AA
S2  = TAG-CTATCAC--GACCGC--GCA
S3  = TAG-CT-------GACCGC--GCT
S4  = TAC----TCAC--GACCGACAGCT
Q1  = -------T-A--AAAC--------
```

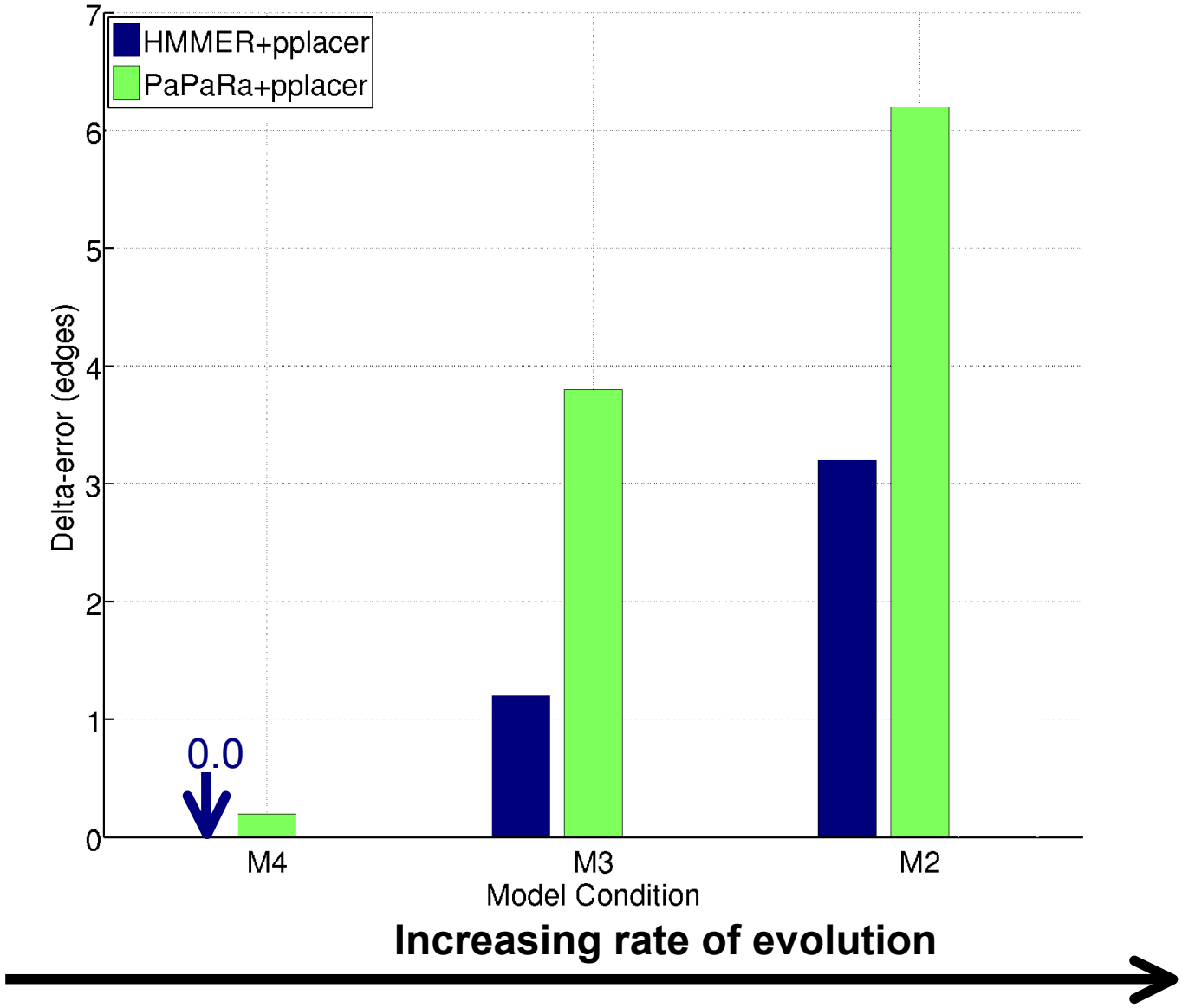


Phylogenetic Placement

- Align each query sequence to backbone alignment
  - HMMALIGN (Eddy, Bioinformatics 1998)
  - PaPaRa (Berger and Stamatakis, Bioinformatics 2011)
- Place each query sequence into backbone tree
  - Pplacer (Matsen et al., BMC Bioinformatics, 2011)
  - EPA (Berger and Stamatakis, Systematic Biology 2011)

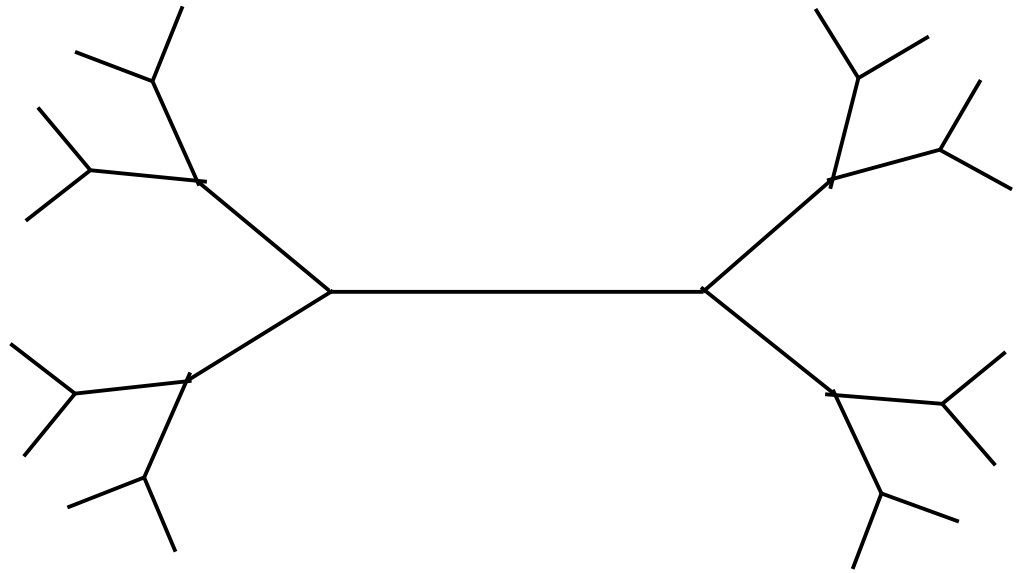
Note: pplacer and EPA use maximum likelihood, and are reported to have the same accuracy.

HMMER vs. PaPaRa

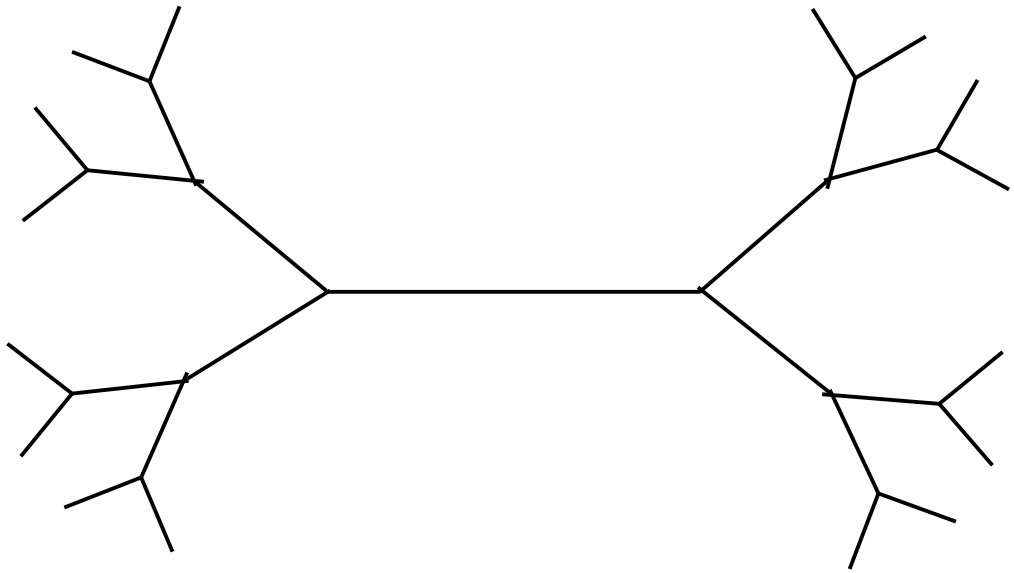


HMMER+pplacer:

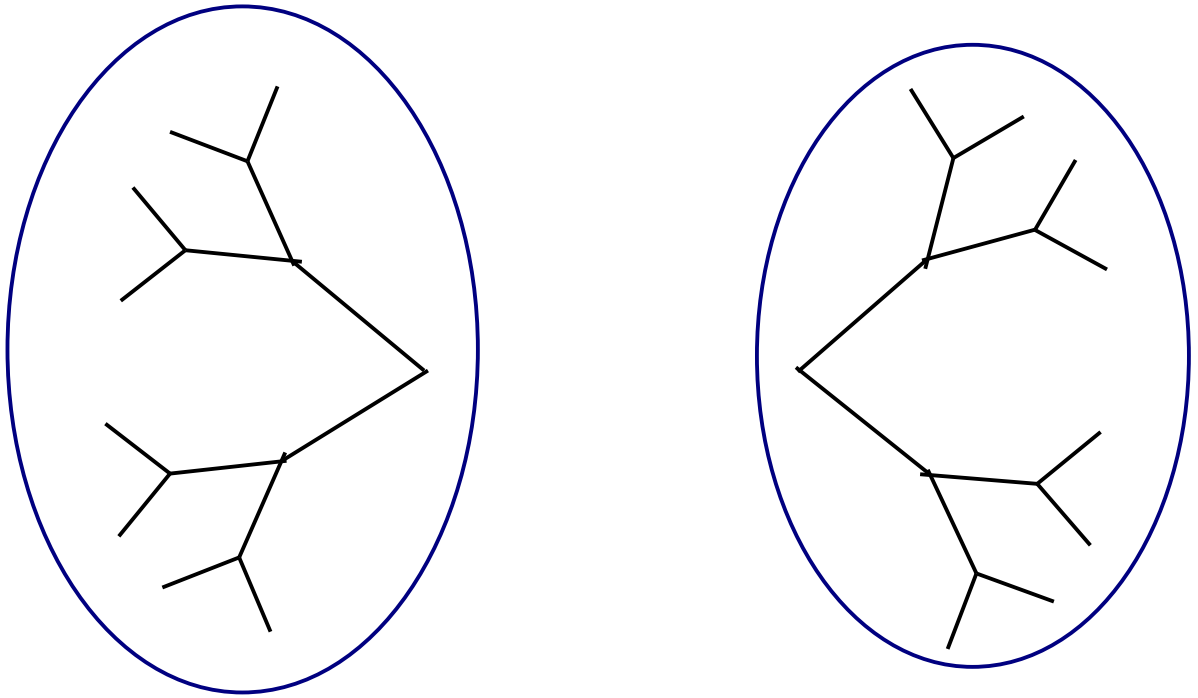
1) build one HMM for the entire alignment
2) Align fragment to the HMM, and insert into alignment
3) Insert fragment into tree to optimize likelihood



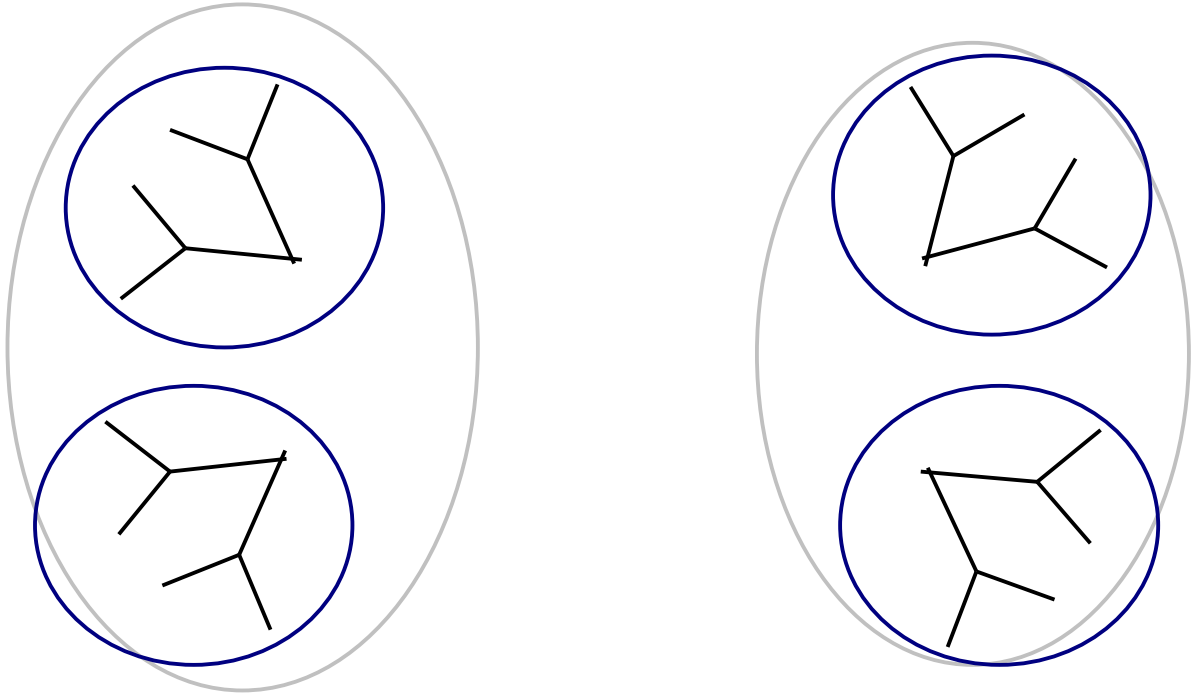
One Hidden Markov Model for the entire alignment?



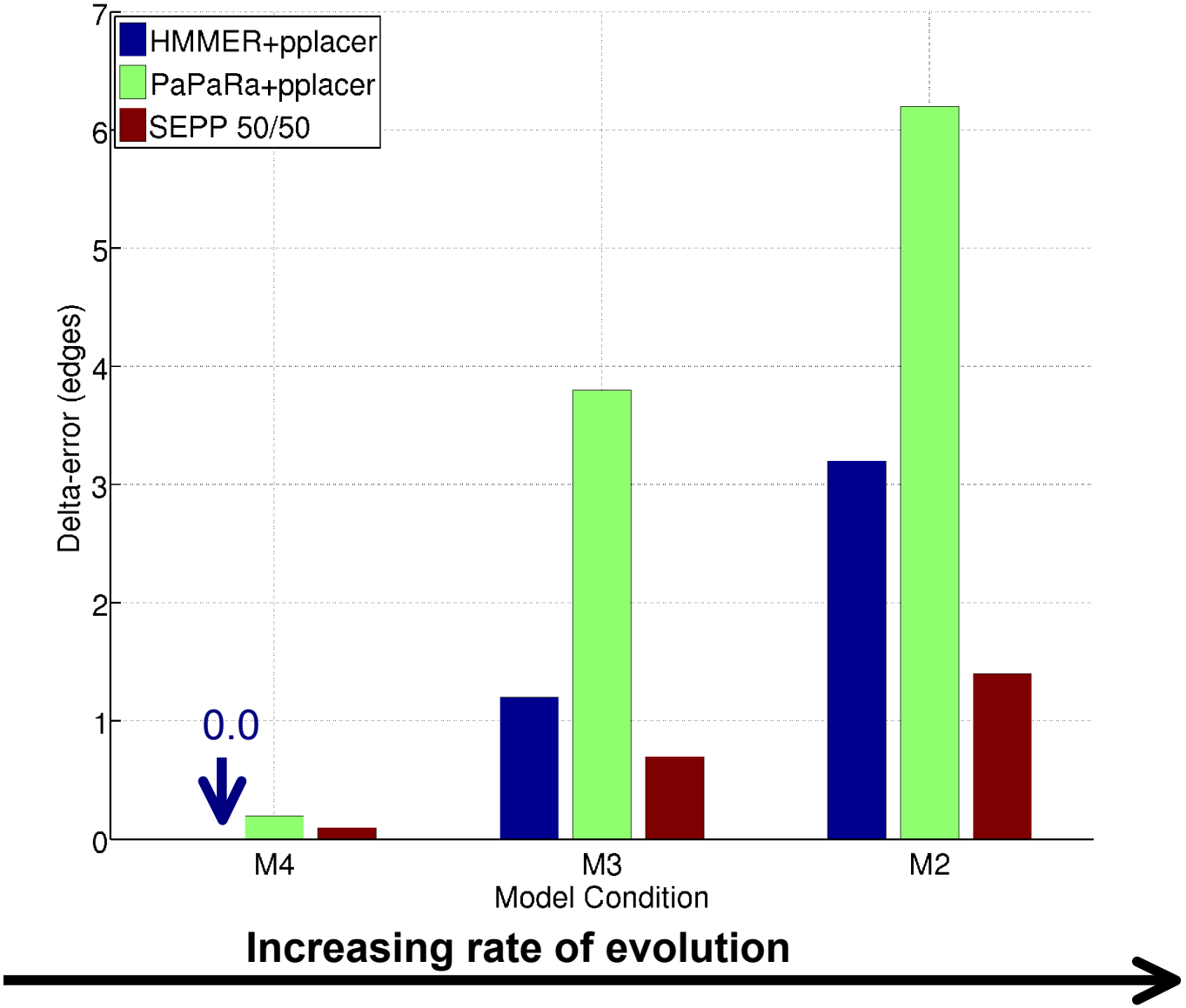
Or 2 HMMs?



Or 4 HMMs?



SEPP(10%), based on ~10 HMMs



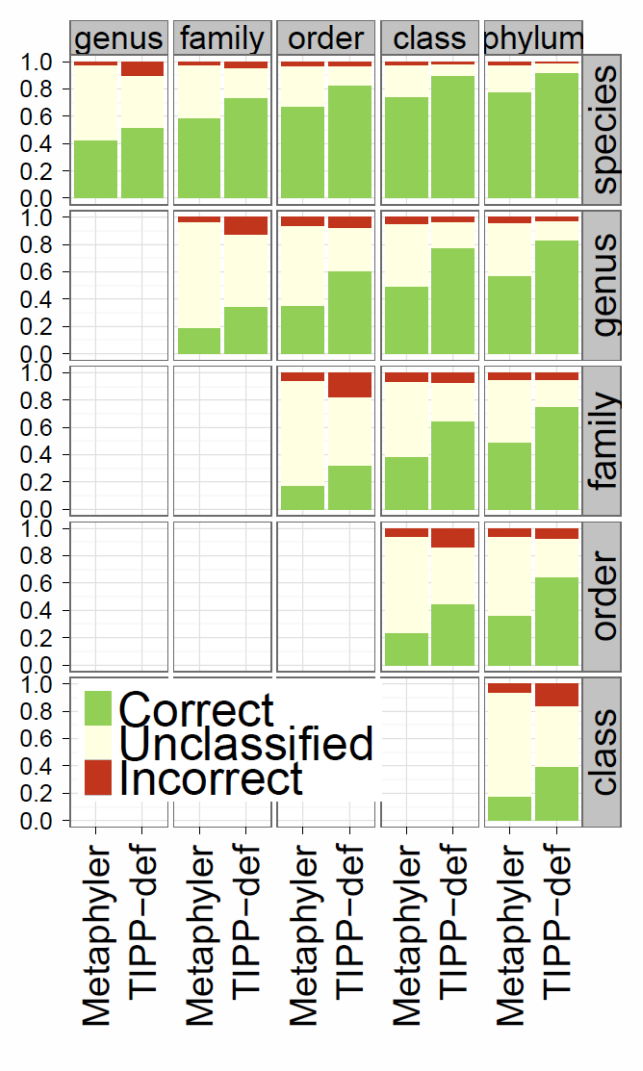
TIPP: SEPP + statistics

SEPP has high recall but low precision (classifies almost everything)

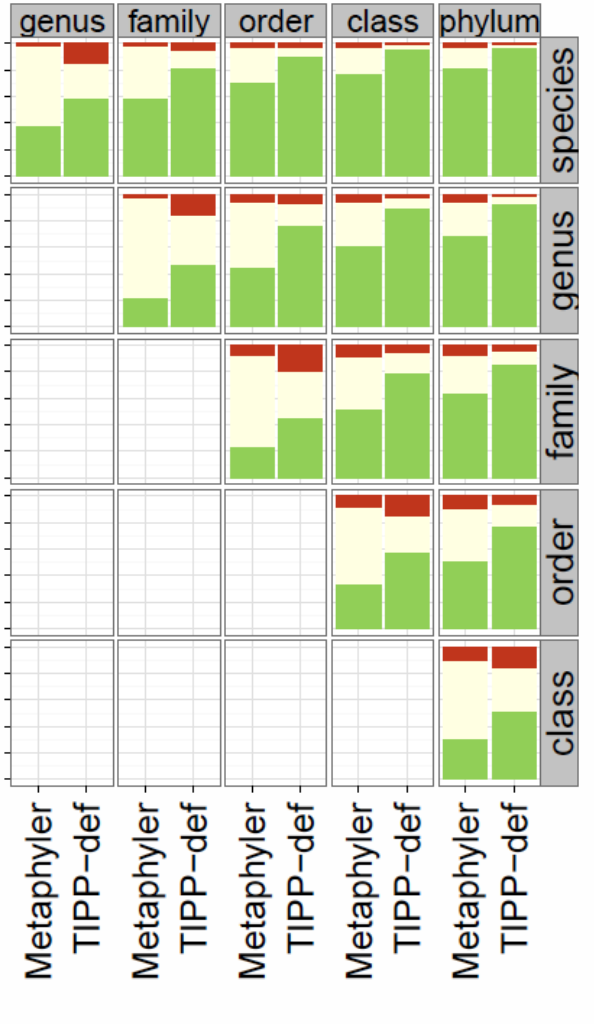
TIPP: dramatically reduces false positive rate with small reduction in true positive rate, by considering uncertainty in alignment (HMMER) and placement (pplacer)

Leave-one-out on 30 marker genes

Illumina Error model



454 Error model



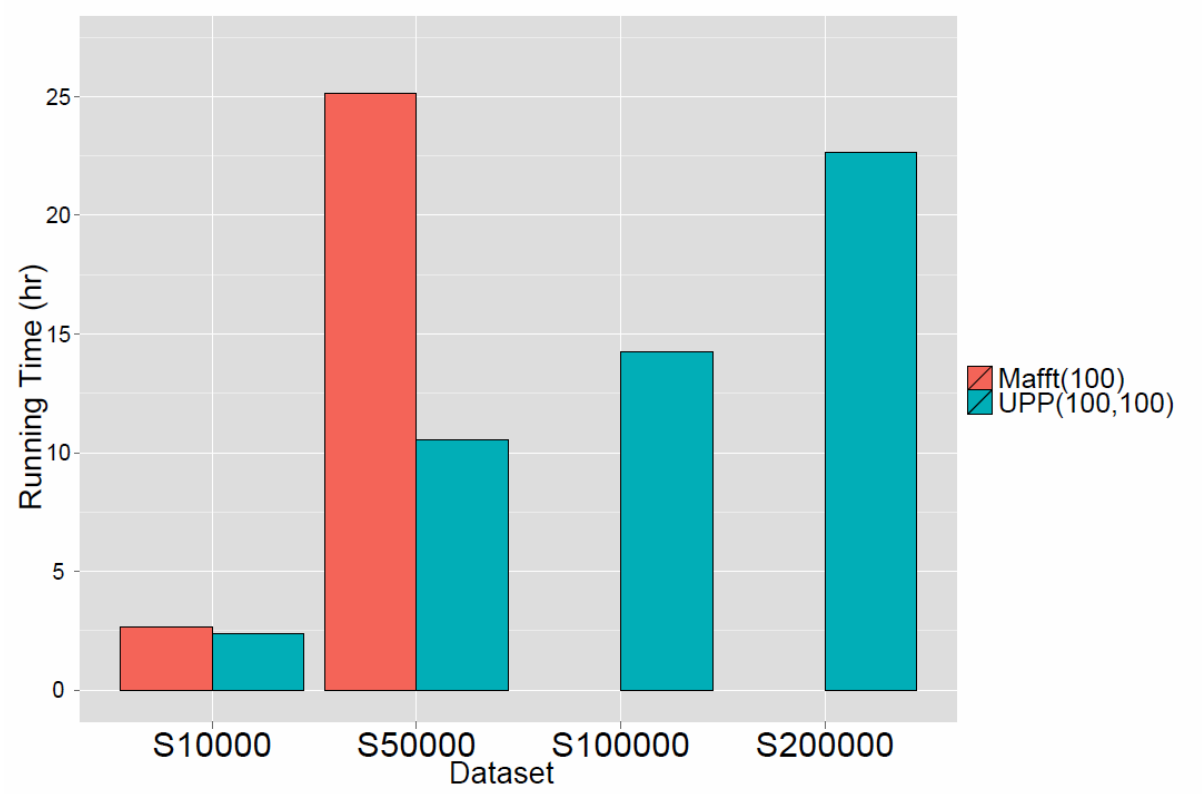
Part III: UPP: Ultra-large alignment using SEPP

Input: set S of unaligned sequences

Output: alignment and tree on S

- Select random subset X of sequences
- Estimate alignment and tree on X
- Run SEPP to align remaining sequences
- Run favorite tree estimation method on alignment
- UPP(x,y) refers to UPP using backbones of size y and alignment subsets of size x

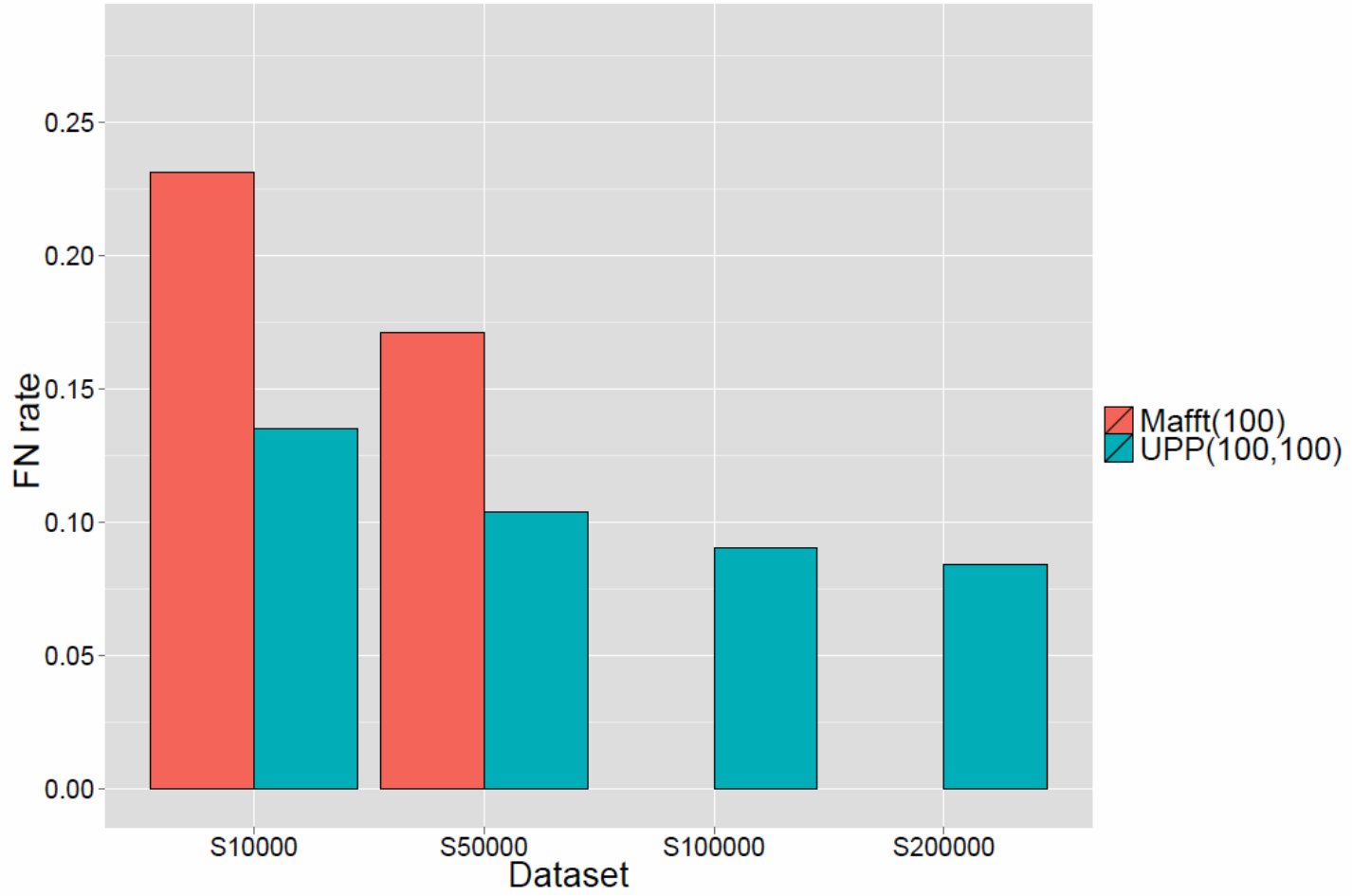
UPP vs. MAFFT: Running time (hr)



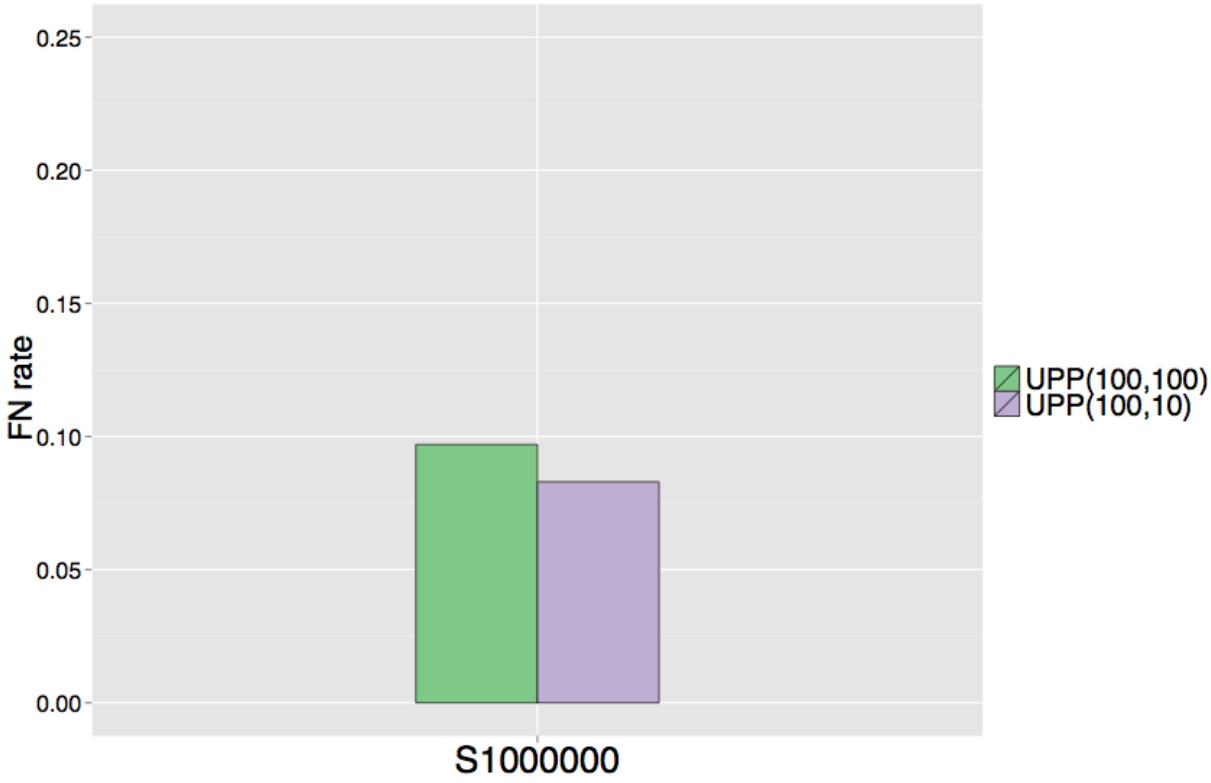
MAFFT-profile cannot run within alotted time (24 hours on 12 processors) on 100,000 sequences

RNASim data generated by Junhyong Kim (Penn)

UPP vs. MAFFT: tree error



One Million Taxa: Tree Error



Note improvement obtained by using SEPP decomposition

“Boosters”

- SATé: co-estimation of alignments and trees
- SEPP: phylogenetic placement of short reads
- TIPP: taxon identification
- UPP: ultra-large alignment

Algorithmic strategies: divide-and-conquer and iteration to improve the accuracy and scalability of a *base method.*

Big Data in Biology

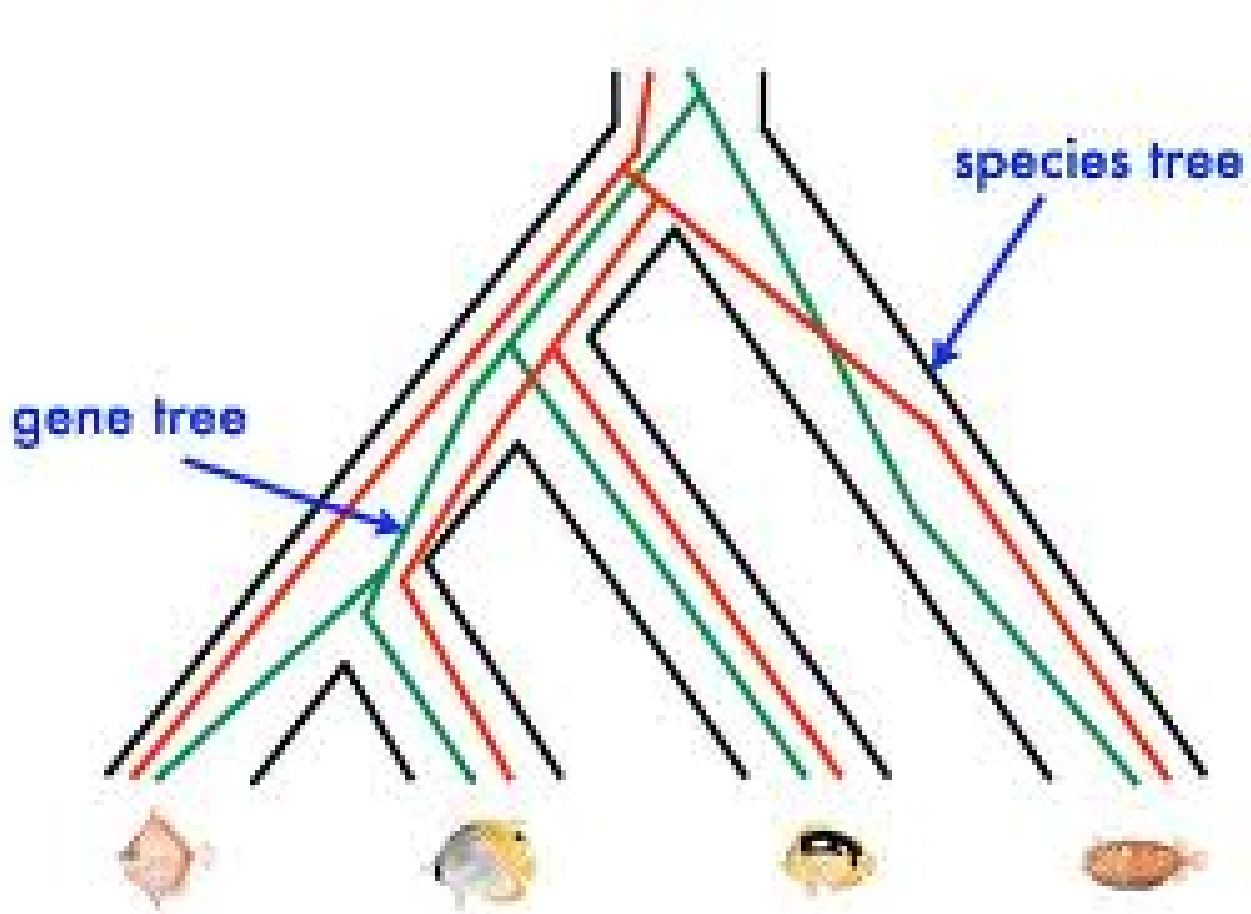
- More than just data management and HPC
- Yes, we need TACC -- but we also need new algorithms
- Important questions can be asked and (possibly) answered
- Collaboration between biologists and computer scientists is essential

Acknowledgments

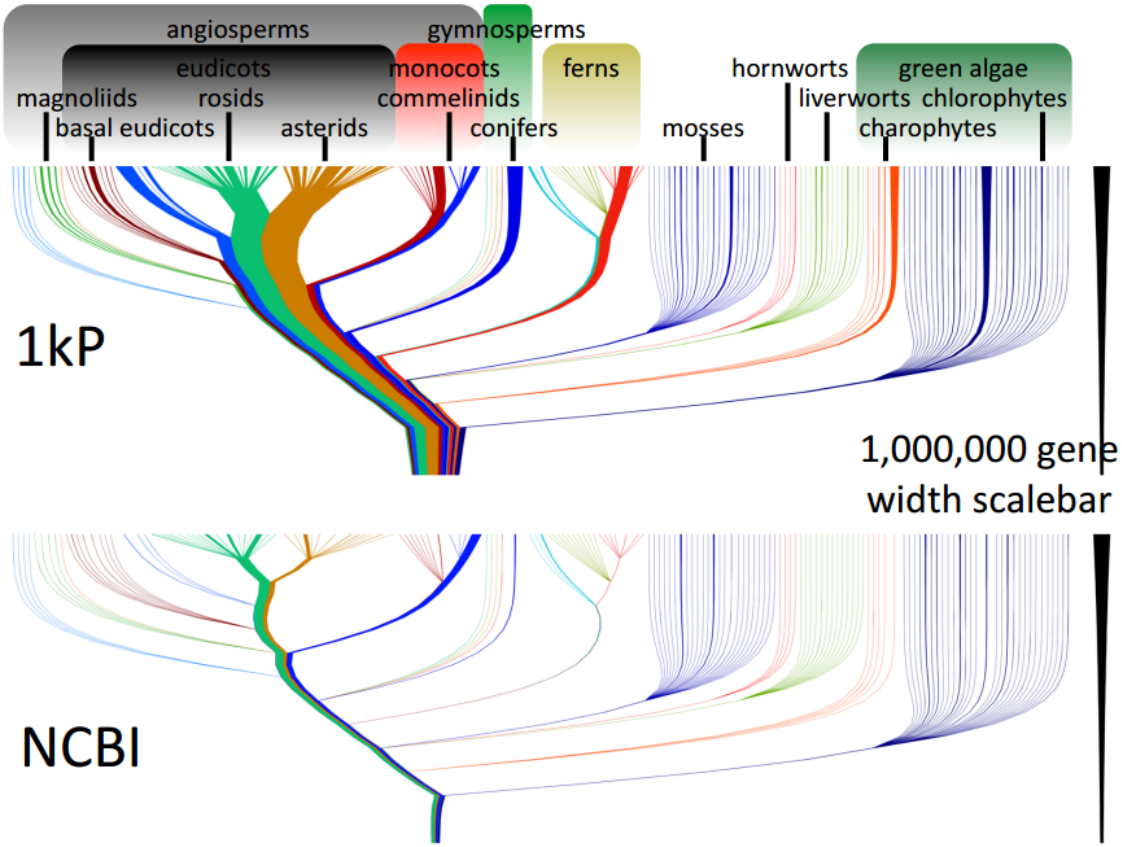
- Funding: Guggenheim Foundation Fellowship, NSF: ATOL, ITR, and IGERT grants, David Bruton Jr. Professorship; HHMI support to Siavash Mirarab, and TACC/IPLANT support to Nam Nguyen
- Computational Resources: TACC and UTCS

- Collaborators:
  - SATé: Kevin Liu, Serita Nelesen, Randy Linder, Mark Holder, Siavash Mirarab, and others
  - SEPP: Siavash Mirarab and Nam Nguyen
  - TIPP: Nam Nguyen, Siavash Mirarab, Mihai Pop, and Bo Liu
  - UPP: Nam Nguyen and Siavash Mirarab

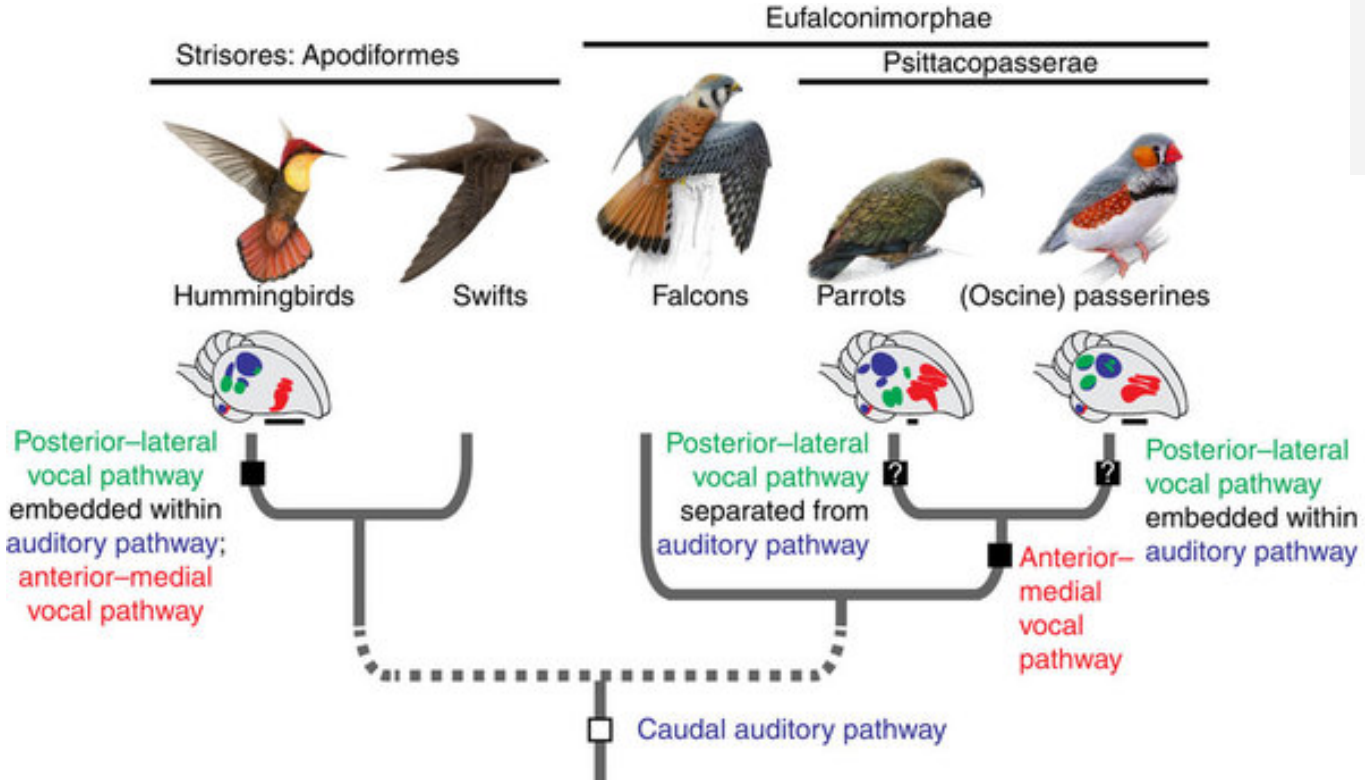
Red gene tree ≠ species tree (green gene tree okay)



1KP - dramatic increase in available sequence

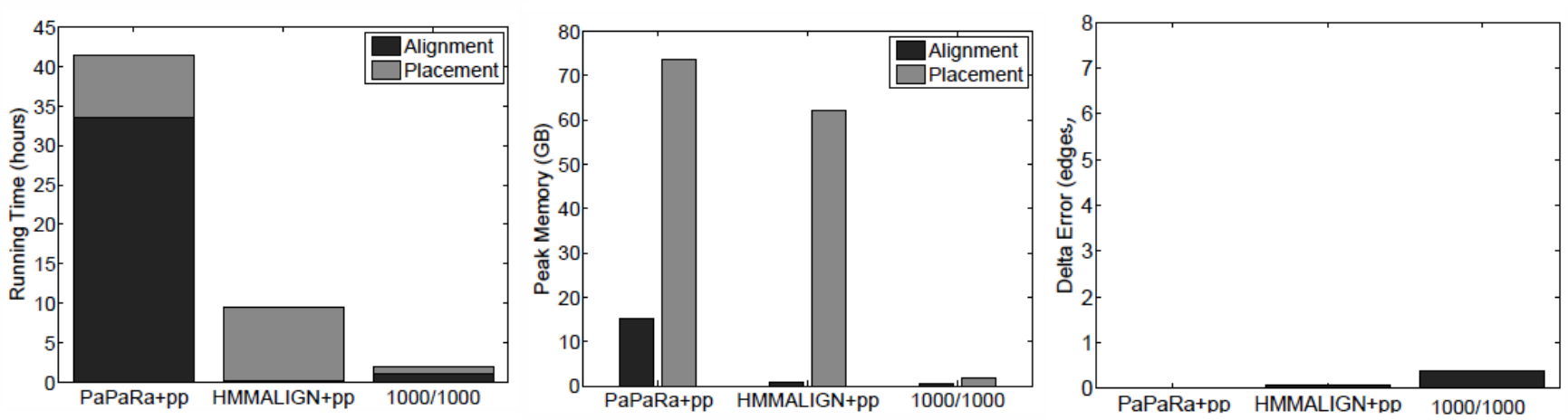


Evolution of Vocal Learning



Erich Jarvis, HHMI

SEPP (10%) on Biological Data



16S.B.ALL dataset, 13k curated backbone tree, 13k total fragments

For 1 million fragments:

PaPaRa+pplacer: ~133 days

HMMALIGN+pplacer: ~30 days

SEPP 1000/1000: ~6 days

Phylogenetic “boosters” (meta-methods)

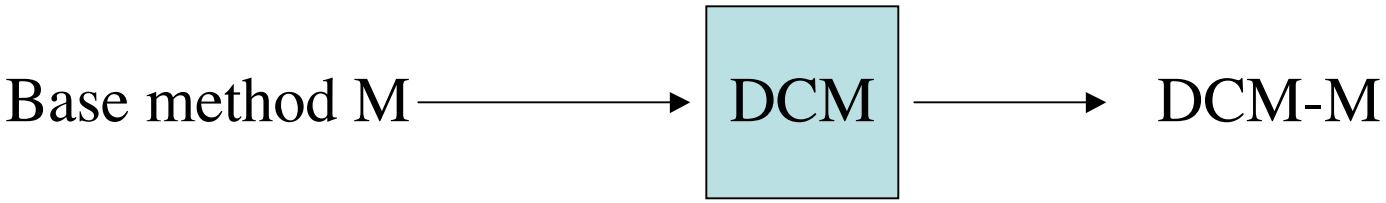
Goal: improve accuracy, speed, robustness, or theoretical guarantees of base methods

Examples:

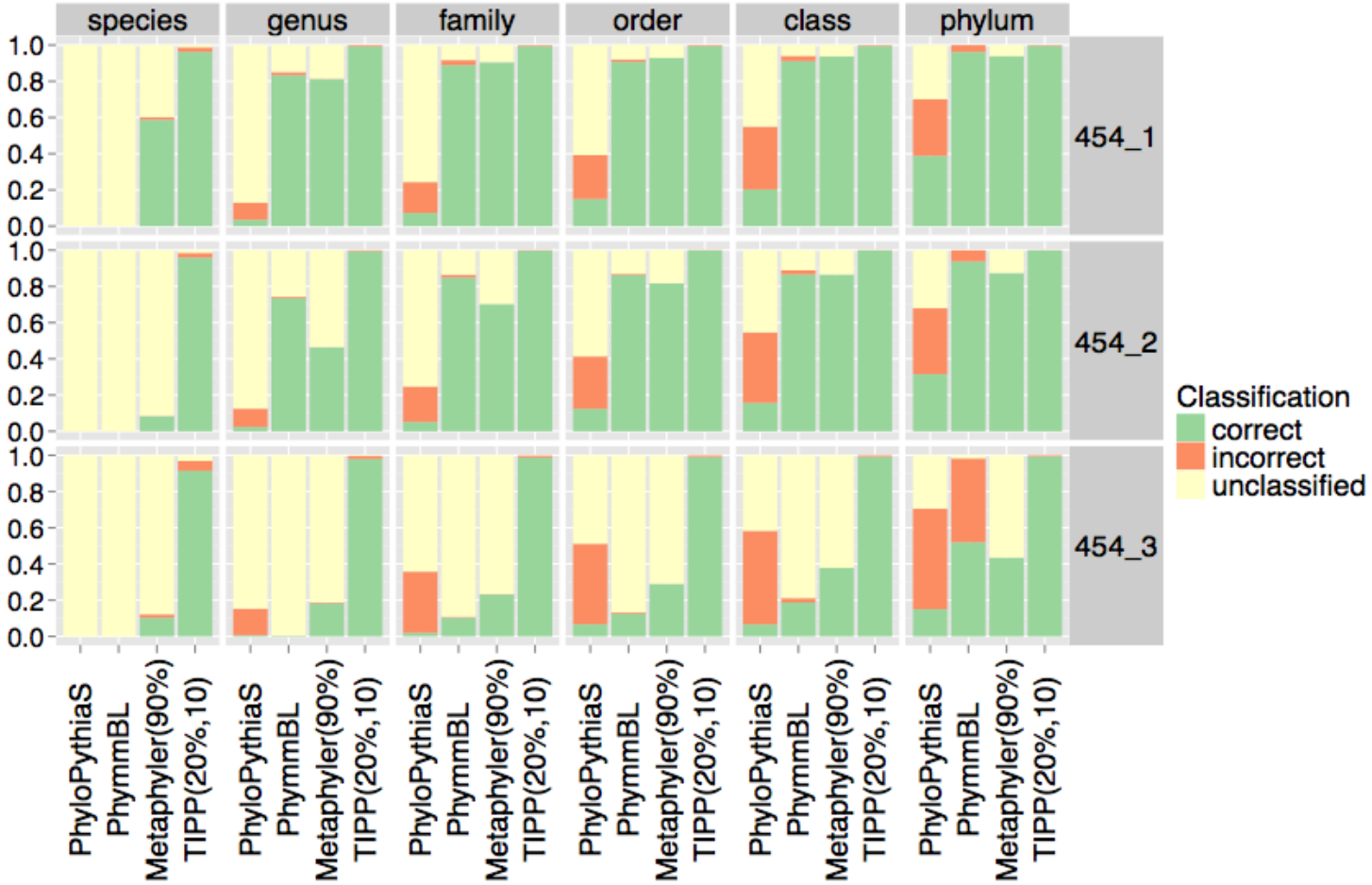
- DCM-boosting for distance-based methods (1999)
- DCM-boosting for heuristics for NP-hard problems (1999)
- SATé-boosting for alignment methods (2009)
- SuperFine-boosting for supertree methods (2011)
- DACTAL-boosting for all phylogeny estimation methods (2011)
- SEPP-boosting for metagenomic analyses (2013)
- SEPP-boosting for ultra-large alignment estimation (2013)

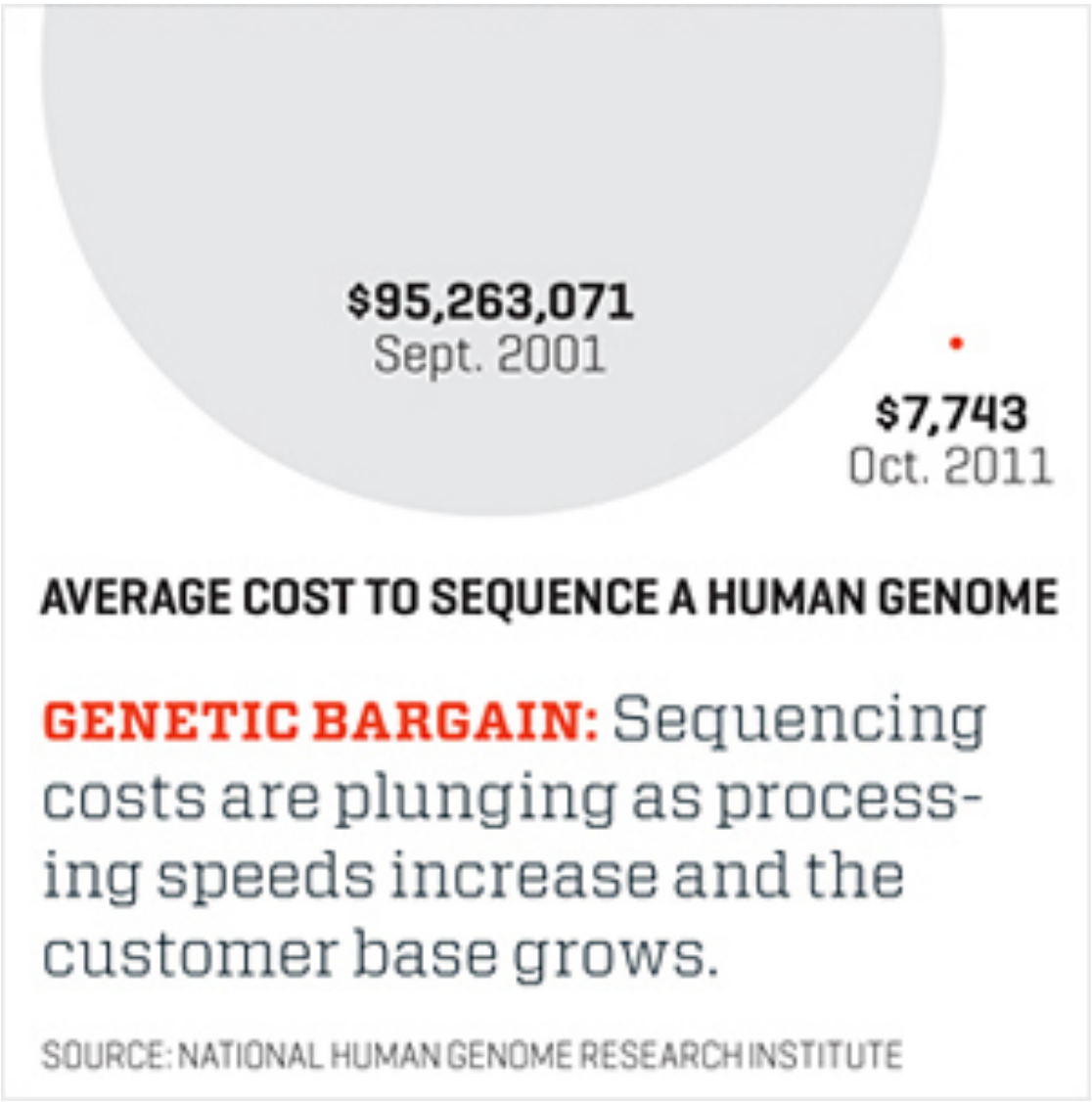
Algorithmic “Boosters” (DCMs)

- These divide-and-conquer techniques “boost” the performance of base methods (for alignment, for tree estimation, for classification, etc.).



TIPP: Taxon Identification using SEPP (30 marker genes, non-leave-out, with 454 sequencing error)




$95,263,071
Sept. 2001
$7,743
Oct. 2011
AVERAGE COST TO SEQUENCE A HUMAN GENOME
GENETIC BARGAIN: Sequencing costs are plunging as processing speeds increase and the customer base grows.
SOURCE: NATIONAL HUMAN GENOME RESEARCH INSTITUTE

BIG DATA FOR BIOLOGY

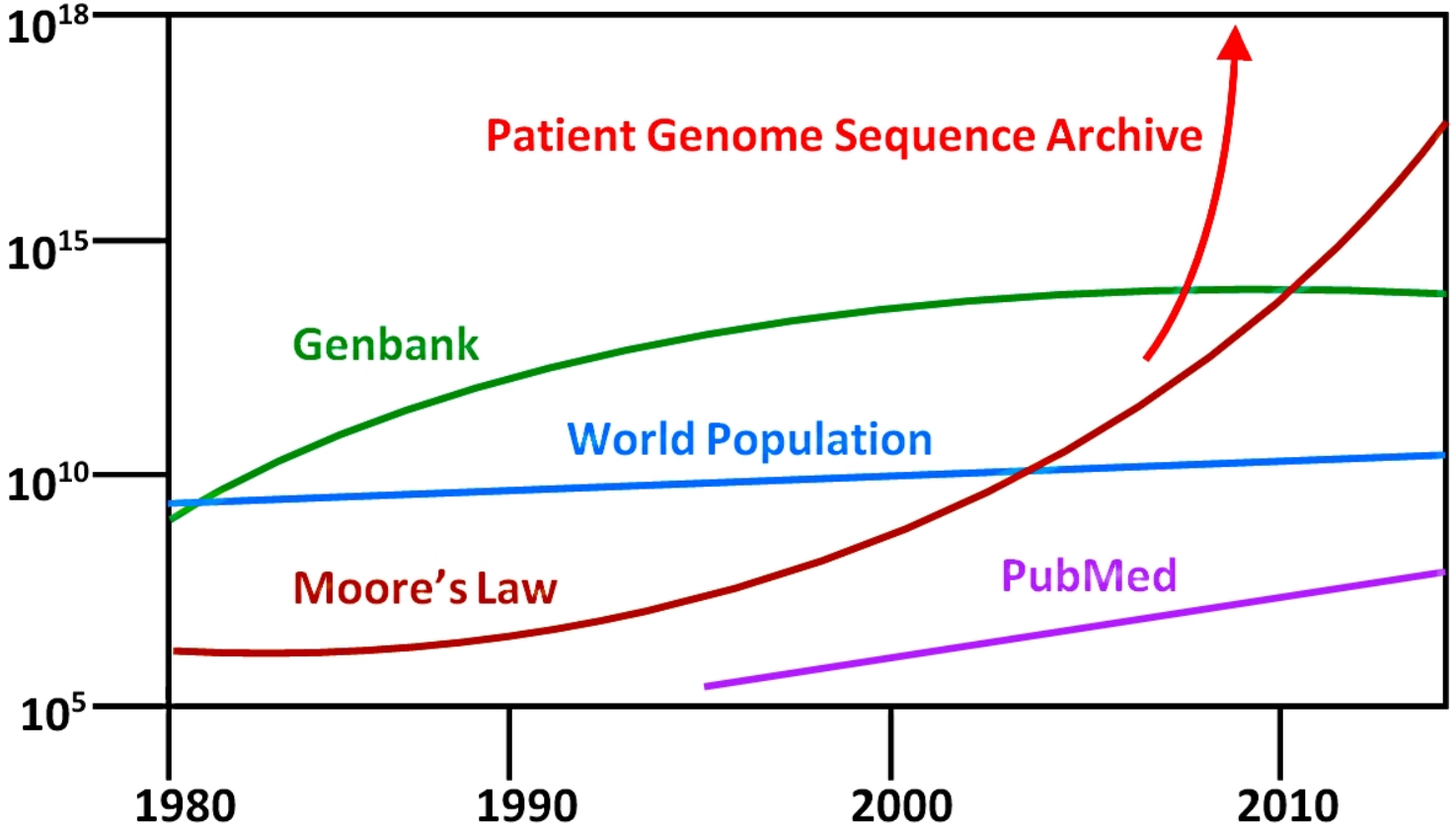
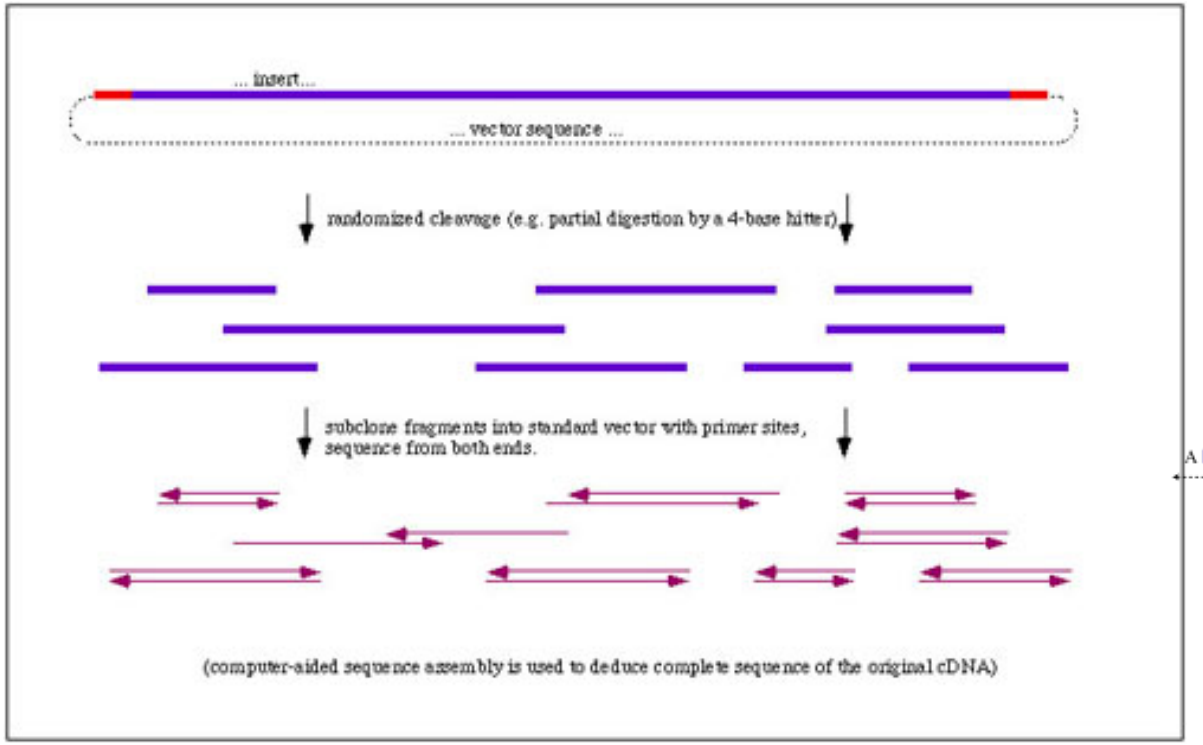
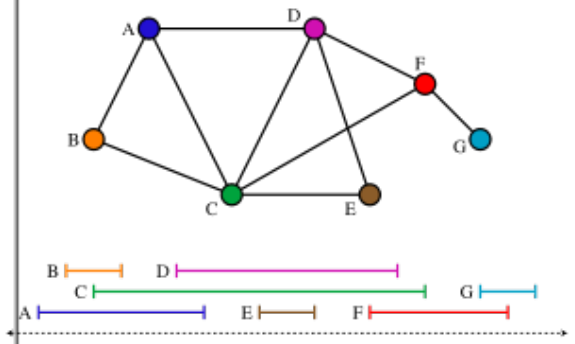


Figure 1: Approximate Growth of Different Data Populations

Whole Genome Assembly

Graph Algorithms and Combinatorial Optimization!

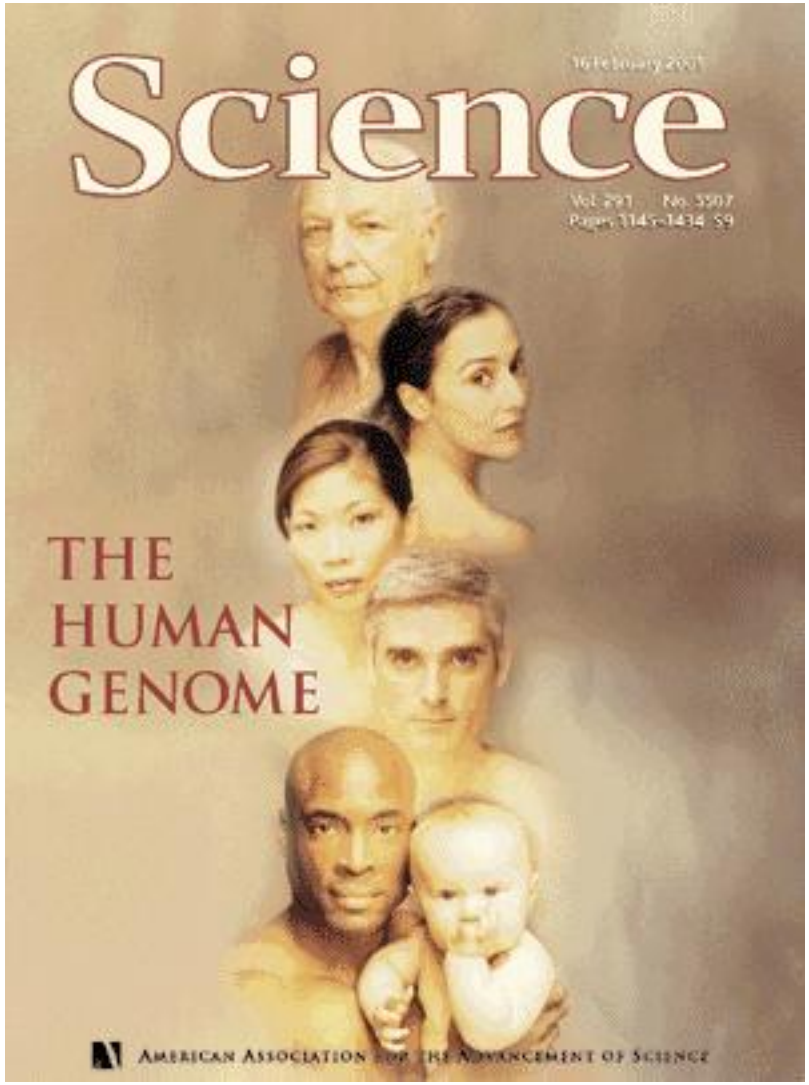




But:

- Modern Sequencing Technologies create short reads with high error rates - need new genome assembly methods
- Metagenome assembly even harder.

BigData for Biology: Genomics

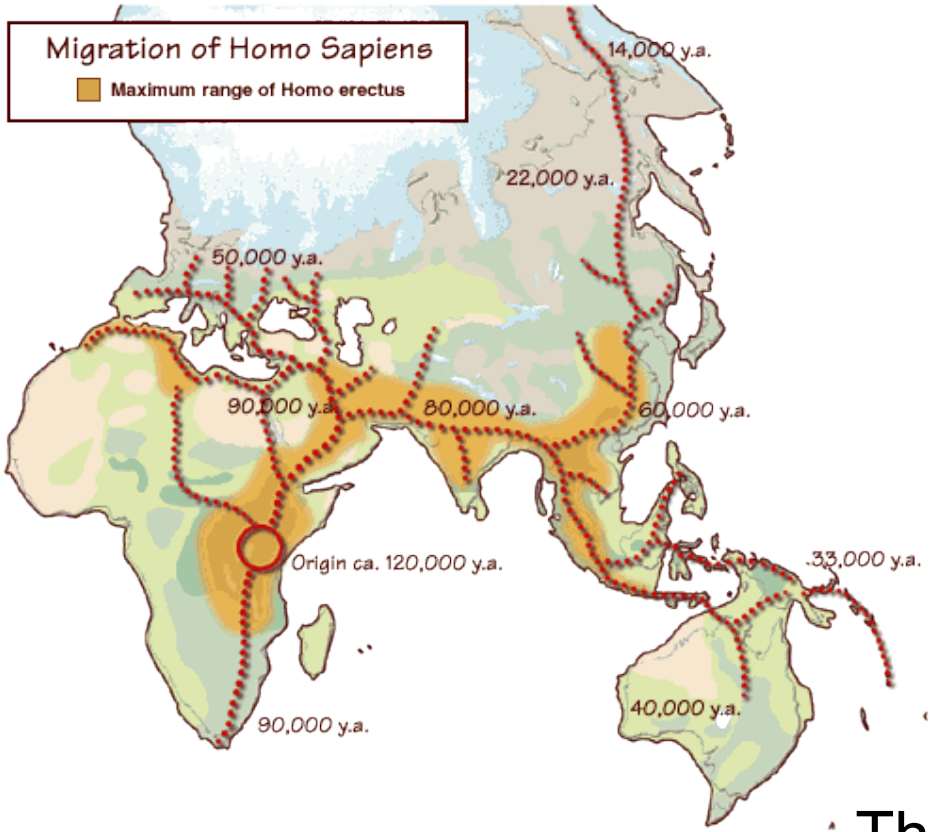


Biology: 21st Century Science!

"When the human genome was sequenced seven years ago, scientists knew that most of the major scientific discoveries of the 21st century would be in biology."

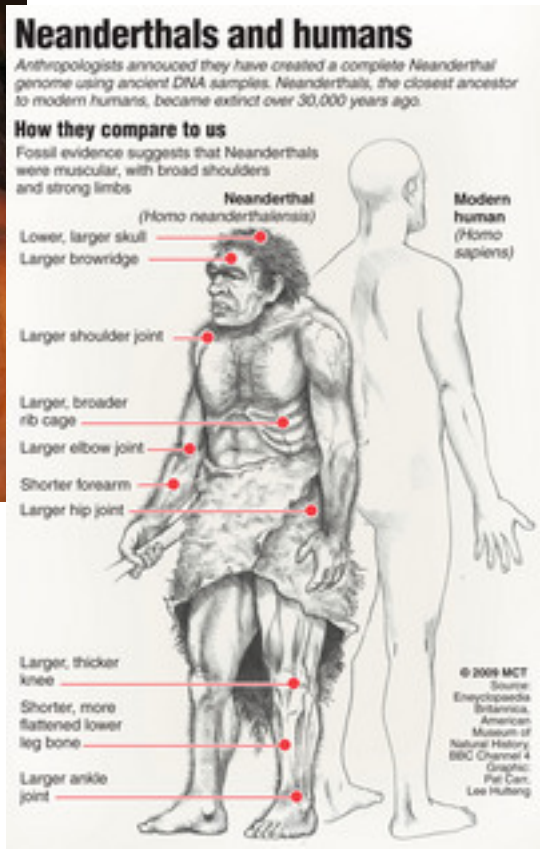
January 1, 2008, guardian.co.uk

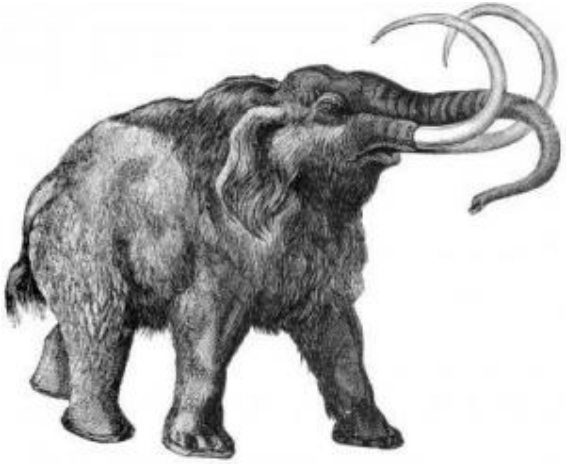
Where did humans come from, and how did they move throughout the globe?



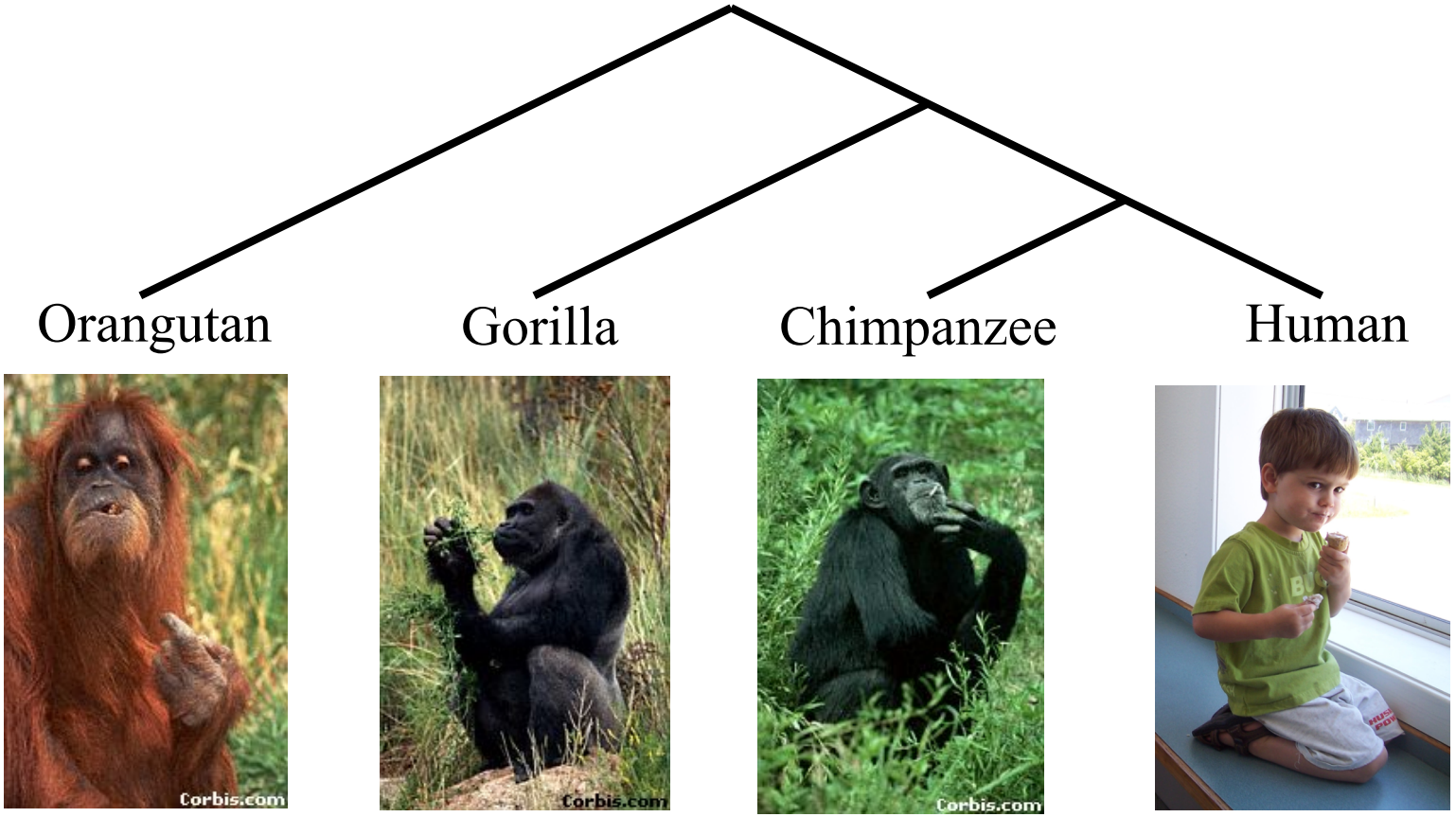
- The 1000 Genome Project: using human genetic variation to better treat diseases

Other Genome Projects! (Neanderthals, Woolly Mammoths, and more ordinary creatures…)





Phylogeny (evolutionary tree)



From the Tree of the Life Website,
University of Arizona

Alignment and Phylogeny Estimation: BigData Problems

- Sequence alignment of many sequences: standard methods do not provide good accuracy, and many cannot run
- Tree estimation for many sequences: no methods are fast enough (many cannot be run) - years of analysis
- Species tree estimation from multiple genes: species trees can differ from gene trees due to incomplete lineage sorting, hybridization, and horizontal gene transfer (among other causes)
- Fragmentary sequences due to sequencing technologies (incomplete assembly)
- Missing data (not all genes in all species)

Alignment and Phylogeny Estimation: BigData Problems

- Sequence alignment of many sequences: standard methods do not provide good accuracy, and many cannot run
- Tree estimation for many sequences: no methods are fast enough (many cannot be run) - years of analysis
- Species tree estimation from multiple genes: species trees can differ from gene trees due to incomplete lineage sorting, hybridization, and horizontal gene transfer (among other causes)
- Fragmentary sequences due to sequencing technologies (incomplete assembly)
- Missing data (not all genes in all species)