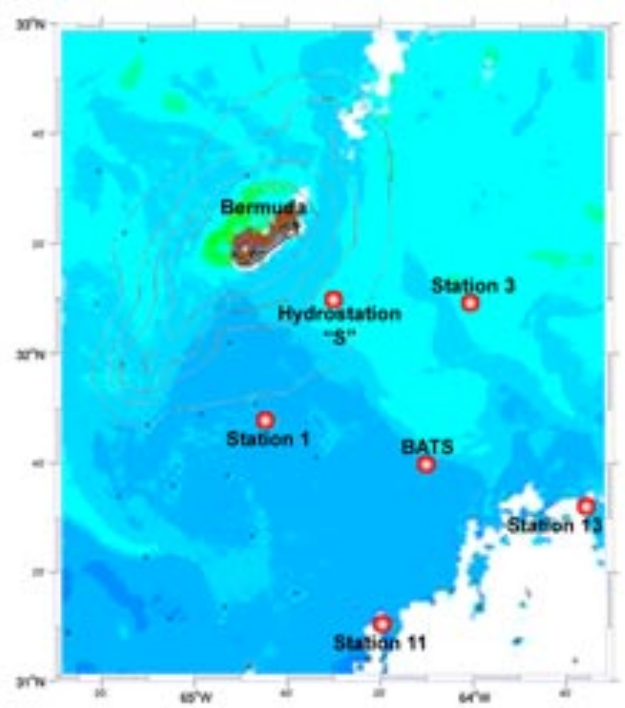
SEPP and TIPP for metagenomic analysis

Tandy Warnow
Department of Computer Science
University of Texas

Metagenomics:

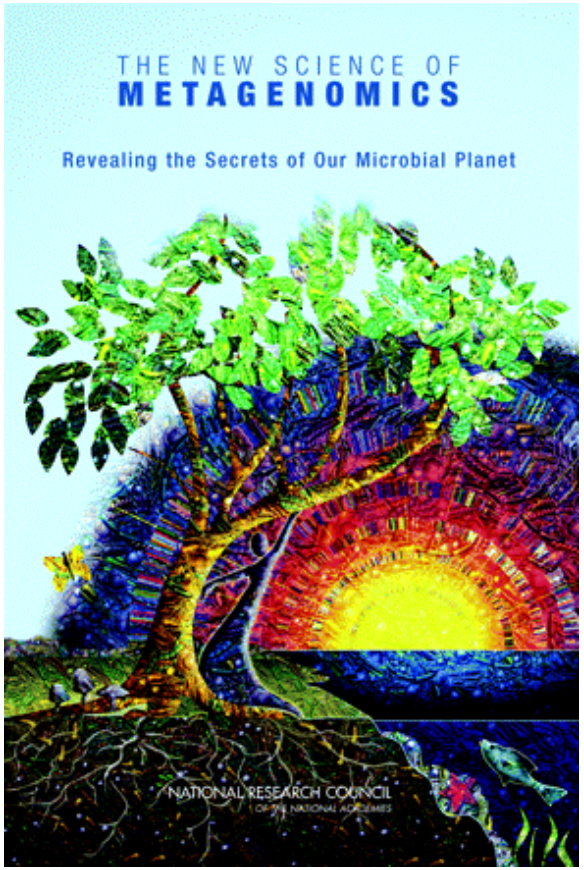
Venter et al., Exploring the Sargasso Sea:

Scientists Discover One Million New Genes in Ocean Microbes



Computational Phylogenetics and Metagenomics

Courtesy of the Tree of Life project



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

Issues: accuracy and speed

Phylogenetic Placement

Input: Backbone alignment and tree on full-length sequences, and a set of query sequences (short fragments)

Output: Placement of query sequences on backbone tree

Phylogenetic placement can be used for taxon identification, but it has general applications for phylogenetic analyses of NGS data.

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)
- **Metagenomic** analyses: methods for species classification of short reads have *poor sensitivity*. Efficient high throughput is necessary (millions of reads).

Today's Talk

- SATé: Simultaneous Alignment and Tree Estimation (Liu et al., Science 2009, and Liu et al. Systematic Biology, 2011)
- SEPP: SATé-enabled Phylogenetic Placement (Mirarab, Nguyen and Warnow, Pacific Symposium on Biocomputing 2012)
- TIPP: Taxon Identification using Phylogenetic Placement (Nguyen, Mirarab, and Warnow, in preparation)

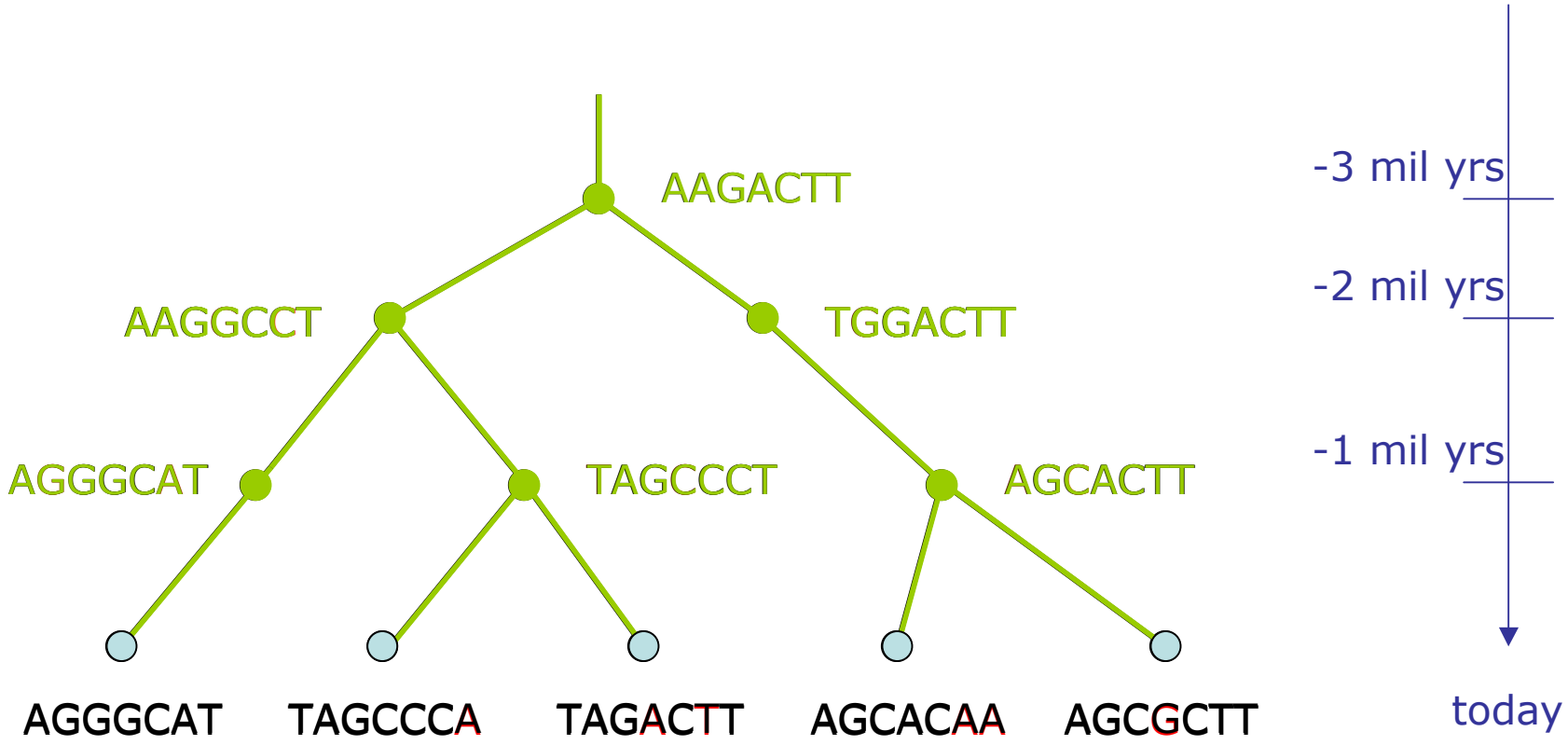
Part 1: SATé

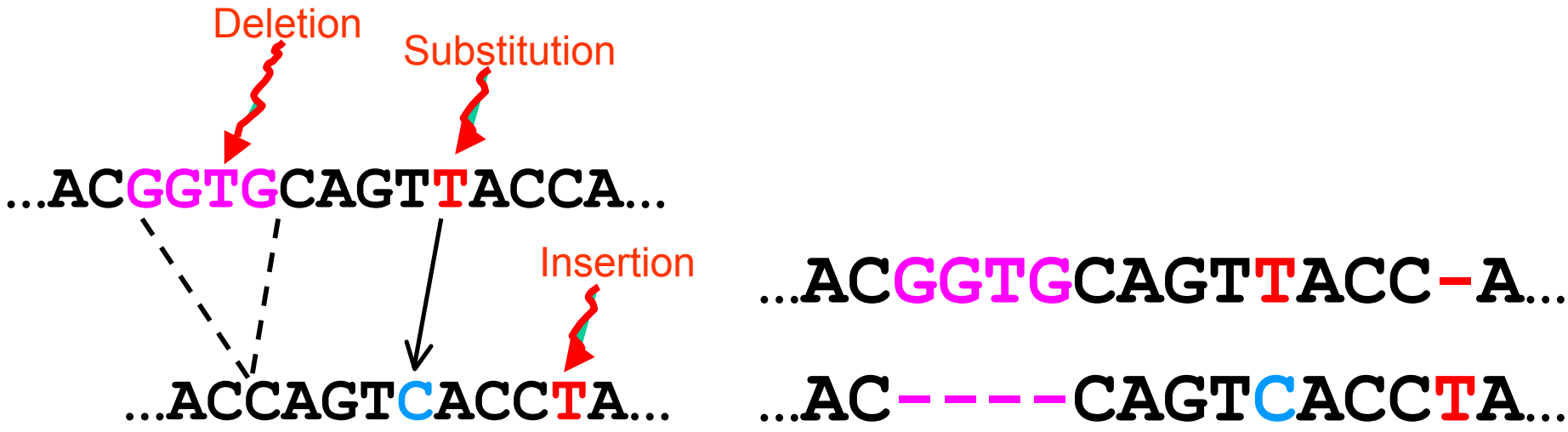
Liu, Nelesen, Raghavan, Linder, and Warnow, *Science*, 19 June 2009, pp. 1561-1564.

Liu et al., Systematic Biology, 2011, 61(1):90-106

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

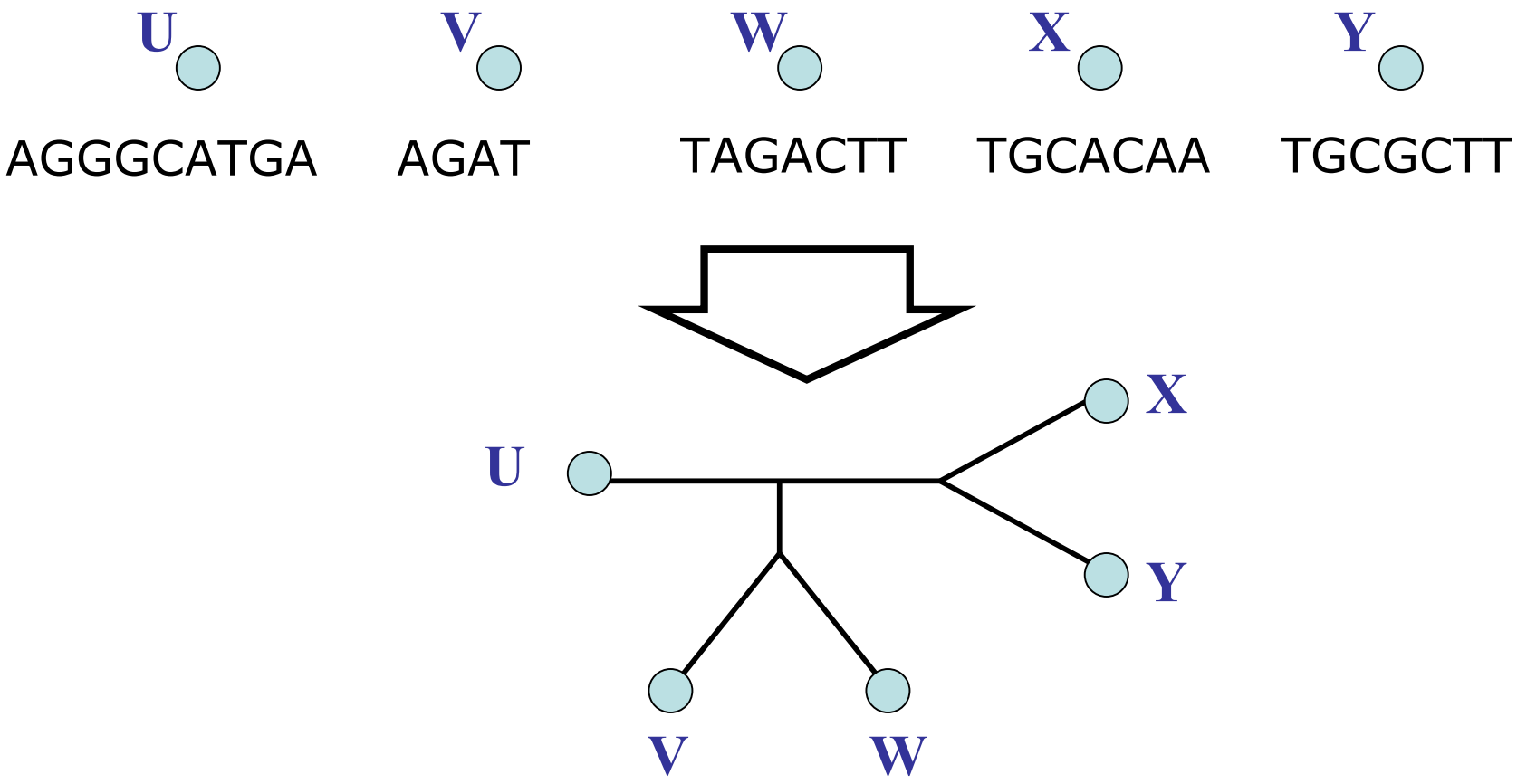
DNA Sequence Evolution





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

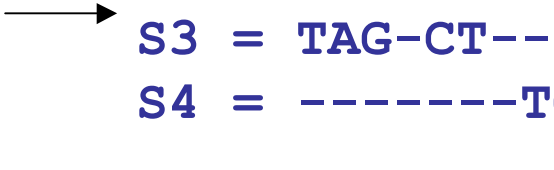
```
S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC
S3 = TAGCTGACCGC
S4 = TCACGACCGACA
```



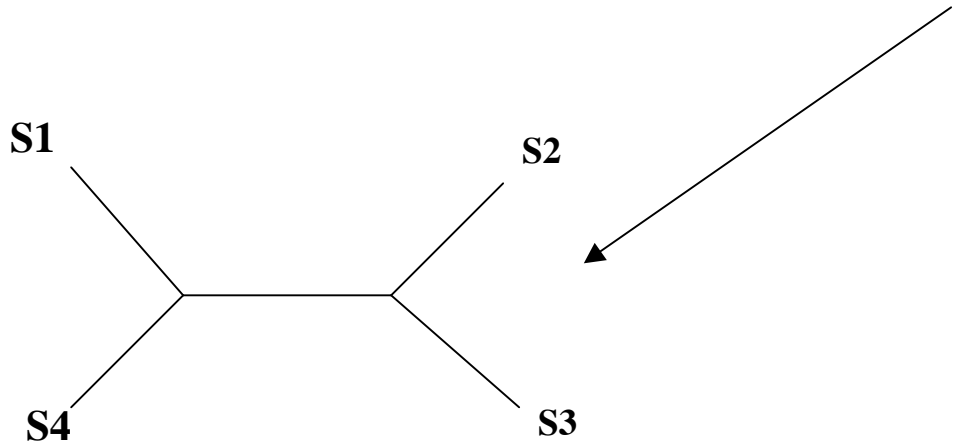
```
S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAC--GACCGC--
S3 = TAG-CT-------GACCGC--
S4 = -------TCAC--GACCGACA
```

Phase 2: Construct tree

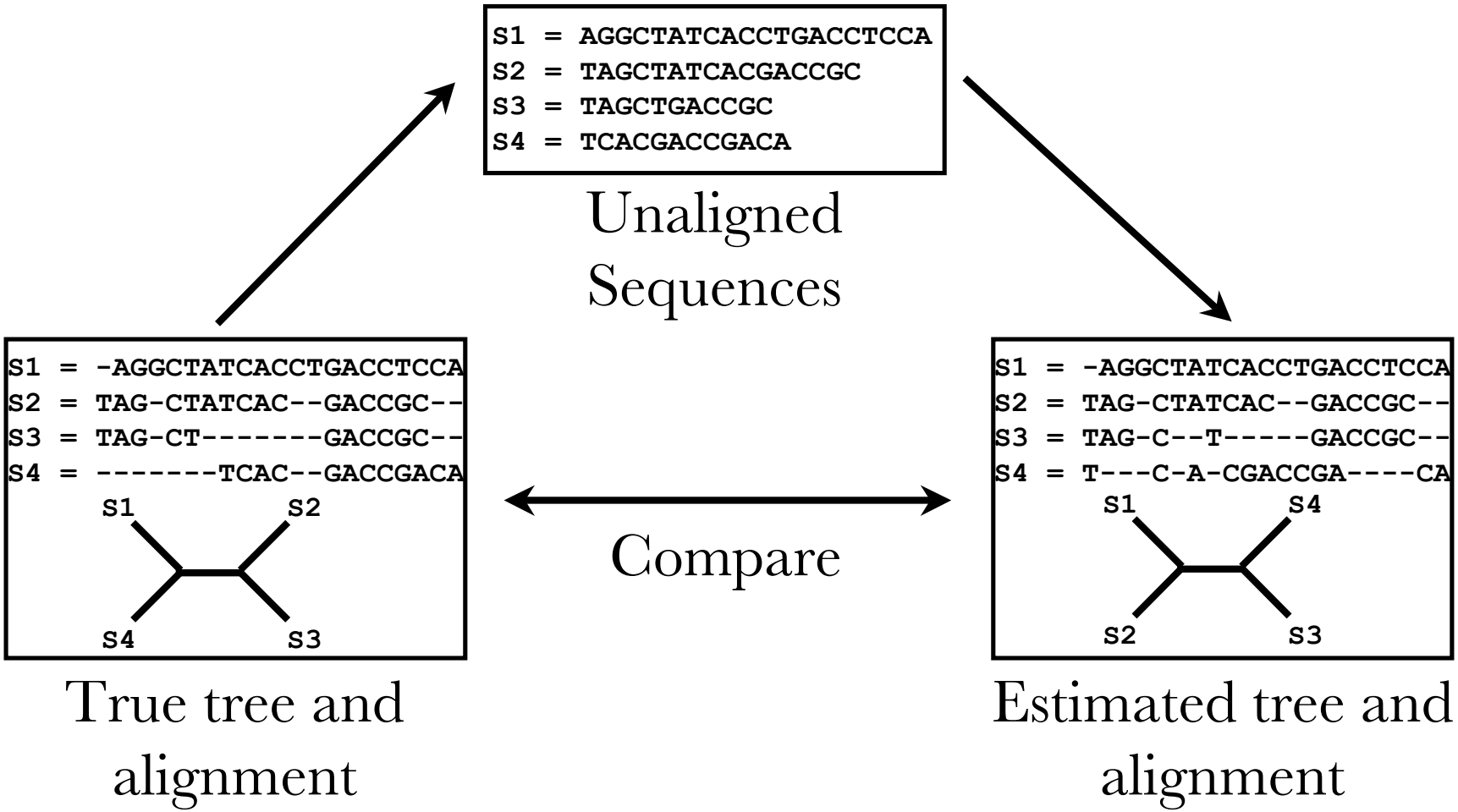
```
S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC
S3 = TAGCTGACCGC
S4 = TCACGACCGACA
```



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



Simulation Studies



Two-phase estimation

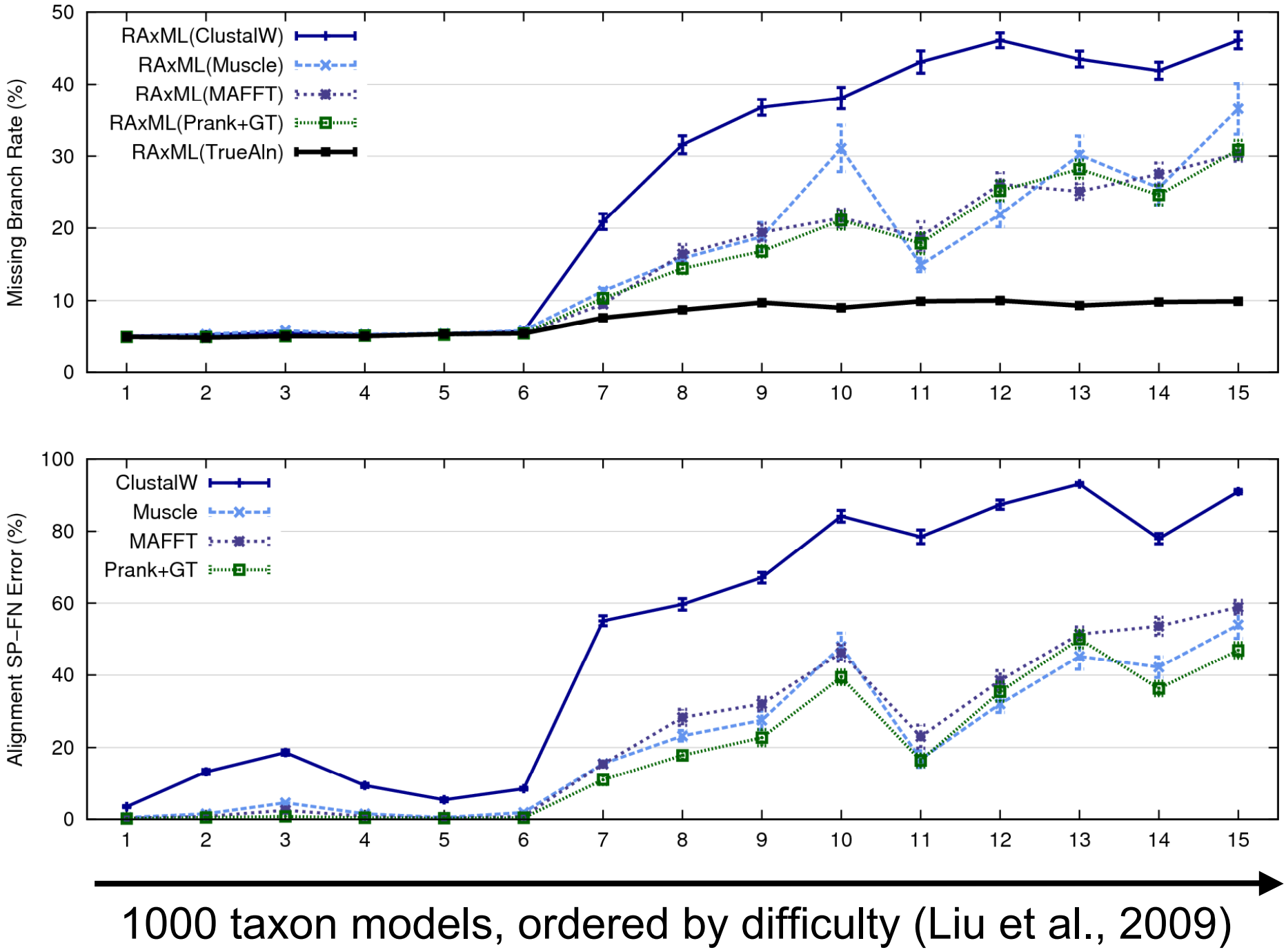
Alignment methods

- 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.

***RAxML**: heuristic for large-scale ML optimization*


Missing Branch Rate (%)
RAxML(ClustalW)
RAxML(Muscle)
RAxML(MAFFT)
RAxML(Prank+GT)
RAxML(TrueAln)
Alignment SP–FN Error (%)
ClustalW
Muscle
MAFFT
Prank+GT
1000 taxon models, ordered by difficulty (Liu et al., 2009)

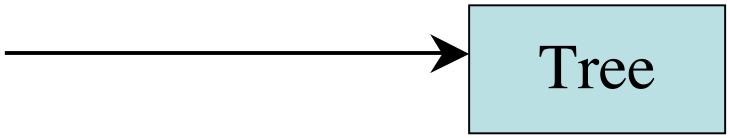
Problems

- Large datasets with high rates of evolution are hard to align accurately, and phylogeny estimation methods produce poor trees when alignments are poor.
- Many phylogeny estimation methods have poor accuracy on large datasets (even if given correct alignments)
- *Potentially useful genes are often discarded* if they are difficult to align.

These issues seriously impact large-scale phylogeny estimation (and Tree of Life projects)

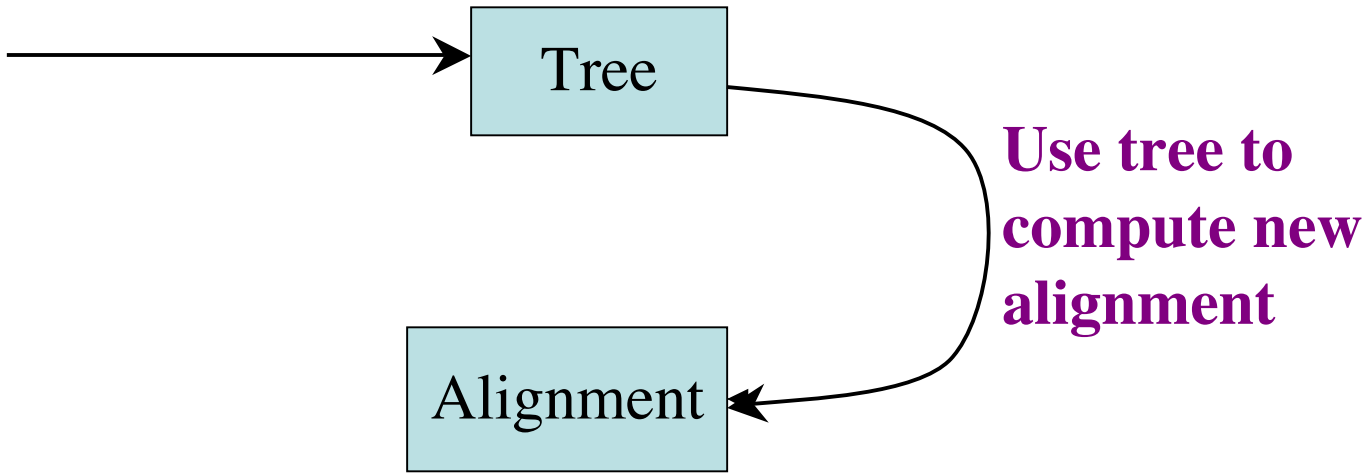
SATé Algorithm

Obtain initial alignment
and estimated ML tree



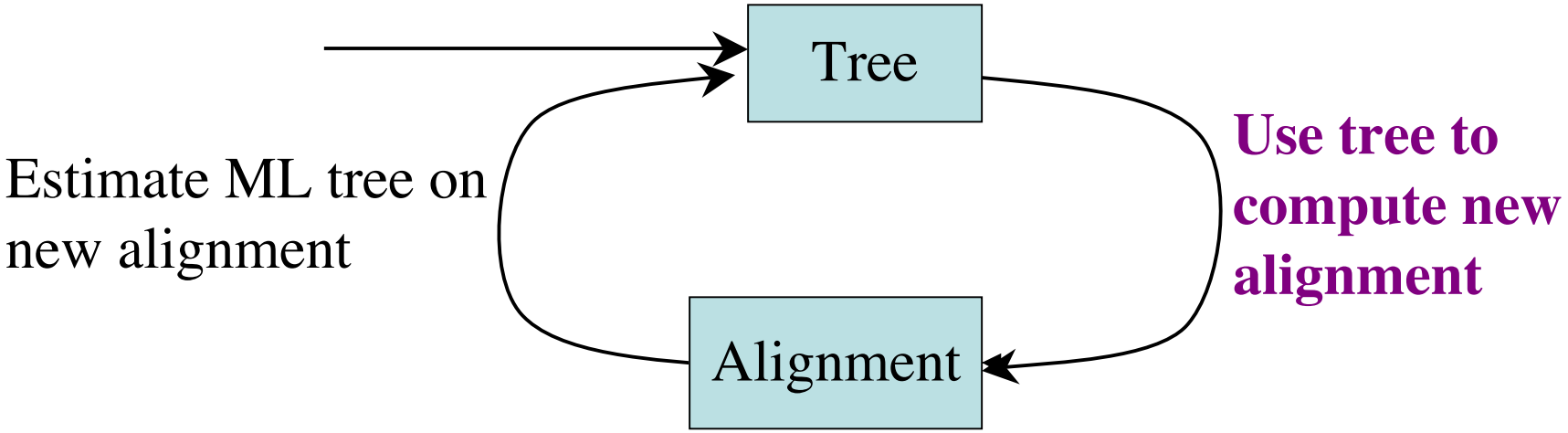
SATé Algorithm

Obtain initial alignment
and estimated ML tree



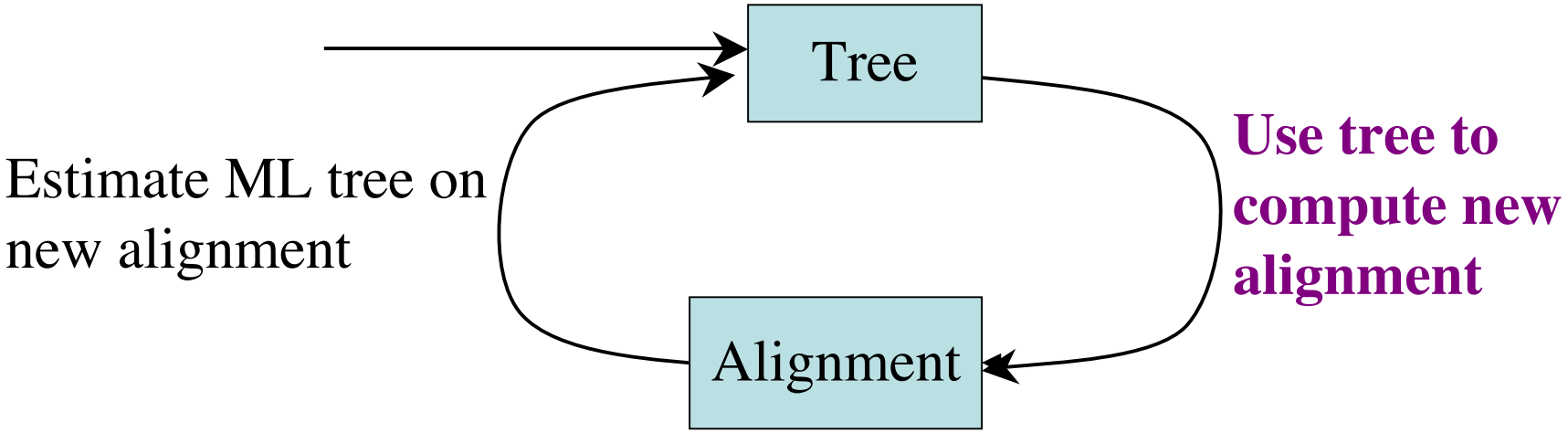
SATé Algorithm

Obtain initial alignment
and estimated ML tree



SATé Algorithm

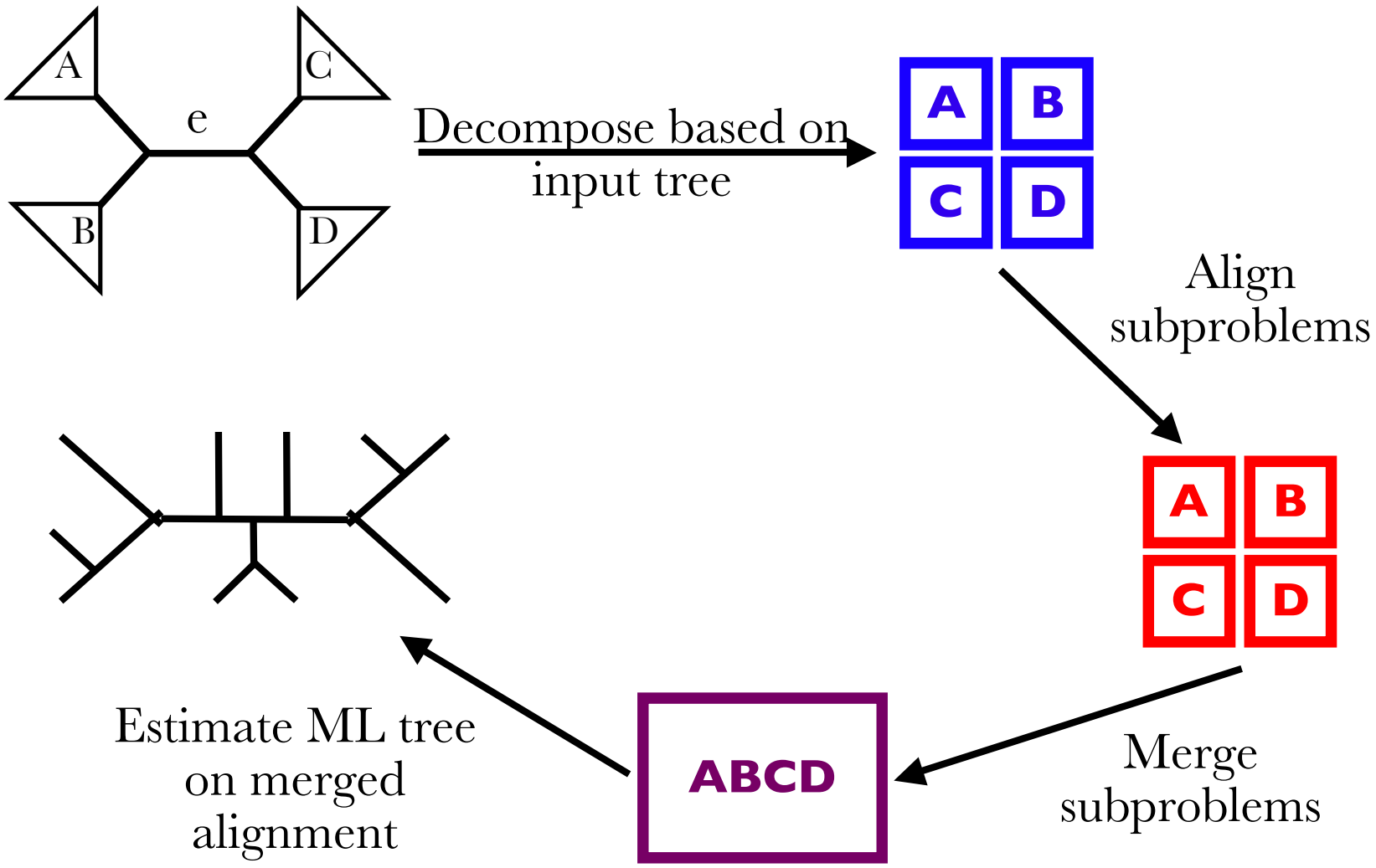
Obtain initial alignment
and estimated ML tree

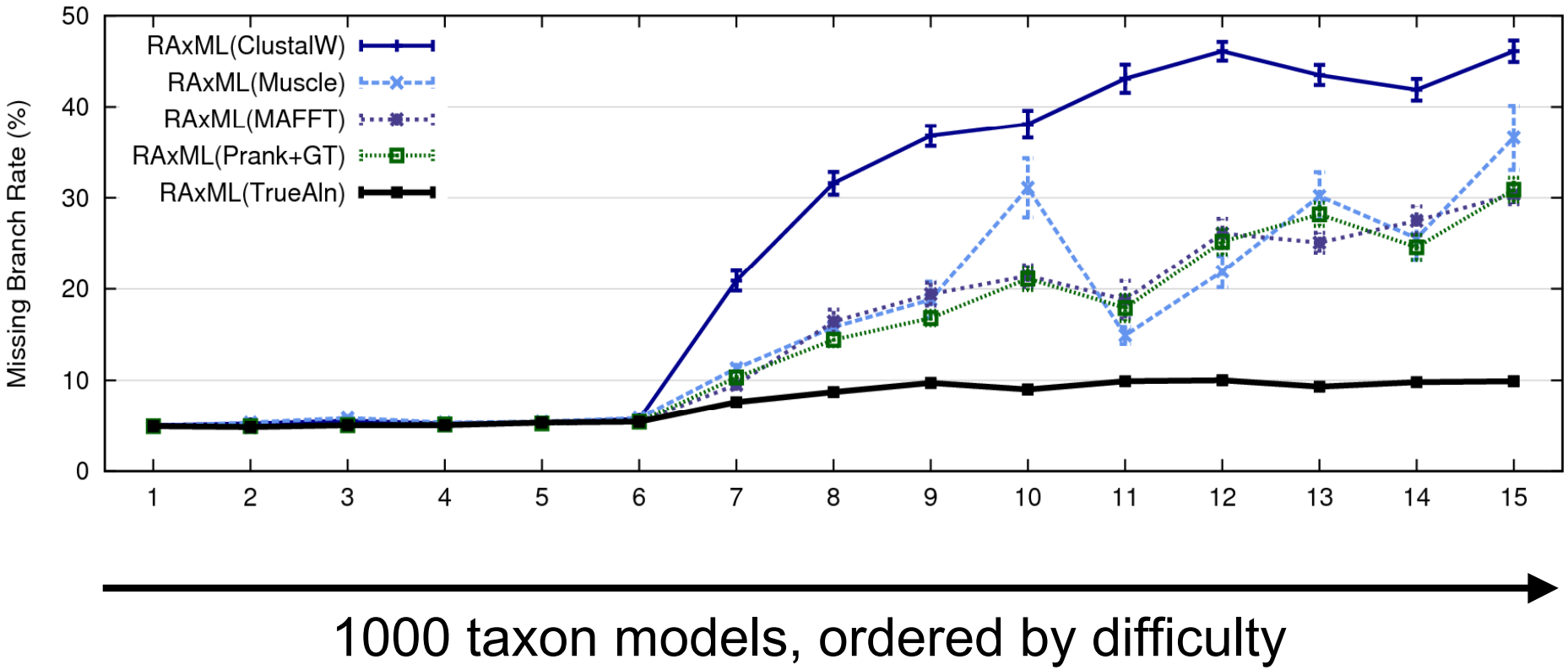


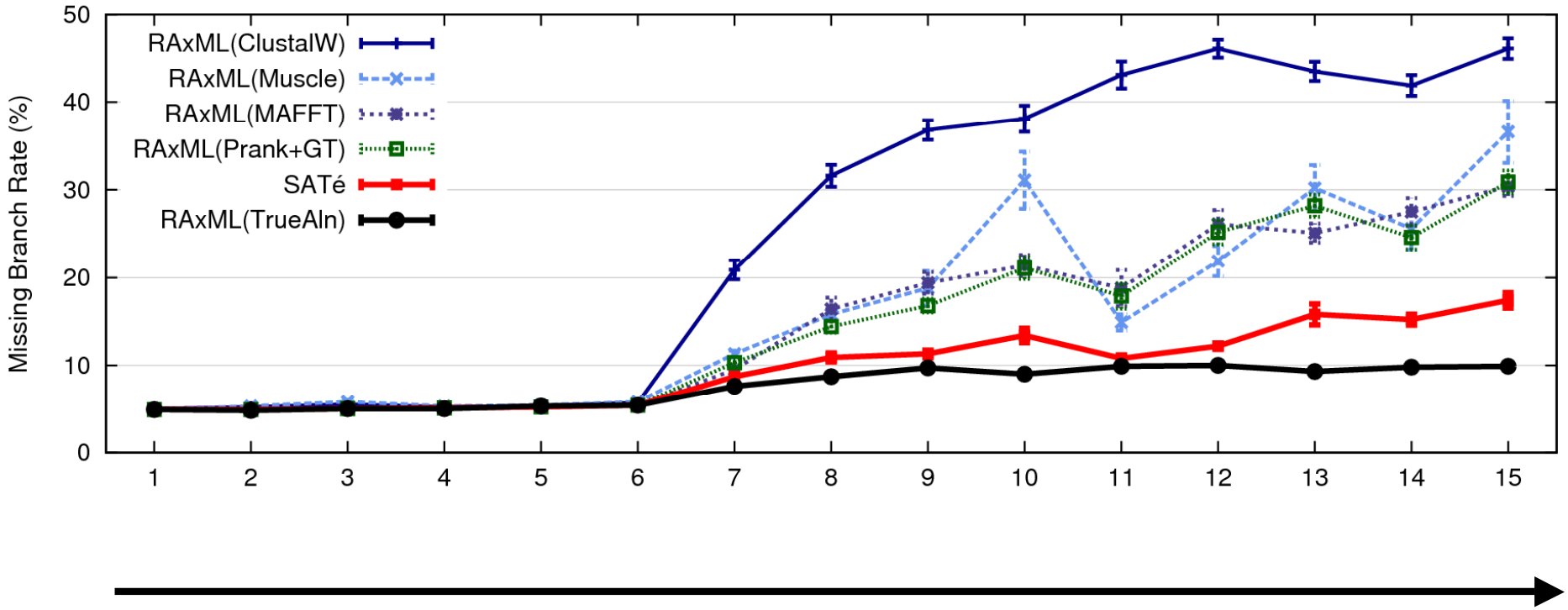
If new alignment/tree pair has worse ML score, realign using a different decomposition

Repeat until termination condition (typically, 24 hours)

One SATé iteration (really 32 subsets)



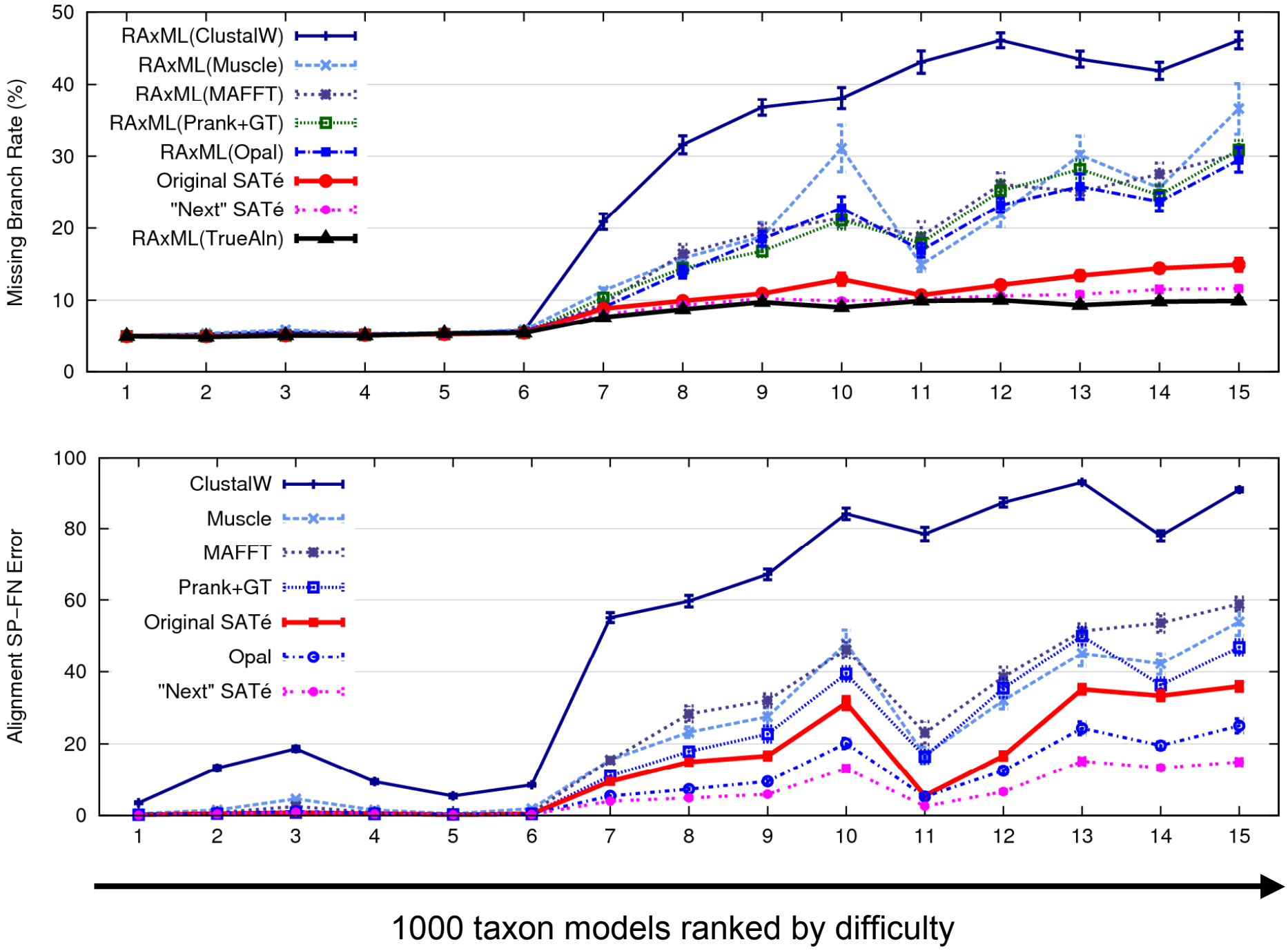

RAxML(ClustalW)
RAxML(Muscle)
RAxML(MAFFT)
RAxML(Prank+GT)
RAxML(TrueAln)
Missing Branch Rate (%)
0
10
20
30
40
50
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
1000 taxon models, ordered by difficulty



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

Part II: SEPP

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

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

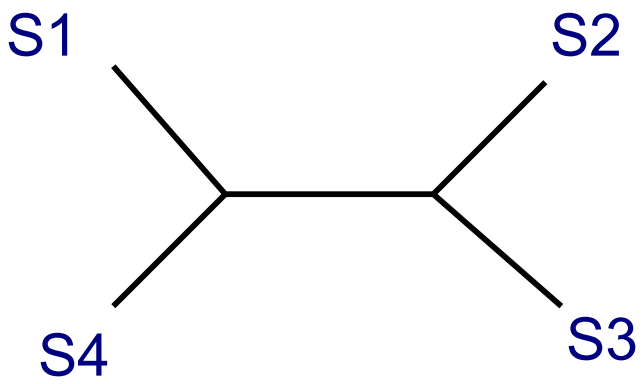
Phylogenetic Placement

- Align each query sequence to backbone alignment

- Place each query sequence into backbone tree, using extended alignment

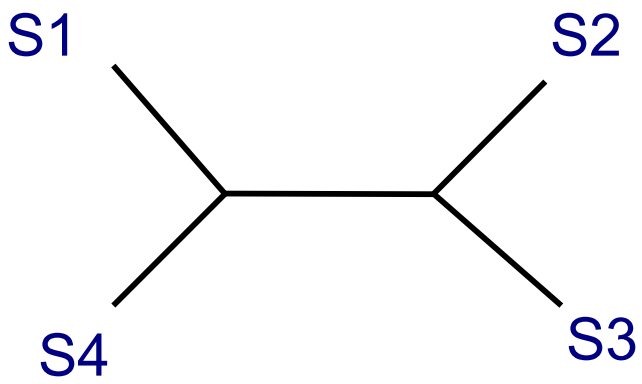
Align Sequence

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



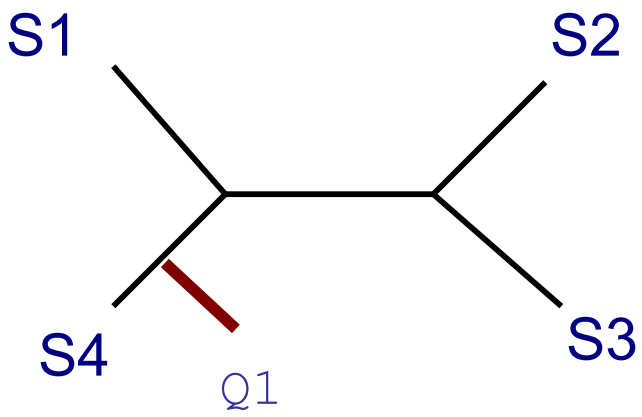
Align Sequence

```
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--------
```

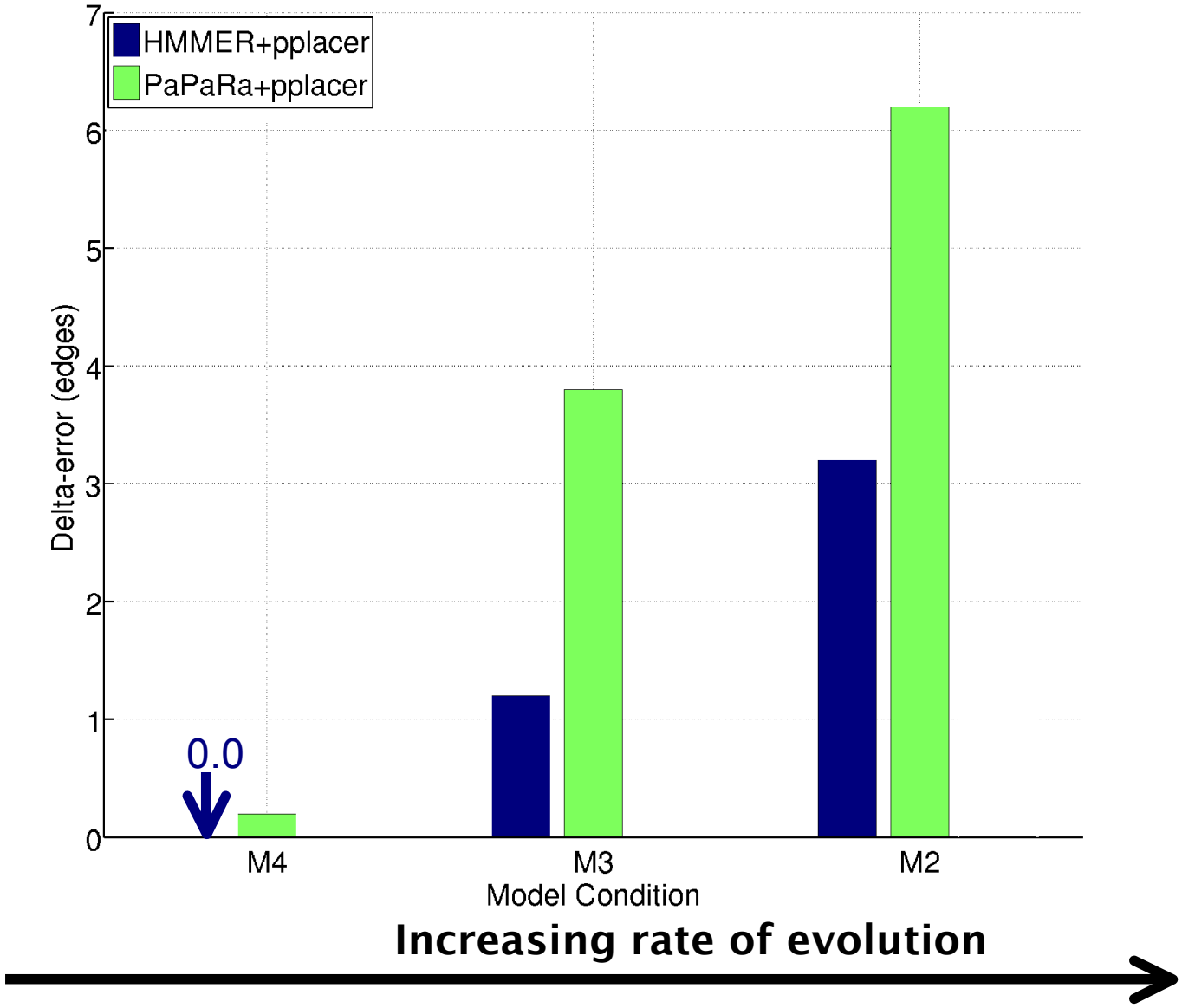


Phylogenetic Placement

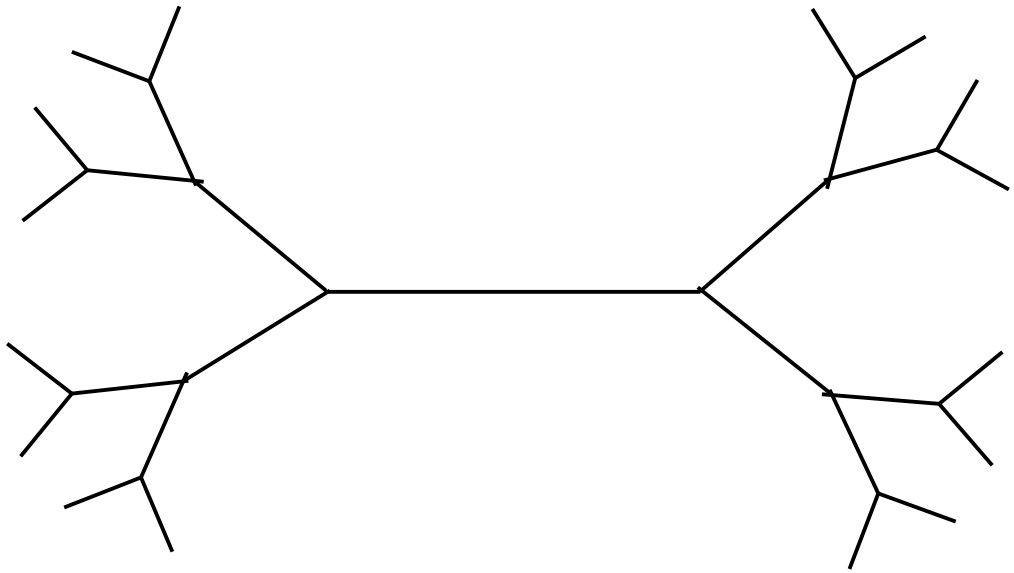
- Align each query sequence to backbone alignment
  - HMMALIGN (Eddy, Bioinformatics 1998)
  - PaPaRa (Berger and Stamatakis, Bioinformatics 2011)
- Place each query sequence into backbone tree
  - Pplacer (Matsen et al., BMC Bioinformatics, 2011)
  - EPA (Berger and Stamatakis, Systematic Biology 2011)

Note: pplacer and EPA use maximum likelihood

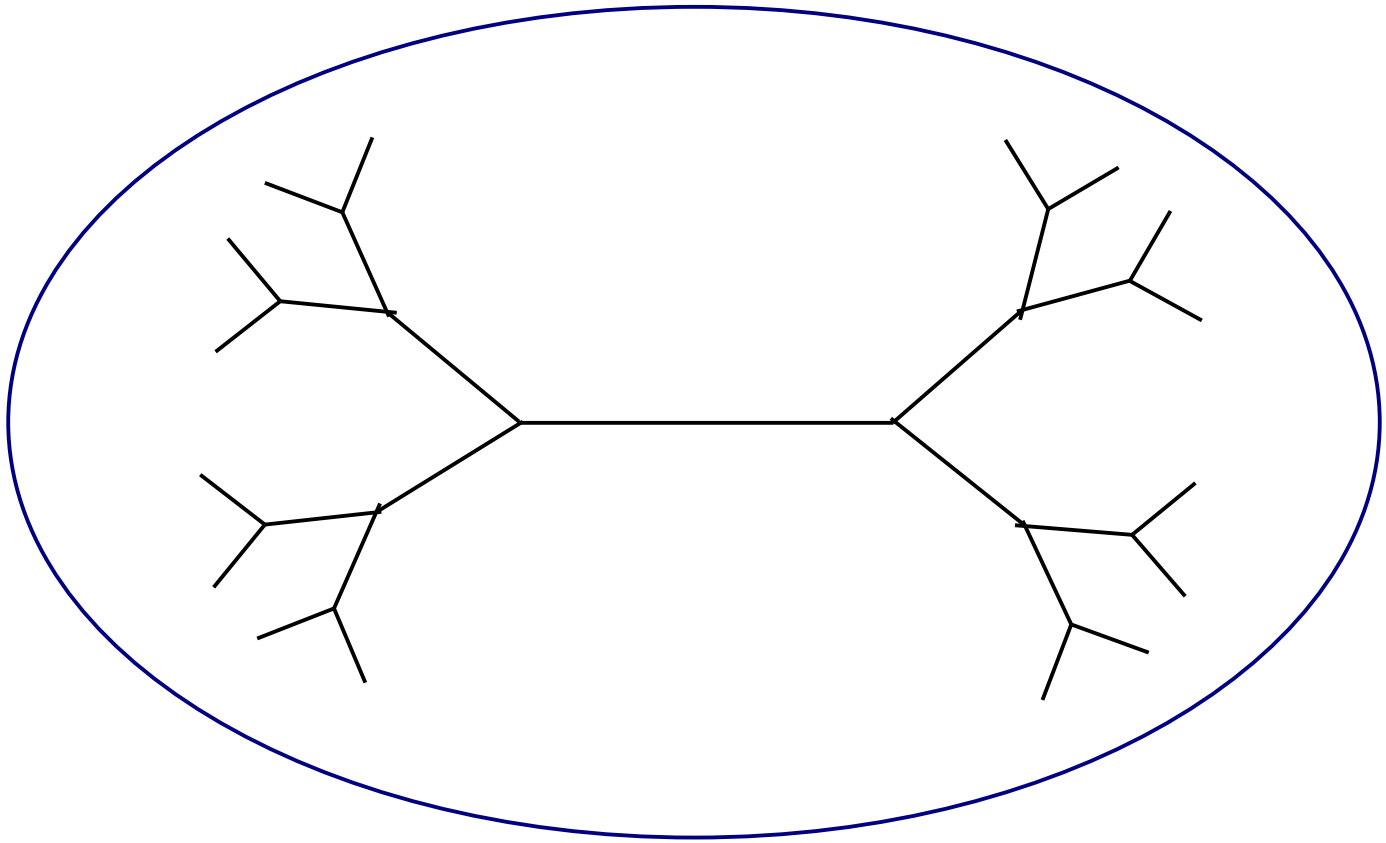
HMMER vs. PaPaRa



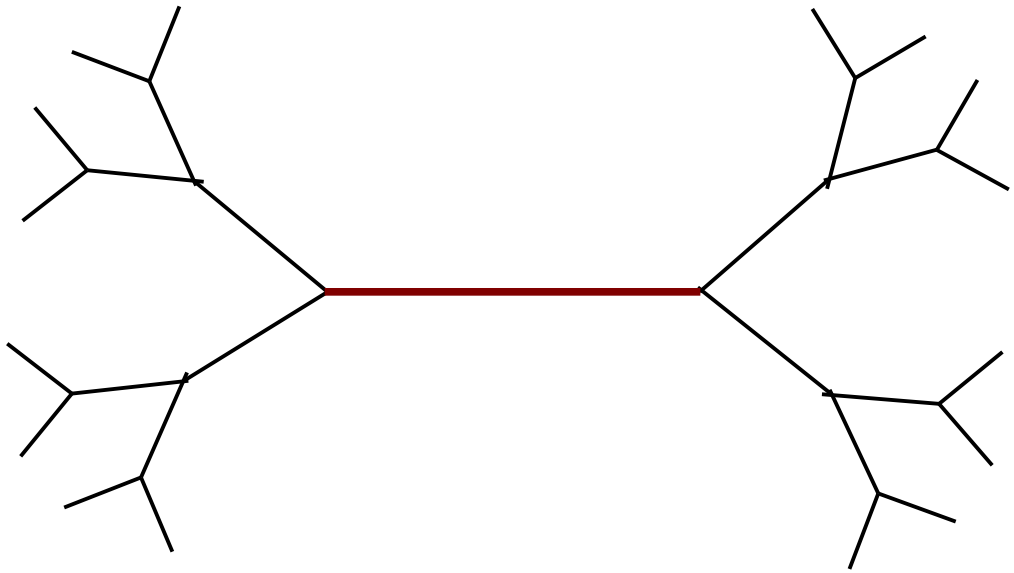
Insights from SATé



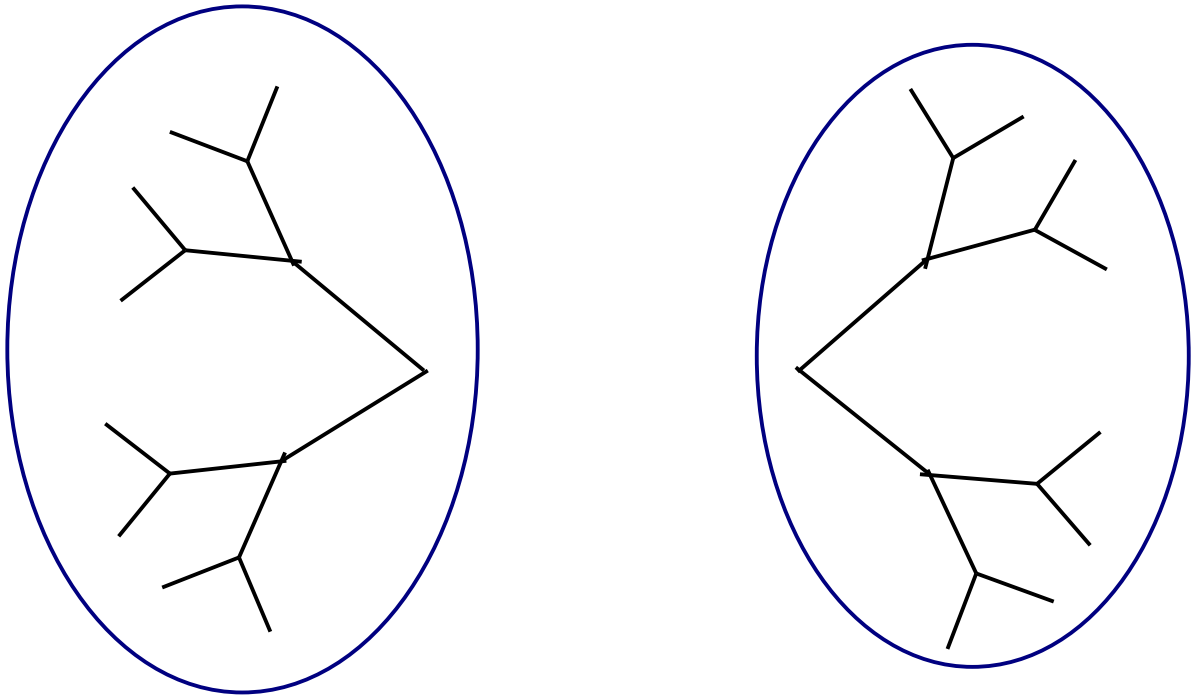
Insights from SATé



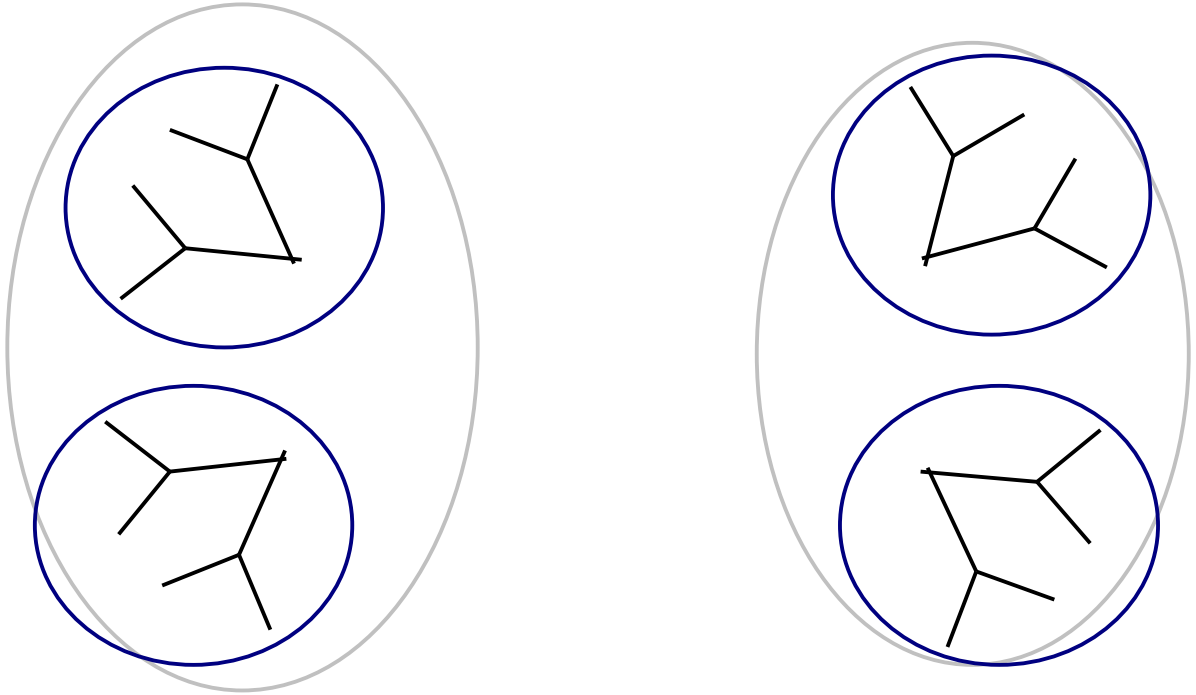
Insights from SATé



Insights from SATé



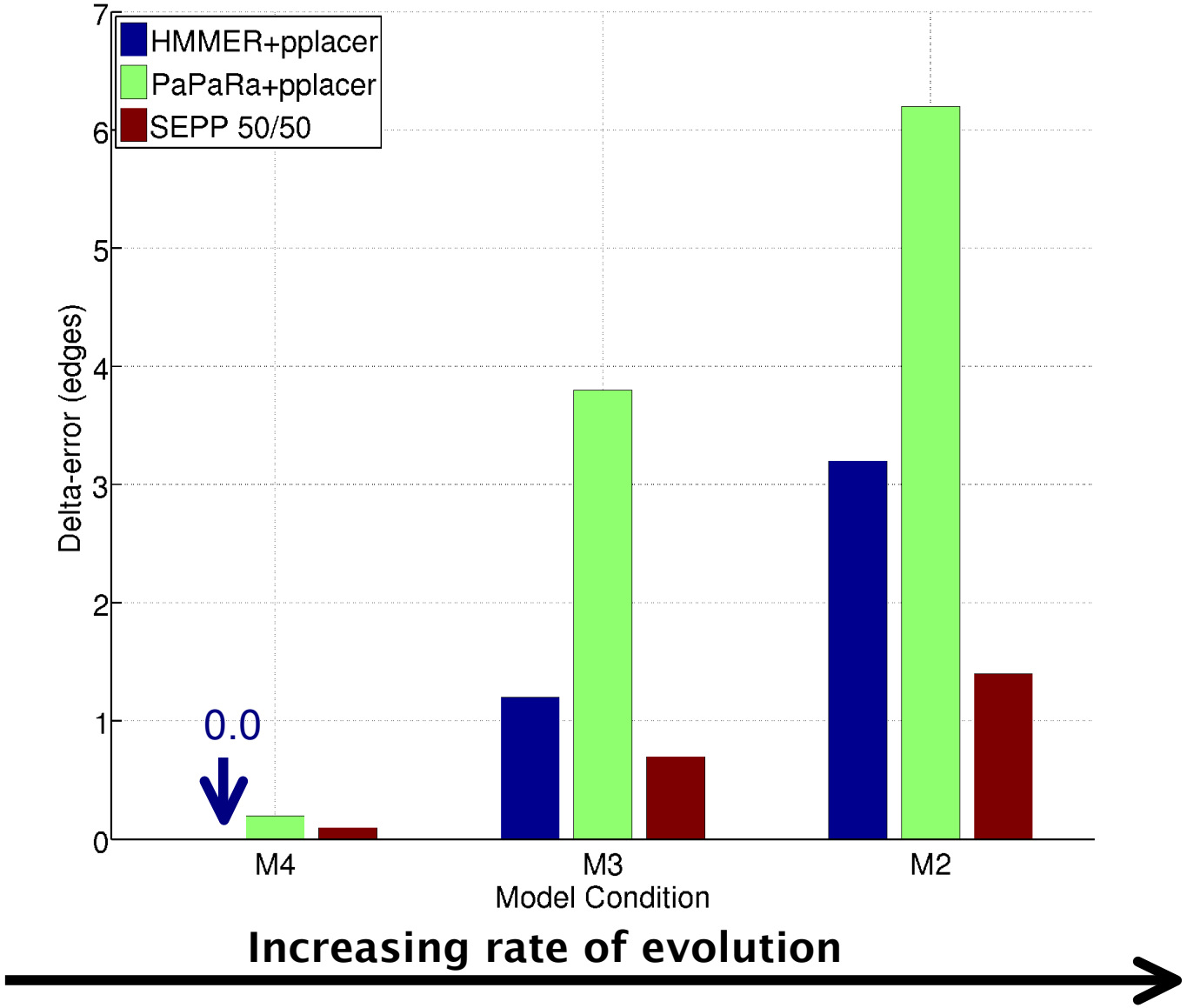
Insights from SATé



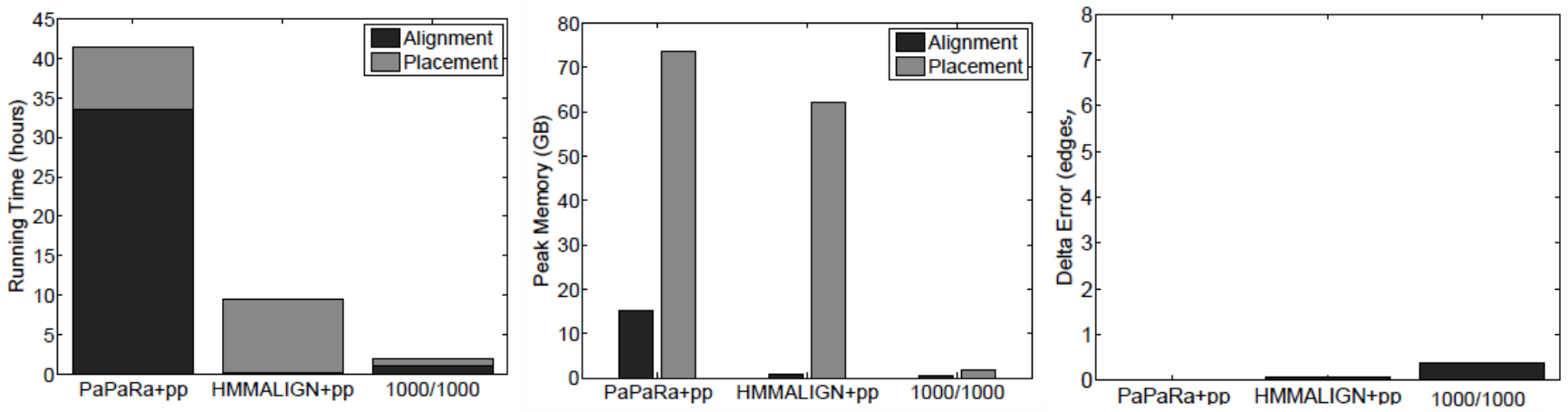
SEPP Parameter Exploration

- Alignment subset size and placement subset size impact the accuracy, running time, and memory of SEPP
- 10% rule (subset sizes 10% of backbone) had best overall performance

SEPP (10%-rule) on simulated data

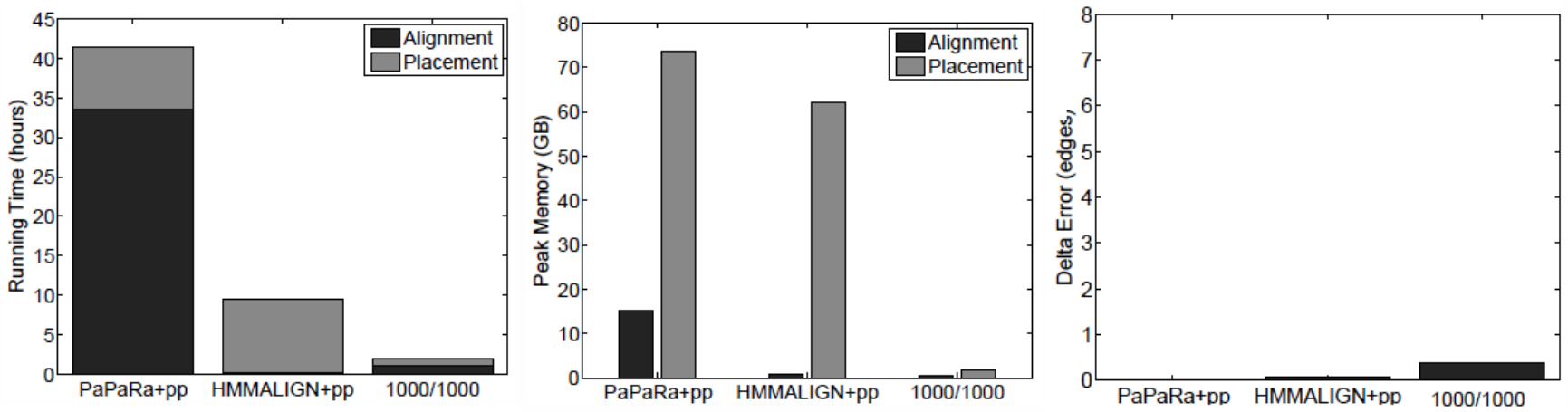


SEPP (10%) on Biological Data



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

SEPP (10%) on Biological Data



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

For 1 million fragments:

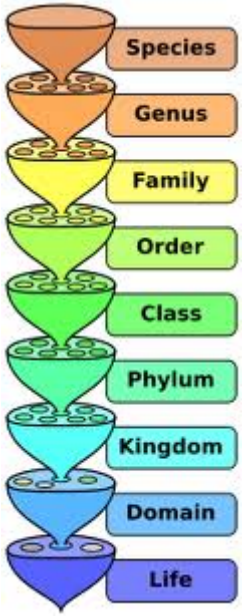
PaPaRa+pplacer: ~133 days

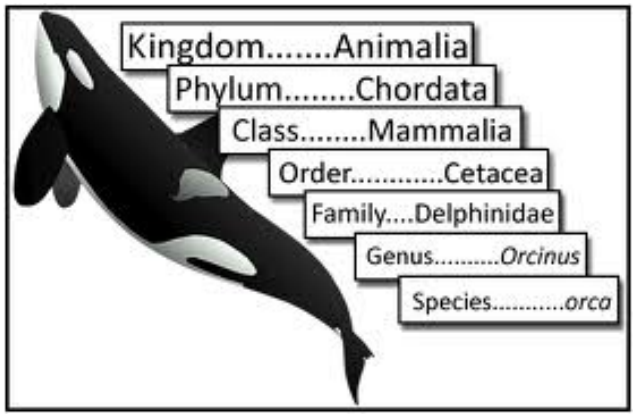
HMMALIGN+pplacer: ~30 days

SEPP 1000/1000: ~6 days

Part III: Taxon Identification

Objective: identify the taxonomy (species, genus, etc.) for each short read (a classification problem)





Taxon Identification

- Objective: identify species, genus, etc., for each short read

- Leading methods: Metaphyler (Univ Maryland), Phylopythia, PhymmBL, Megan

Megan vs MetaPhyler on 60bp error-free reads from rpsB gene

OBSERVATIONS

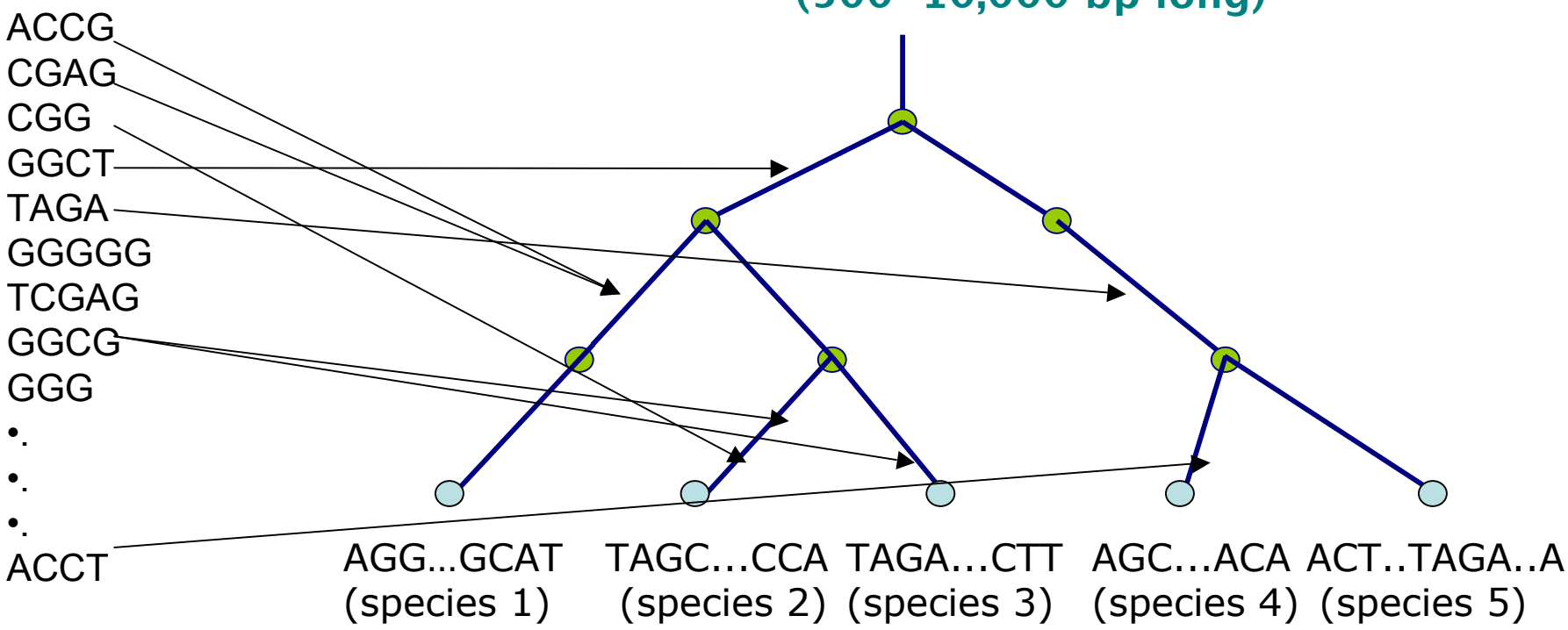
- MEGAN is very conservative
- MetaPhyler makes more correct predictions than MEGAN
- Other methods not as sensitive on these 31 marker genes as MetaPhyler (see MetaPhyler study in Liu et al, BMC Bioinformatics 2011)

Thus, the best taxon identification methods have **high precision** (few false positives), but **low sensitivity** (i.e., they **fail to classify** a large portion of reads) even at higher taxonomy levels.

TIPP: Taxon Identification using Phylogenetic Placement

Fragmentary Unknown Reads:
(60–200 bp long)

Estimated alignment and tree (gene tree or taxonomy) on known full length sequences
(500–10,000 bp long)



TIPP - Version 1

Given a set Q of query sequences for some gene, a taxonomy T*, and a set of full-length sequences for the gene,

- Compute backbone alignment/tree pair (T,A) on the full-length sequences, using SATé
- Use SEPP to place query sequence into T*
  - Compute extended alignment for each query sequence, using (T,A)
  - Place query sequence into T* using pplacer (maximizing likelihood score)

But … *TIPP version 1 too aggressive (over-classifies)*

TIPP version 2:

Use statistical support to reduce over-classification:

- Find 2 or more backbone alignment/tree pairs of full-length sequences
- For each backbone alignment/tree pair, produce many extended alignments using **HMMER statistical support**
- For each extended alignment, use **pplacer statistical support** to place fragment within taxonomy
- Classify each fragment at the **LCA** of all placements obtained for the fragment

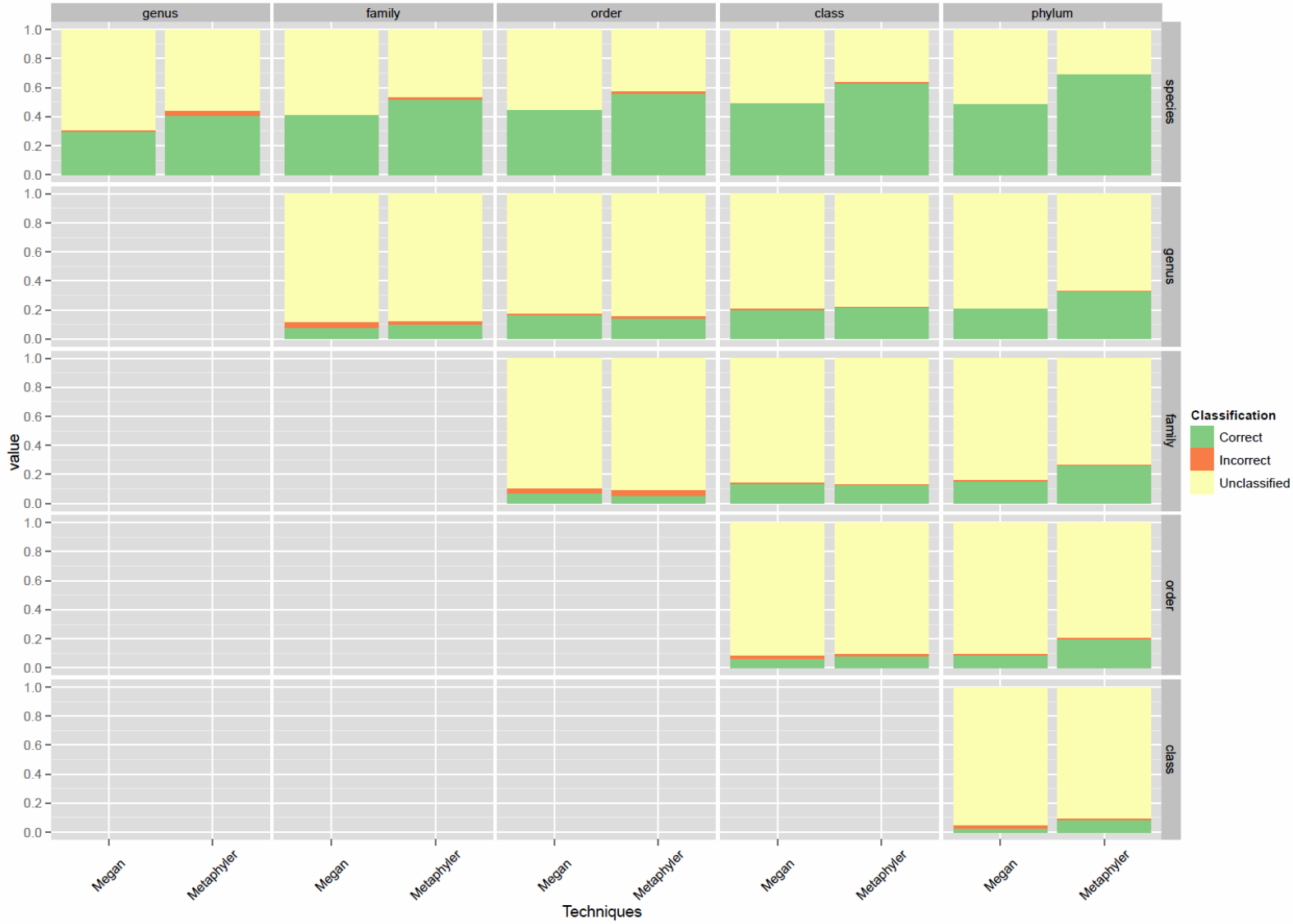
TIPP version 2 dramatically reduces false positive rate with small reduction in true positive rate by considering uncertainty, using statistical techniques.

Experiments

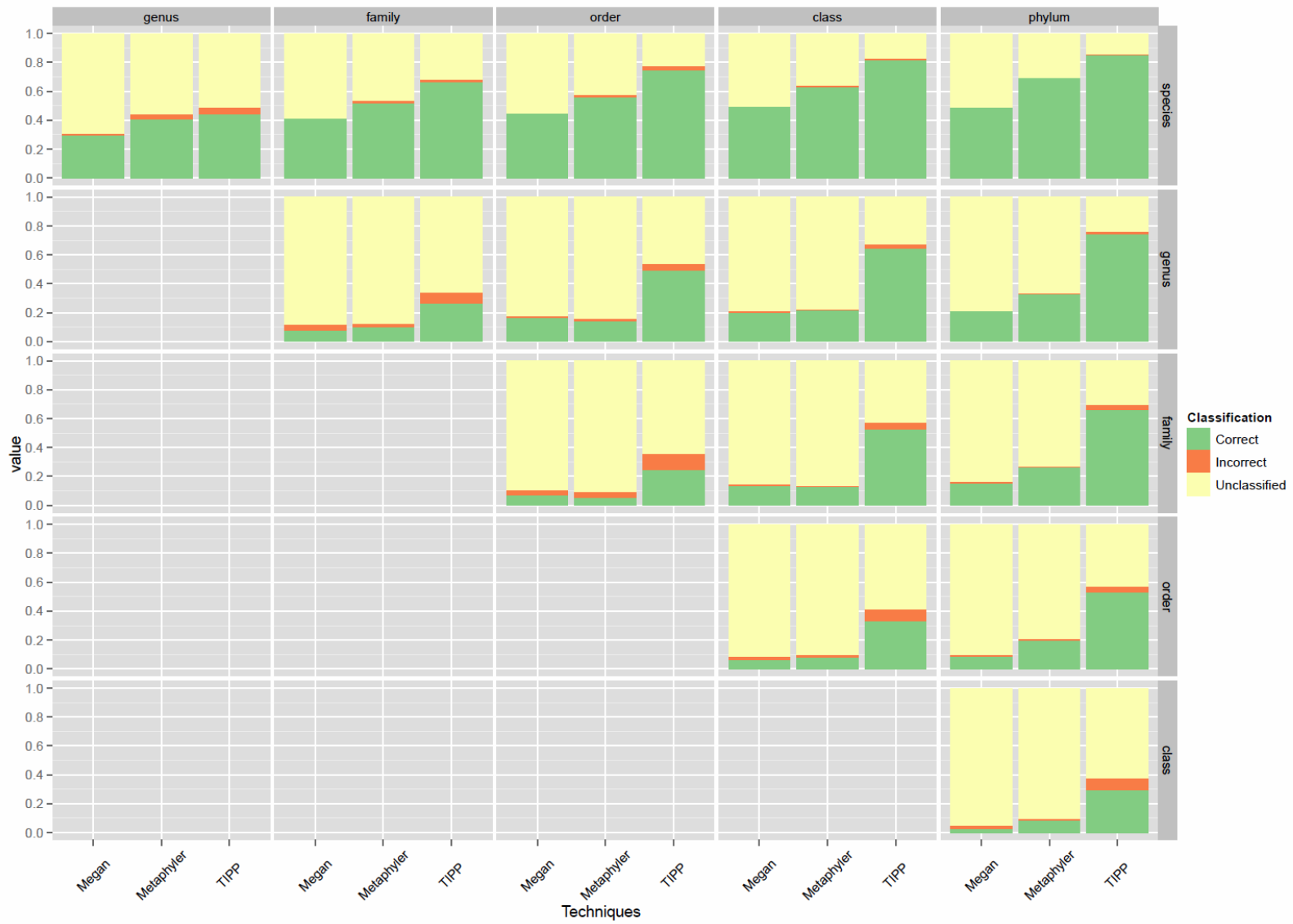
Tested TIPP and Metaphyler on marker genes:

- Leave-one-out experiments on 60bp reads without error
- Leave-one-out experiments on 100bp reads with simulated Illumina errors
- Leave-one-out experiments on 300bp reads with simulated 454 errors (1% error rate with indels and substitutions)

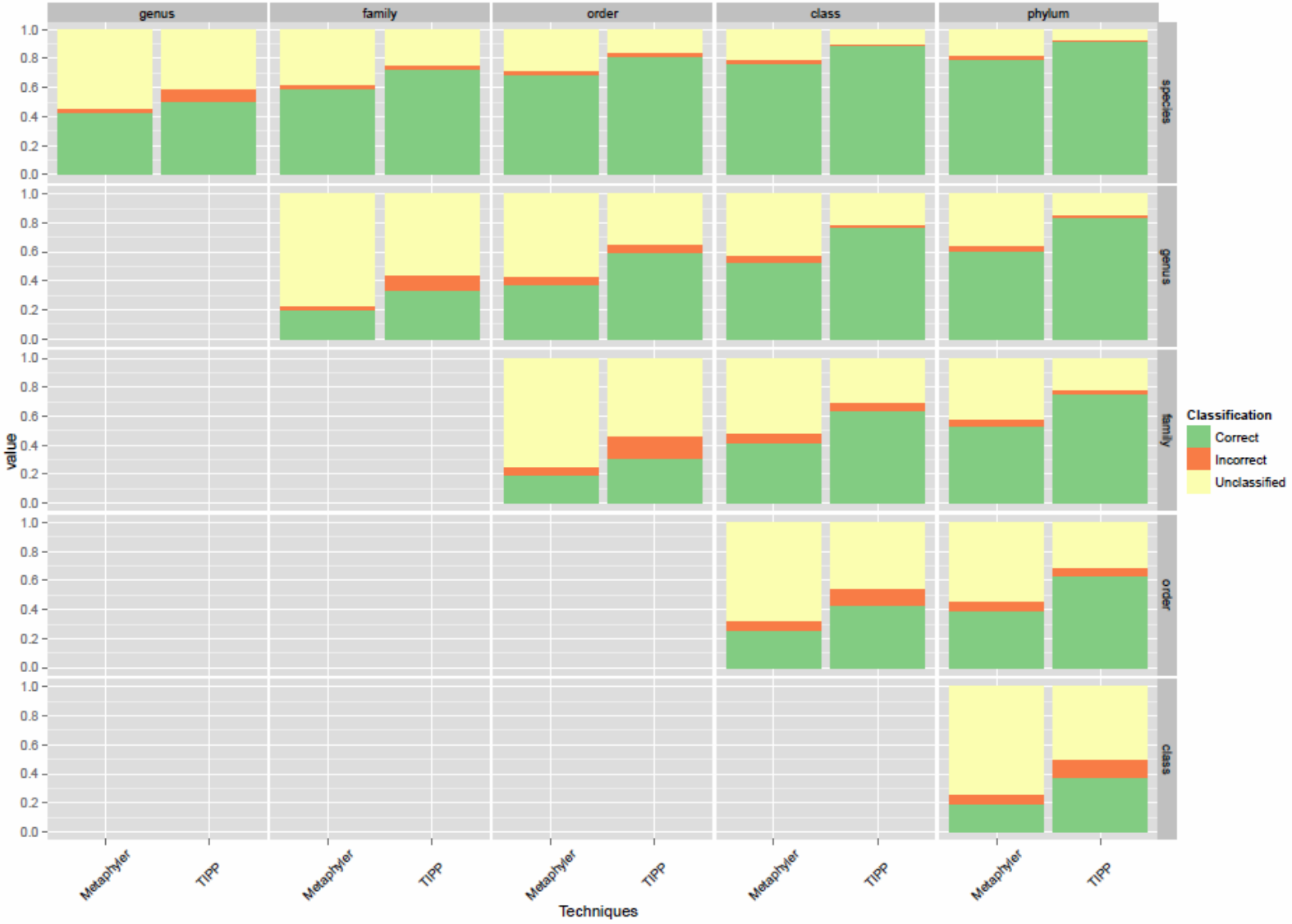
60bp error-free reads on rpsB marker gene



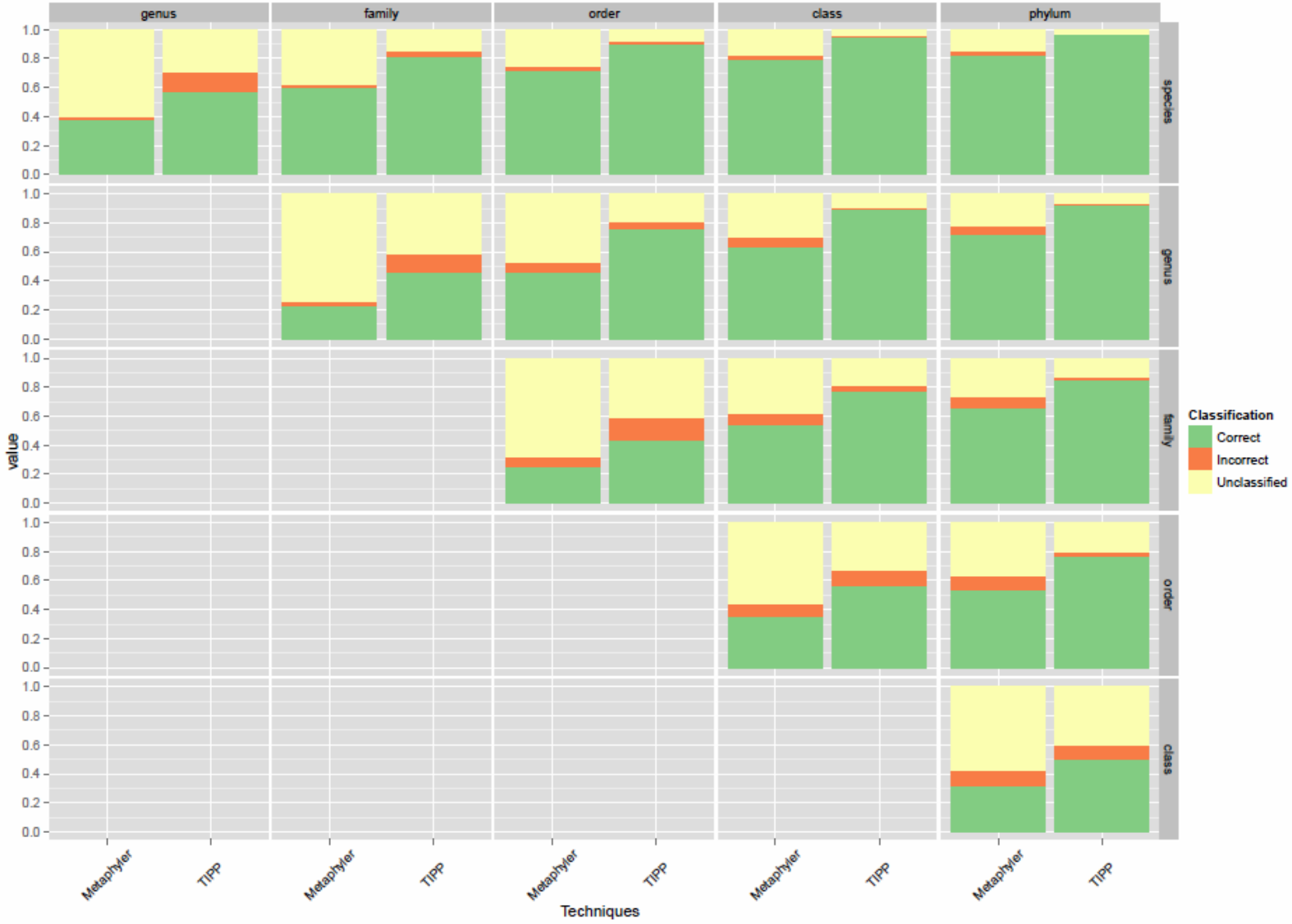
60bp error-free reads on rpsB marker gene



MetaPhyler versus TIPP on 100bp Illumina reads across 29 marker genes



MetaPhyler versus TIPP on 300bp 454 reads across 29 marker genes



Summary

- SATé gives better alignments and trees
- SEPP yields improved alignment of short (fragmentary) sequences into alignments of full-length sequences, and results in more accurate phylogenetic placement
- TIPP gives improved taxon identification of short reads
- *Key insight: improved alignment through careful divide-and-conquer*

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)
- SEPP-boosting for metagenomic analyses (2012)
- DACTAL-boosting for all phylogeny estimation methods (in prep)

Overall message

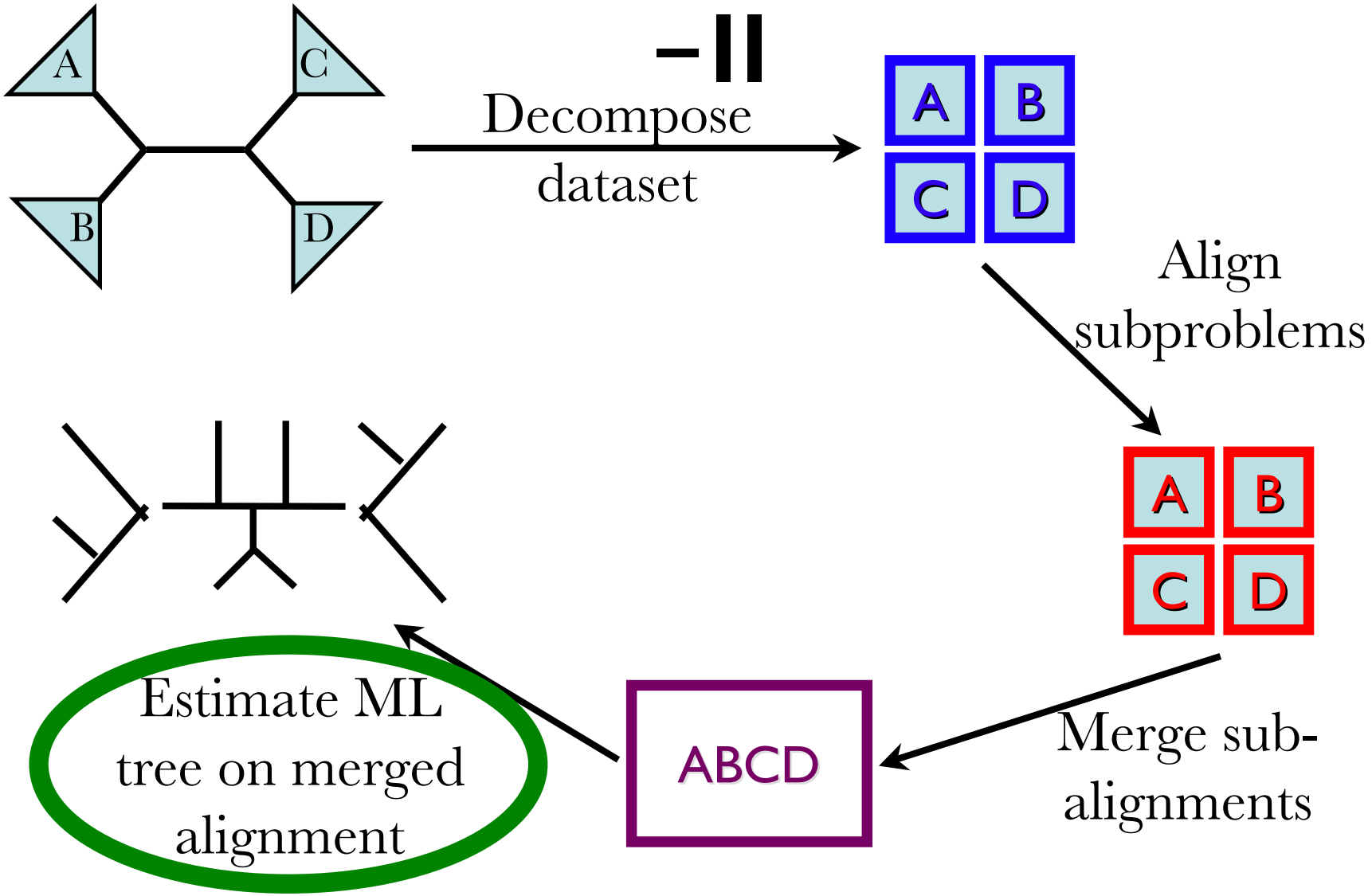
- When data are difficult to analyze, develop better methods - don’t throw out the data.

Acknowledgments

- Guggenheim Foundation Fellowship, Microsoft Research New England, National Science Foundation: Assembling the Tree of Life (ATOL), ITR, and IGERT grants, and David Bruton Jr. Professorship
- NSERC support to Siavash Mirarab

- Collaborators:
  - SATé: Kevin Liu, Serita Nelesen, Sindhu Raghavan, and Randy Linder
  - SEPP/TIPP: Siavash Mirarab and Nam Nguyen

Limitations of SATé-I and -II

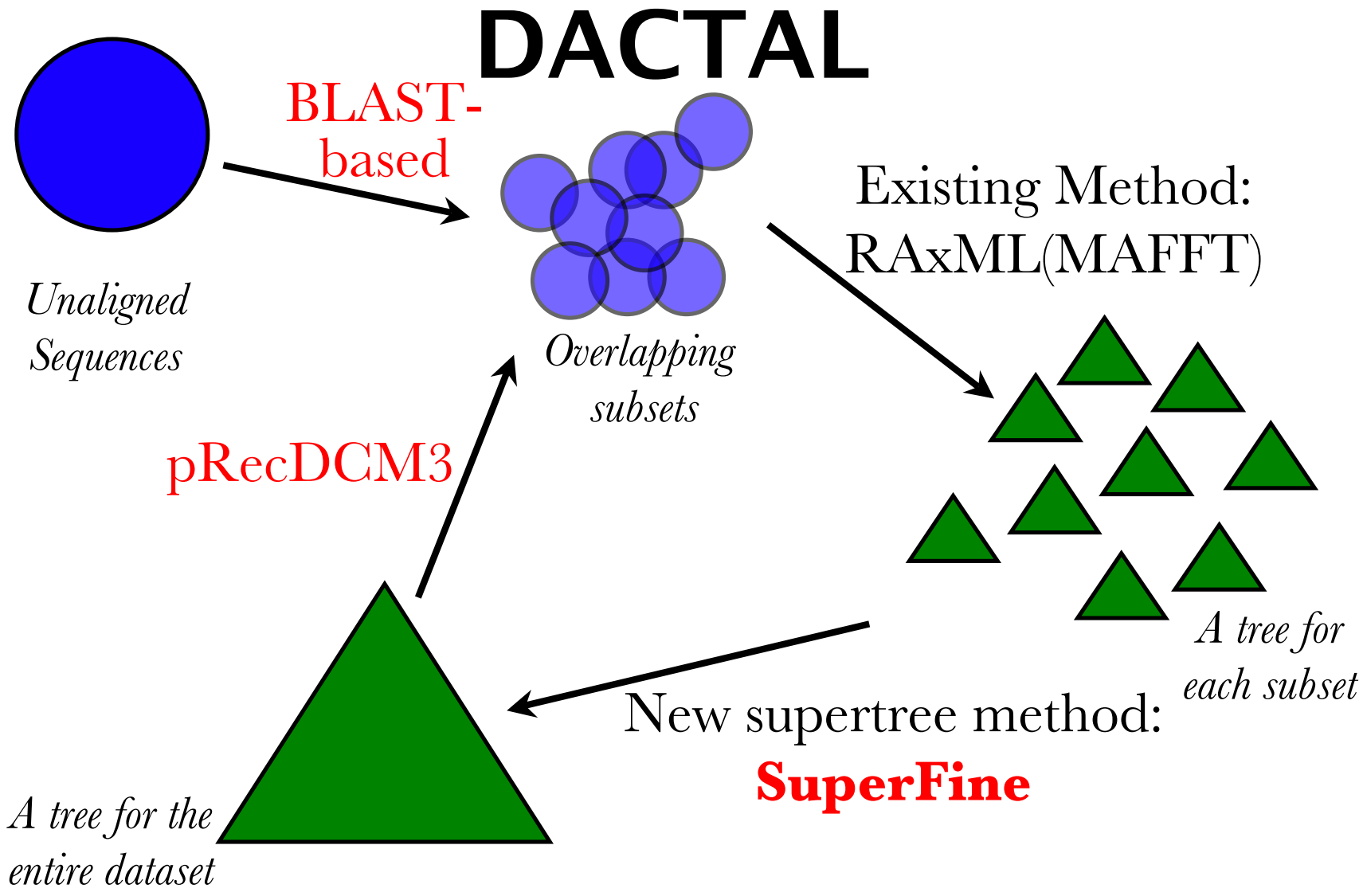


Part II: DACTAL

(Divide-And-Conquer Trees (without) ALignments)

- Input: set S of unaligned sequences
- Output: tree on S (but no alignment)

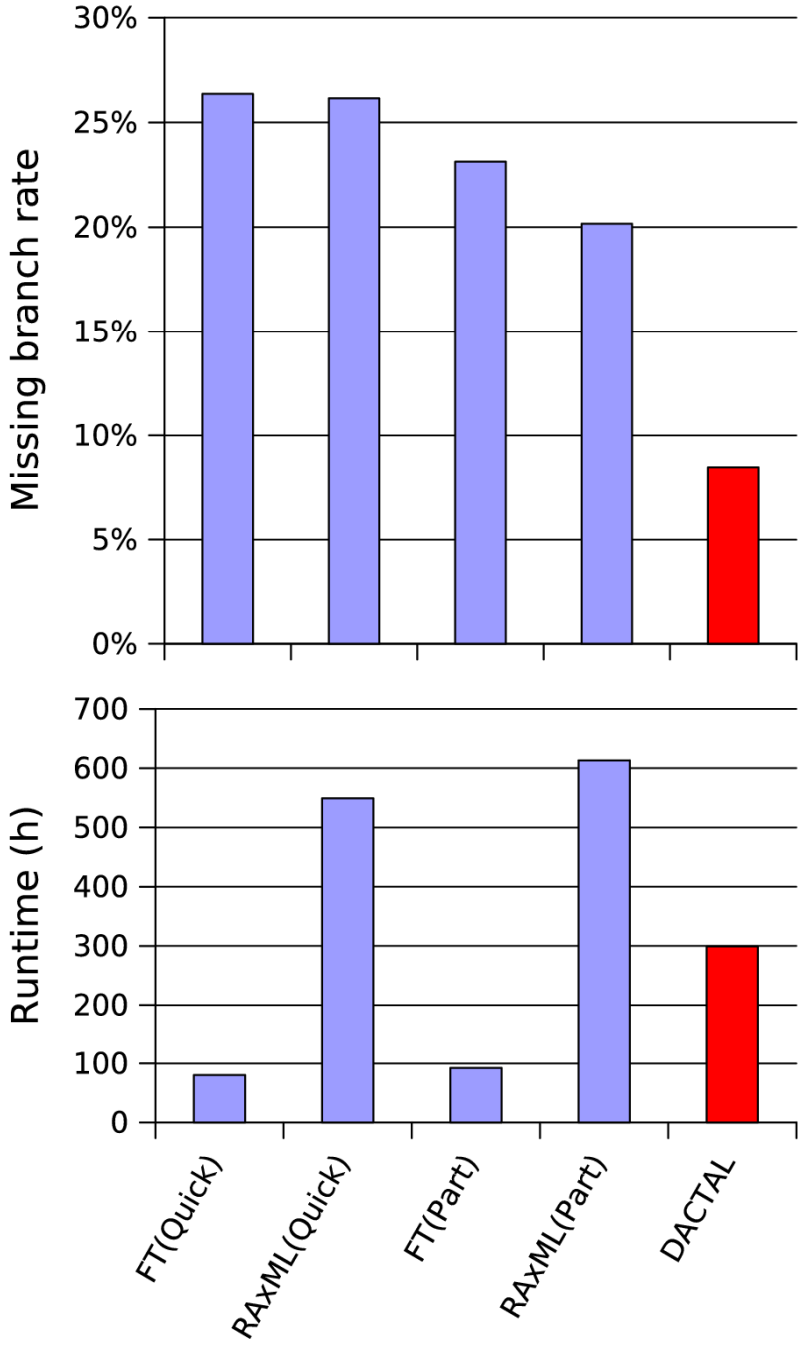
(Nelesen, Liu, Wang, Linder, and Warnow, in preparation)


DACTAL
BLAST-based
Unaligned Sequences
Overlapping subsets
Existing Method: RAxML(MAFFT)
A tree for each subset
New supertree method:
SuperFine
pRecDCM3
A tree for the entire dataset

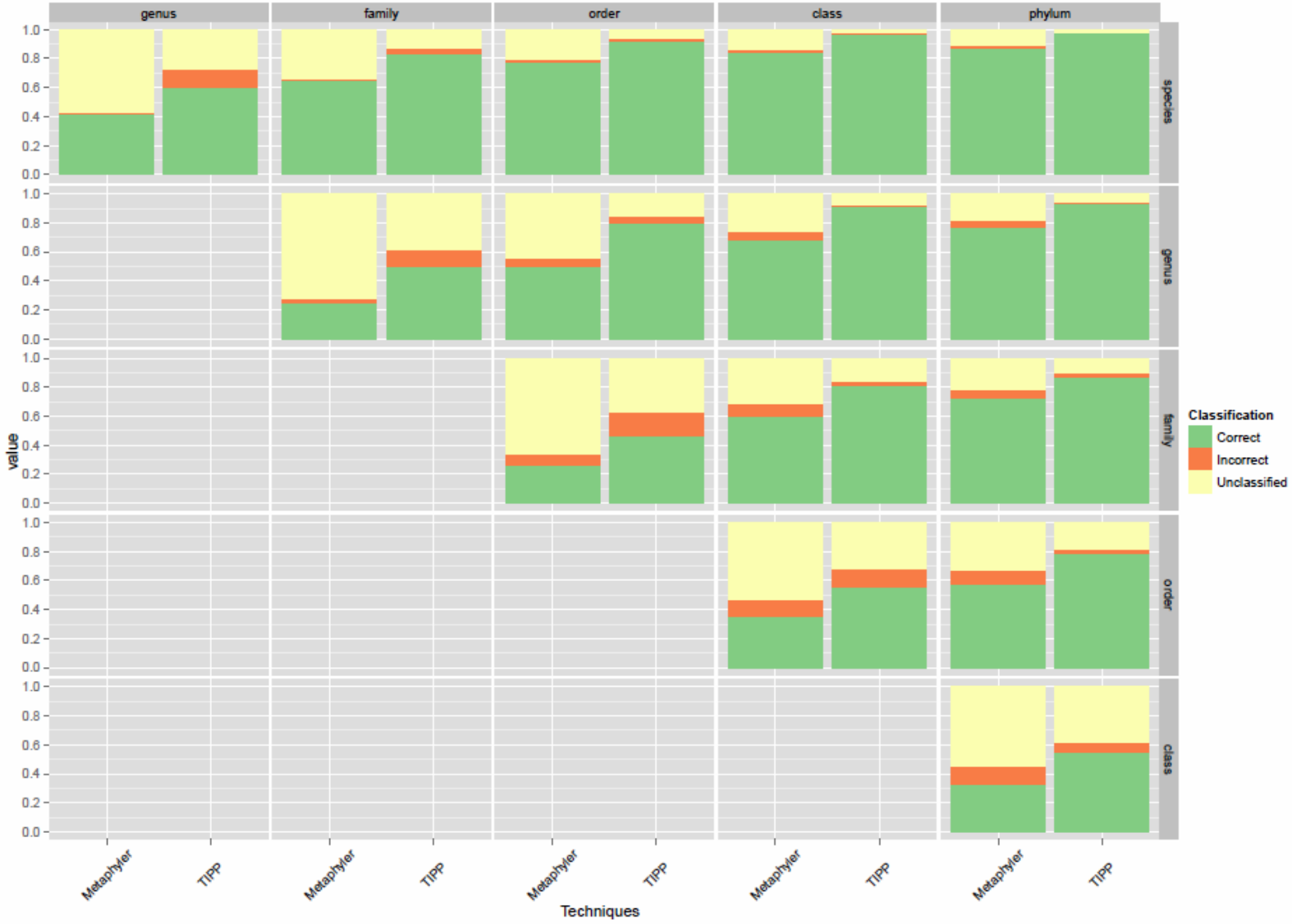
Average of 3 Largest CRW Datasets

- CRW: Comparative RNA database, datasets 16S.B.ALL, 16S.T, and 16S.3
- 6,323 to 27,643 sequences
- These datasets have curated alignments based on secondary structure
- Reference trees are 75% RAxML bootstrap trees

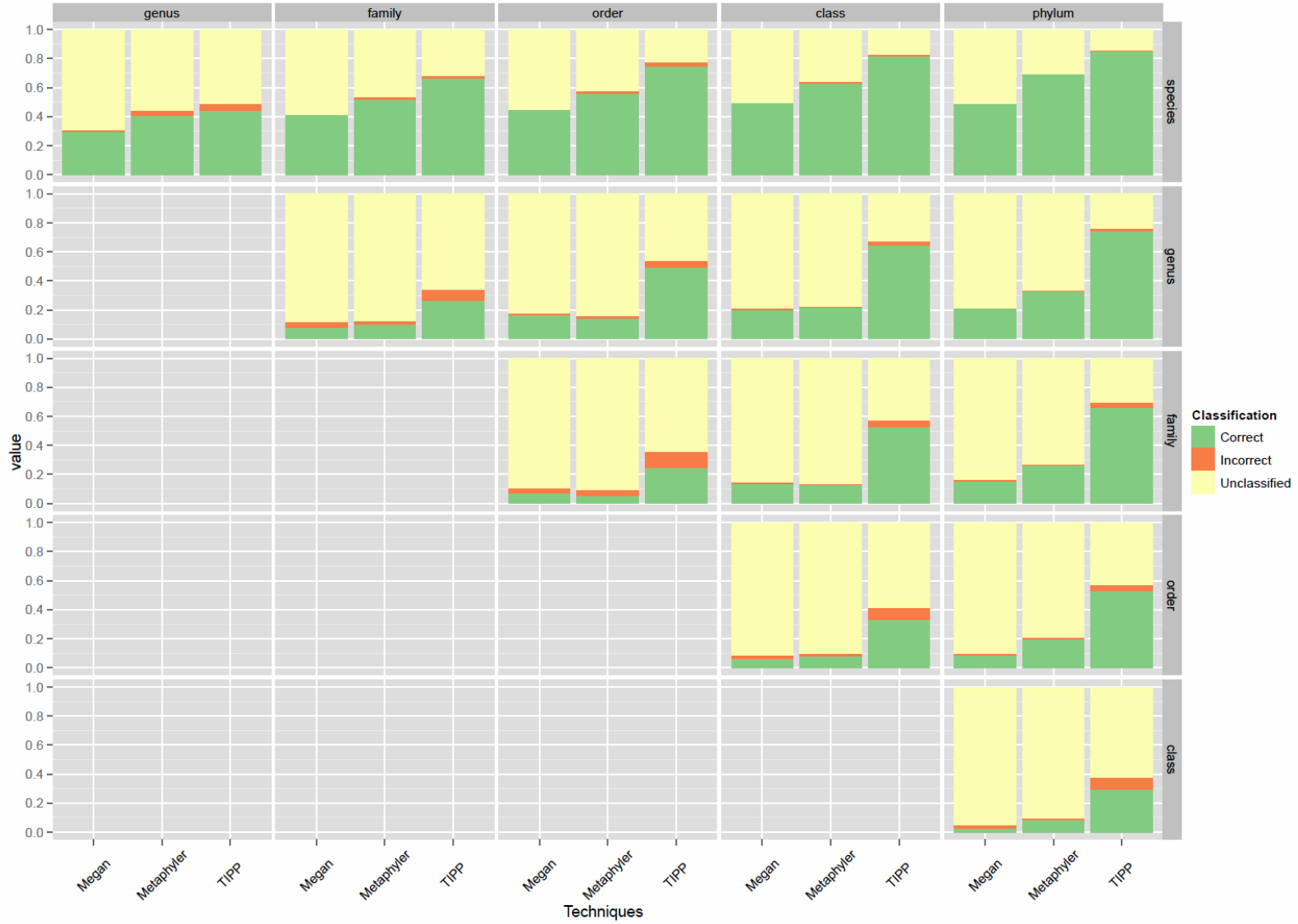
- DACTAL (shown in red) run for 5 iterations starting from FT(Part)
- DACTAL is robust to starting trees
- PartTree and Quicktree are the only MSA methods that run



MetaPhyler versus TIPP on 300bp 454 reads across on rpsB marker gene



60bp error free reads on rpsB marker gene

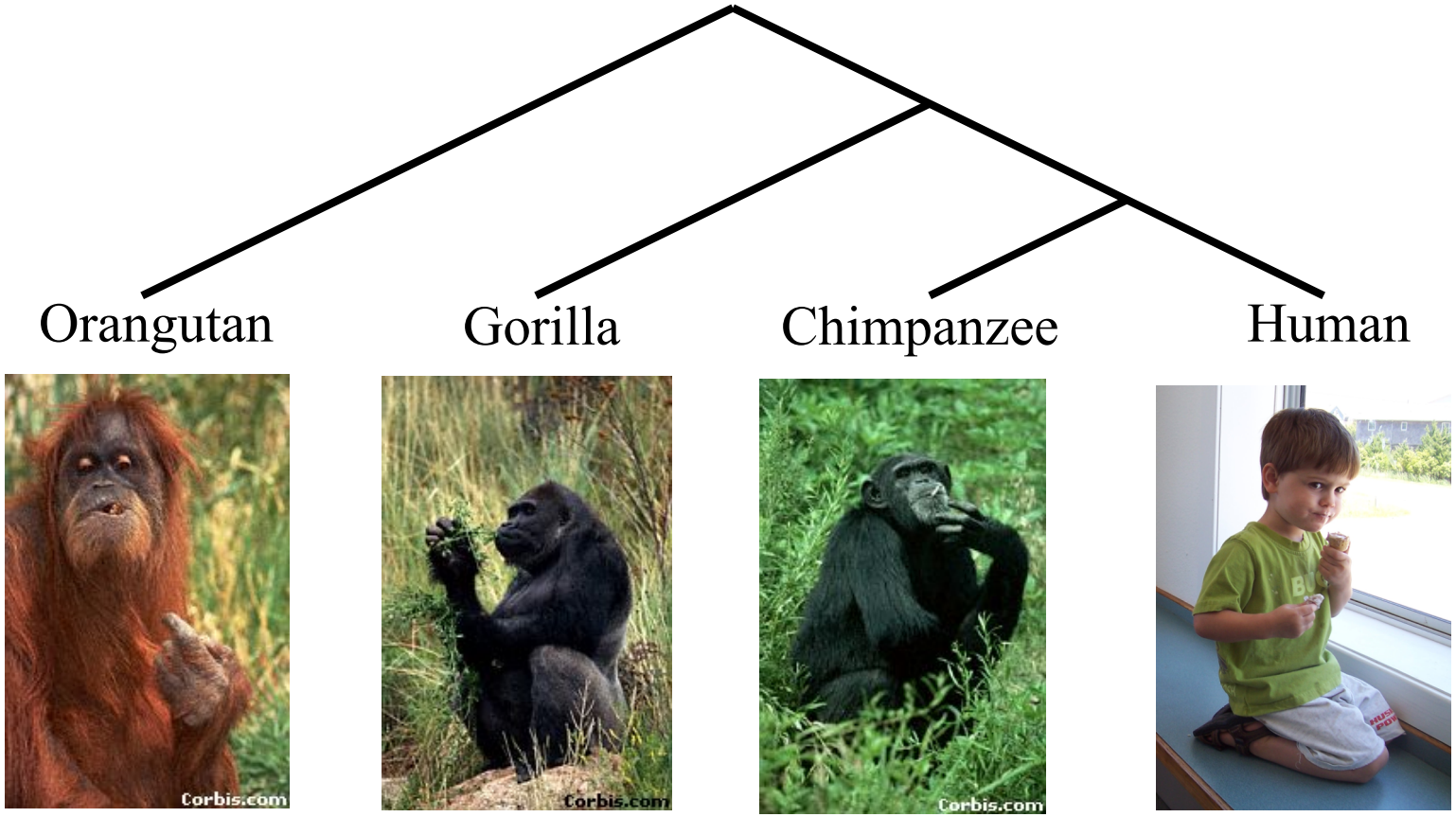


Phylogenetic "Boosters"

- SATé: co-estimation of alignments and trees
- SEPP/TIPP: phylogenetic analysis of fragmentary data

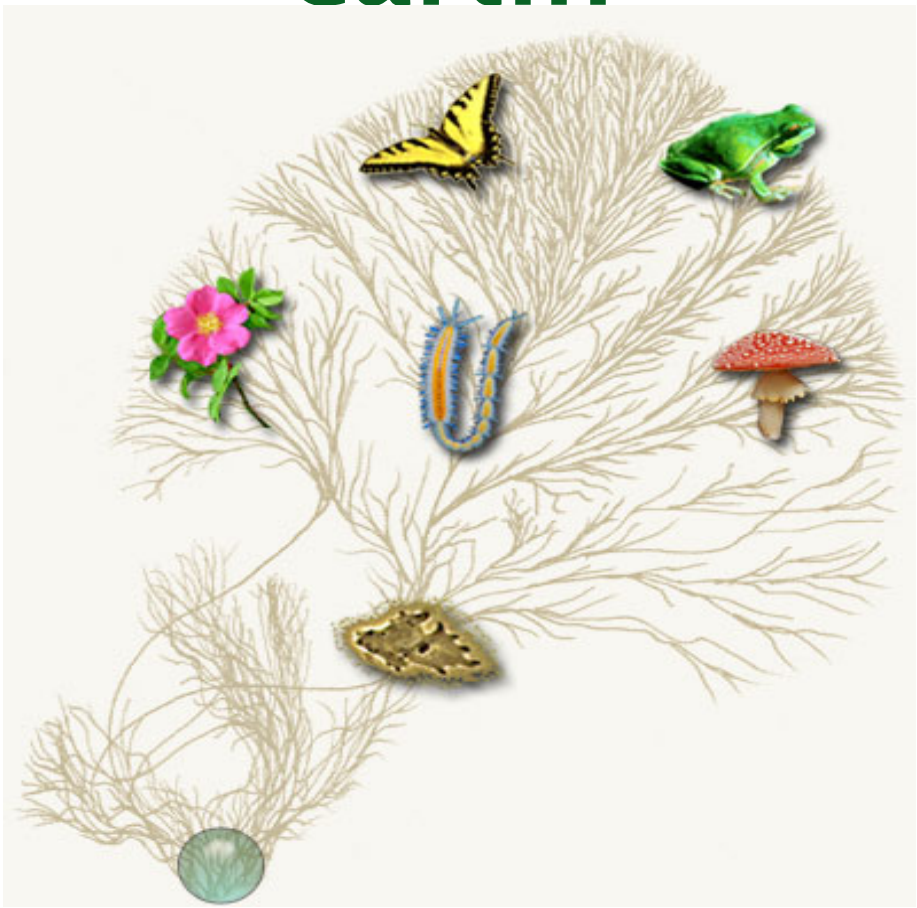
Algorithmic strategies: divide-and-conquer and iteration to improve the accuracy and scalability of a *base method*

Phylogeny (evolutionary tree)



From the Tree of the Life Website,
University of Arizona

How did life evolve on earth?



Courtesy of the Tree of Life project

MetaPhyler versus TIPP on 100bp Illumina reads on rpsB marker gene