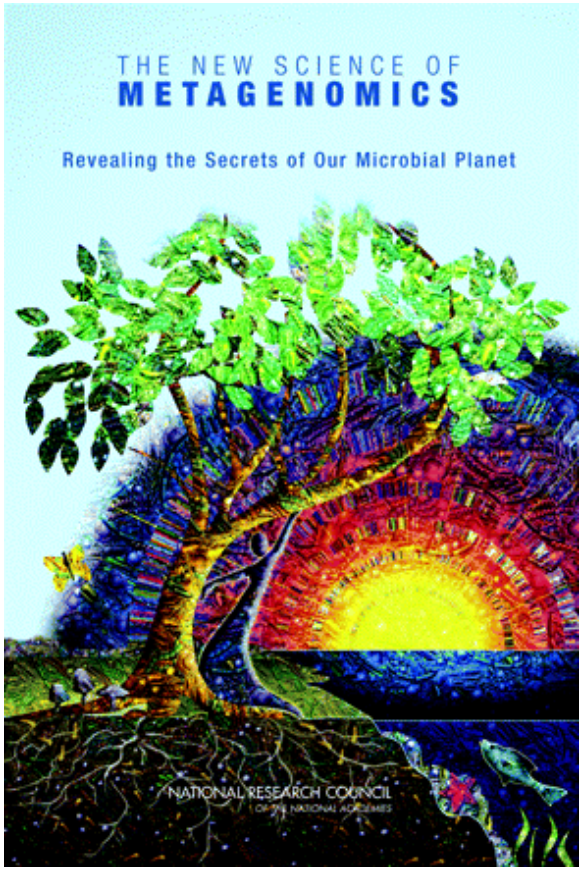
SEPP: SATé-enabled phylogenetic placement (and metagenomics)

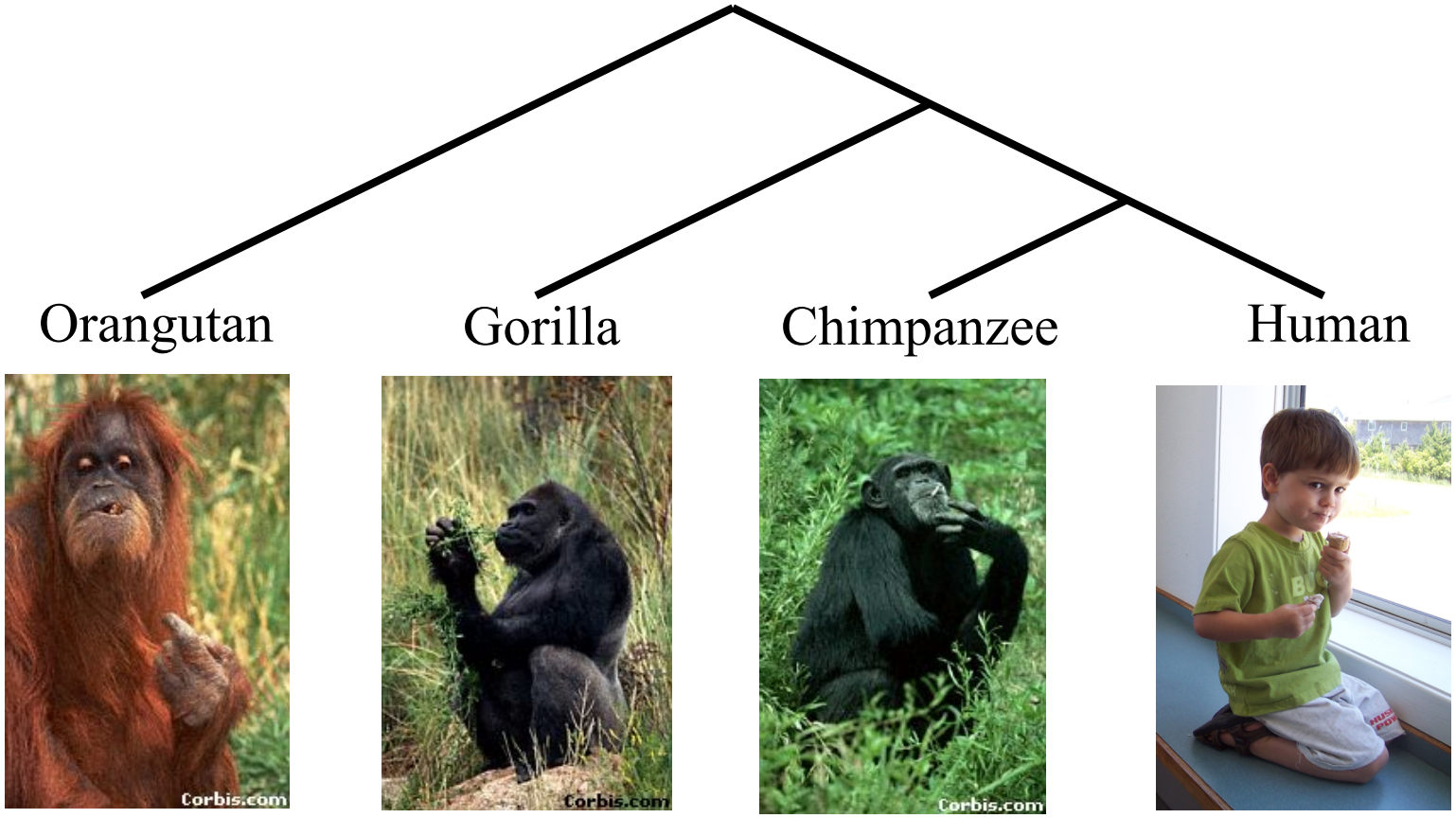
Tandy Warnow
Department of Computer Science
University of Texas

Computational Phylogenetics and Metagenomics



Courtesy of the Tree of Life project

Phylogeny (evolutionary tree)



From the Tree of the Life Website,
University of Arizona

The New Science of Metagenomics

"Metagenomics will generate knowledge of microbial interactions so that they can be harnessed to improve human health, food security, and energy production."

"Metagenomics combines the power of genomics, bioinformatics, and systems biology."

"Metagenomics will be the systems biology of the biosphere."

"There is no doubt that its concepts and methods will ultimately transform all biology."

National Academies Press

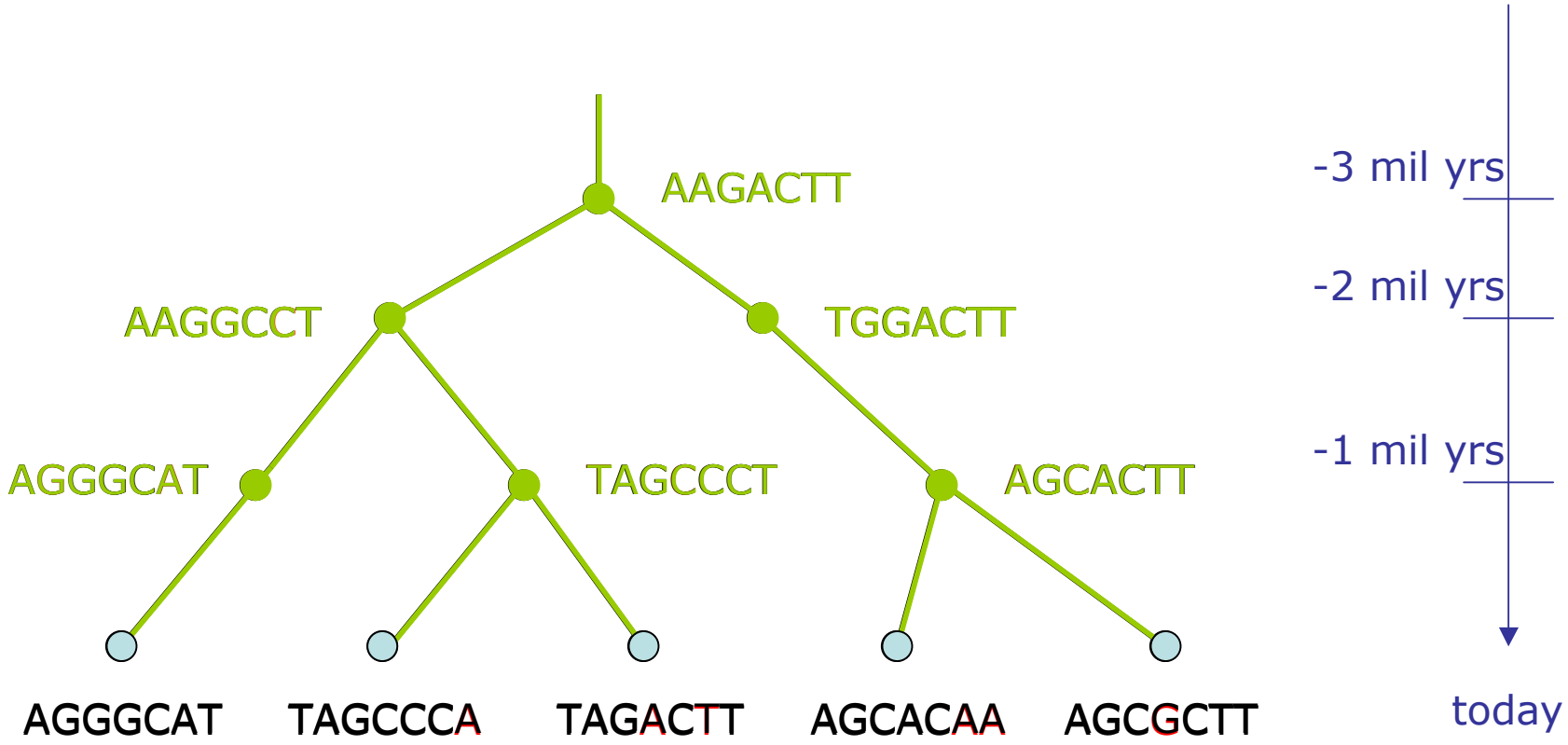
Basic questions

- Who is there?
- What are they doing?
- What is being done by the microbial community?

Major Challenges

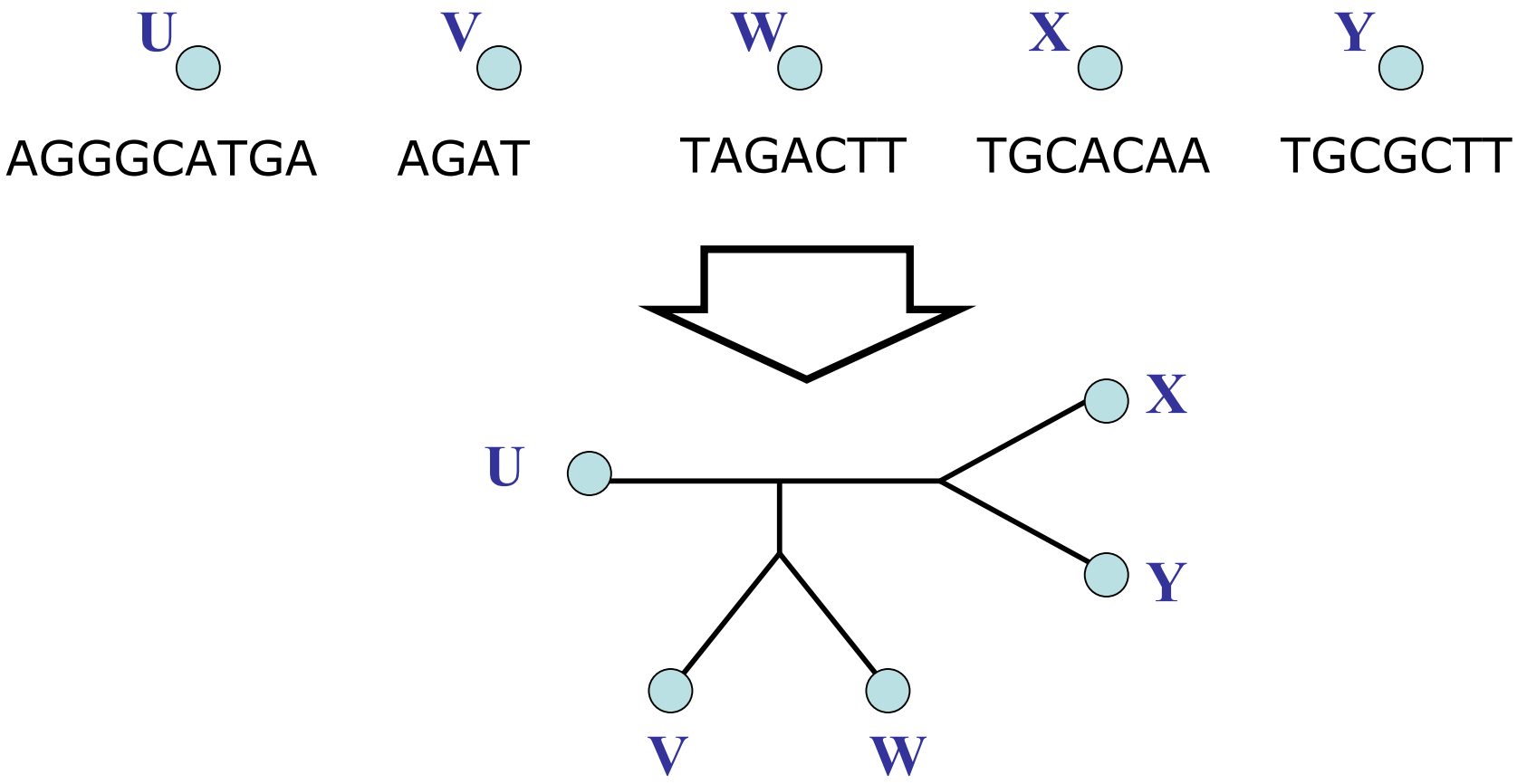
- **Phylogenetic analyses:** standard methods have *poor accuracy* on even moderately large datasets, and the most accurate methods are enormously *computationally intensive* (weeks or months, high memory requirements)
- **Metagenomic** analyses: methods for species classification of short reads have *poor sensitivity*. Efficient high throughput is necessary (millions of reads).

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

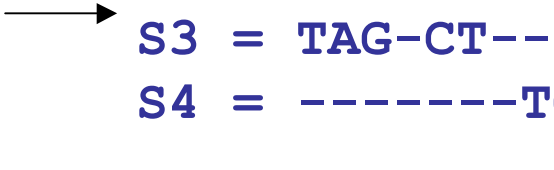
```
S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC
S3 = TAGCTGACCGC
S4 = TCACGACCGACA
```



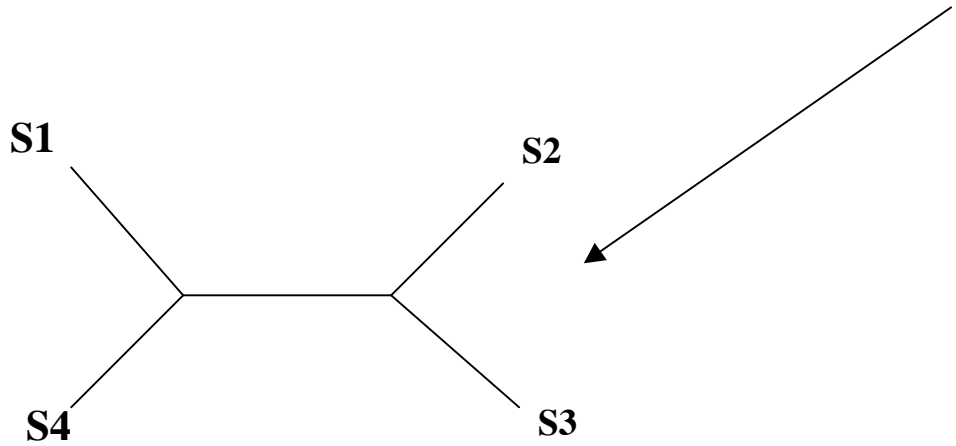
```
S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAC--GACCGC--
S3 = TAG-CT-------GACCGC--
S4 = --------TCAC--GACCGACA
```

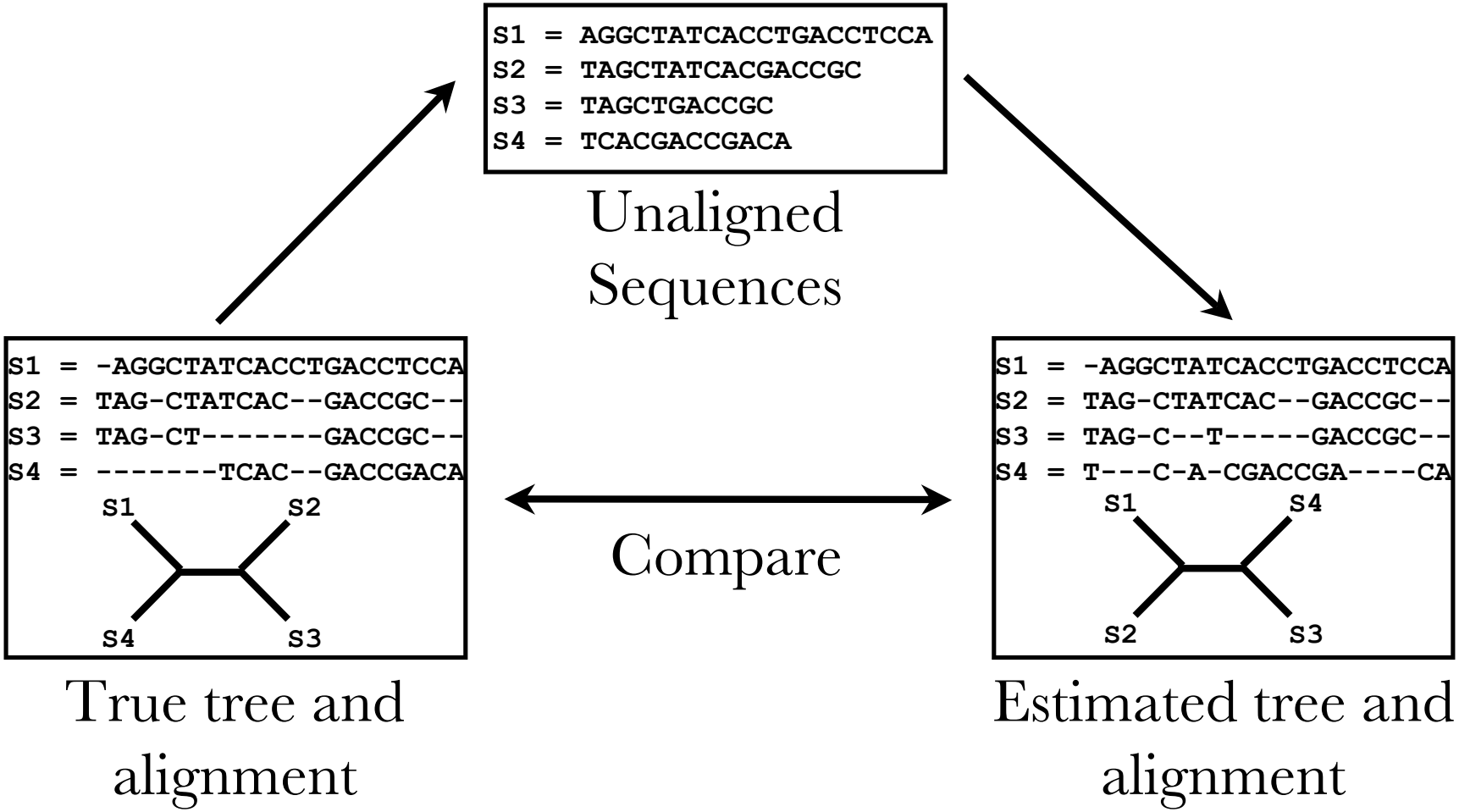
Phase 2: Construct tree

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

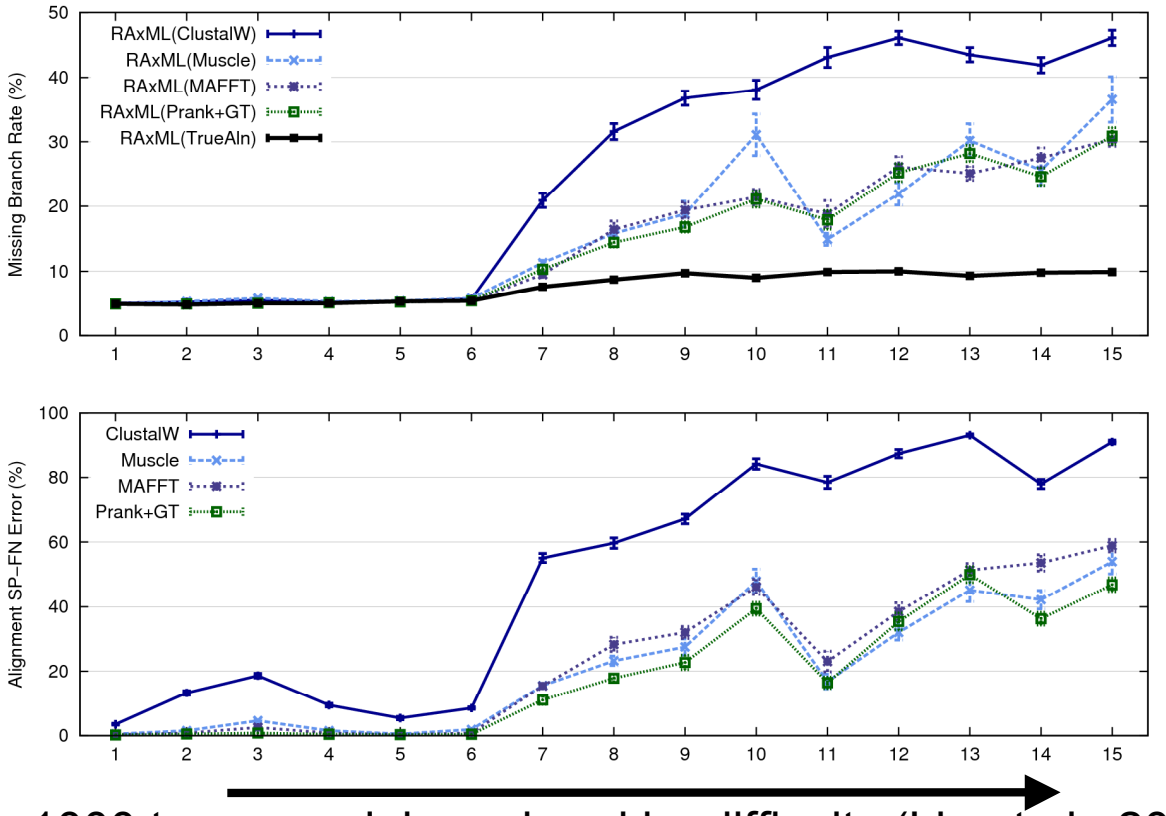


```
S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAC--GACCGC--
S3 = TAG-CT-------GACCGC--
S4 = -------TCAC--GACCGACA
```





Simulation Study Protocol



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)

Major Challenges

- Current phylogenetic datasets contain hundreds to thousands of taxa, with multiple genes.
- Future datasets will be substantially larger (e.g., iPlant plans to construct a tree on 500,000 plant species)
- *Current methods have poor accuracy or cannot run on large datasets.*

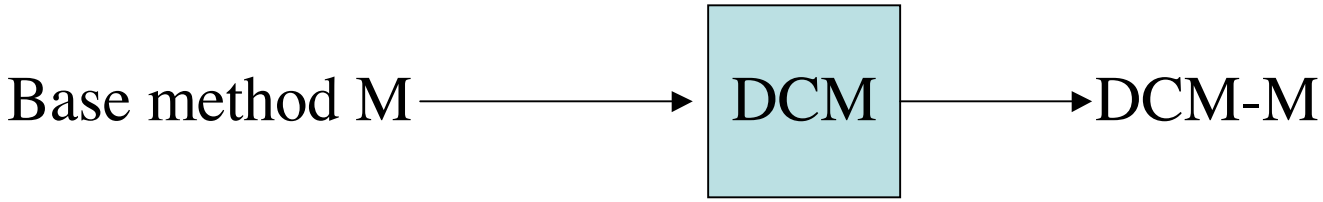
Phylogenetic “boosters”

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 (2011)

- DCMs “boost” the performance of phylogeny reconstruction methods.



Today's Talk

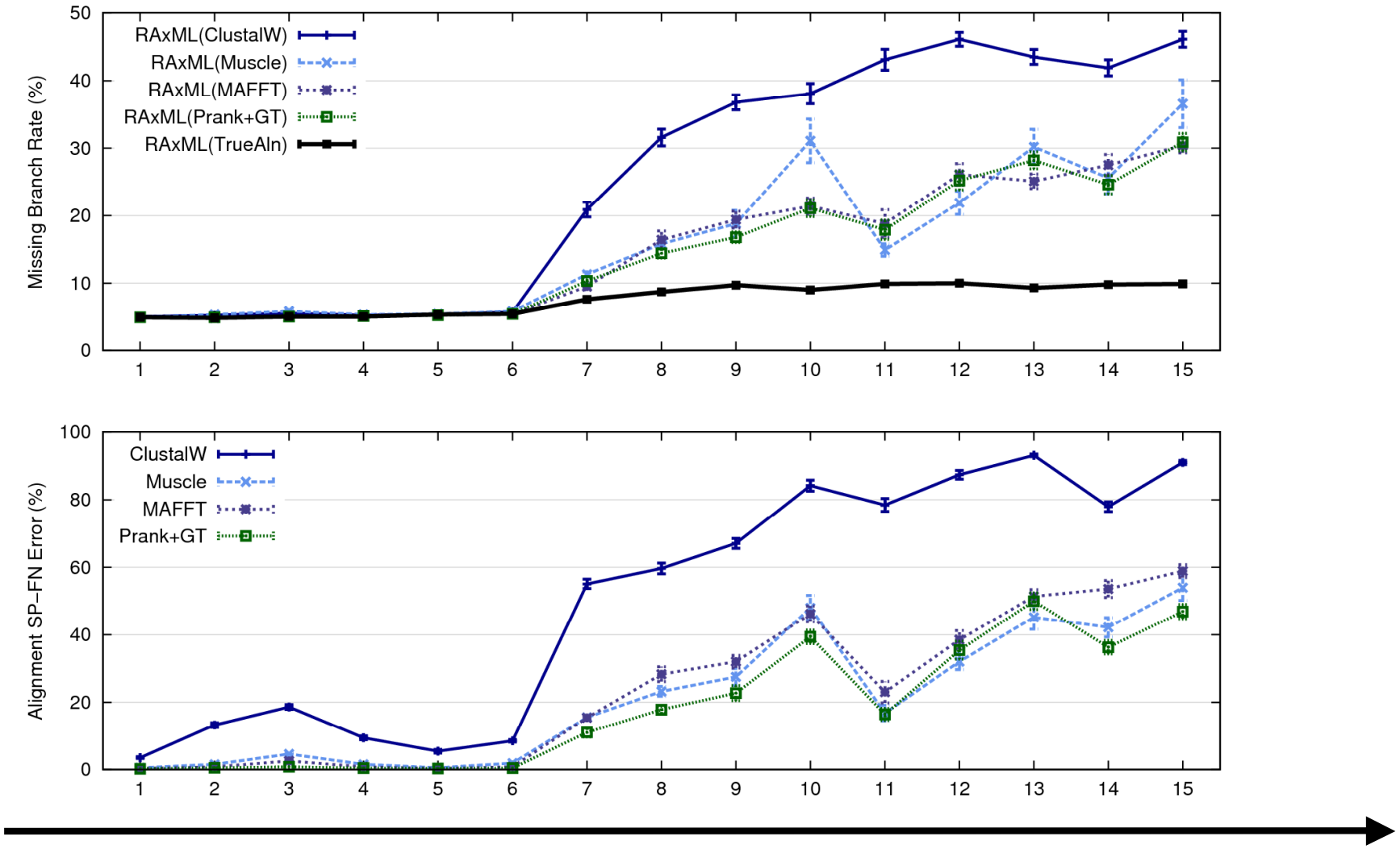
- **SATé**: Simultaneous Alignment and Tree Estimation (Liu et al., Science 2009, and Liu et al. Systematic Biology, in press)
- **SEPP**: SATé-enabled Phylogenetic Placement (Mirarab, Nguyen and Warnow, to appear, PSB 2012)
- **Taxon identification using SEPP** (Warnow, Pop, Mirarab, Nguyen, and Liu, in progress)

Part 1: SATé

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

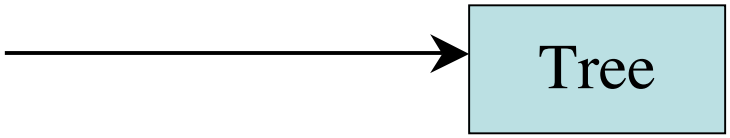
Liu et al., Systematic Biology (in press)

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


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)

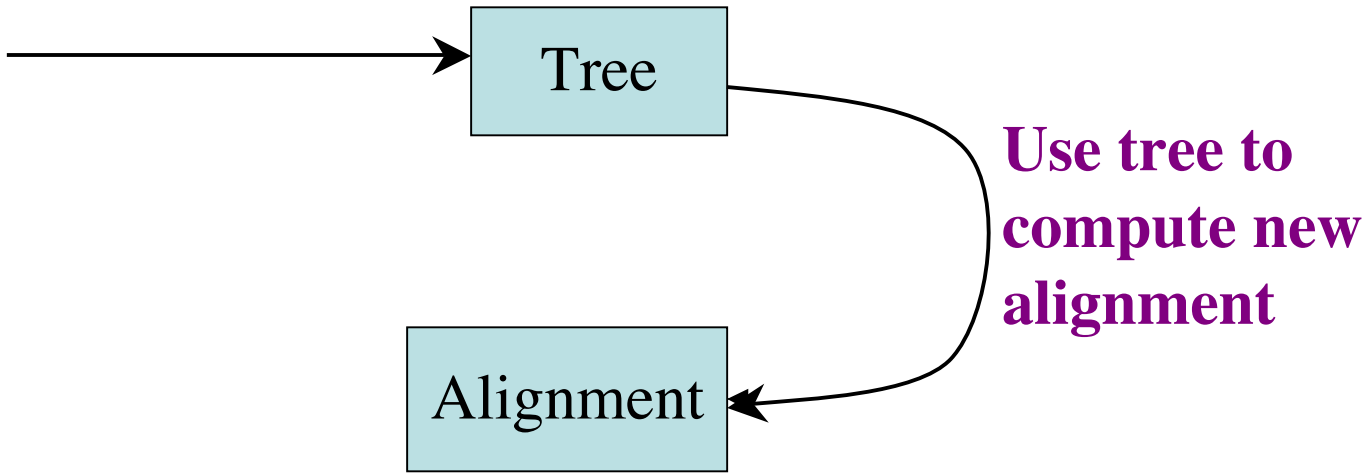
SATé Algorithm

Obtain initial alignment
and estimated ML tree



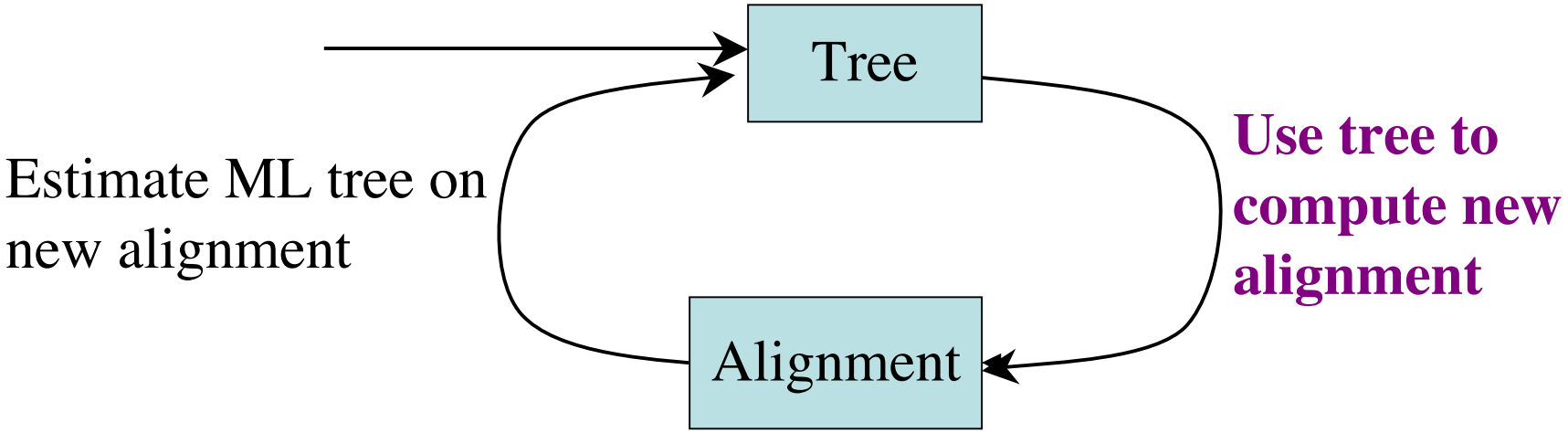
SATé Algorithm

Obtain initial alignment
and estimated ML tree

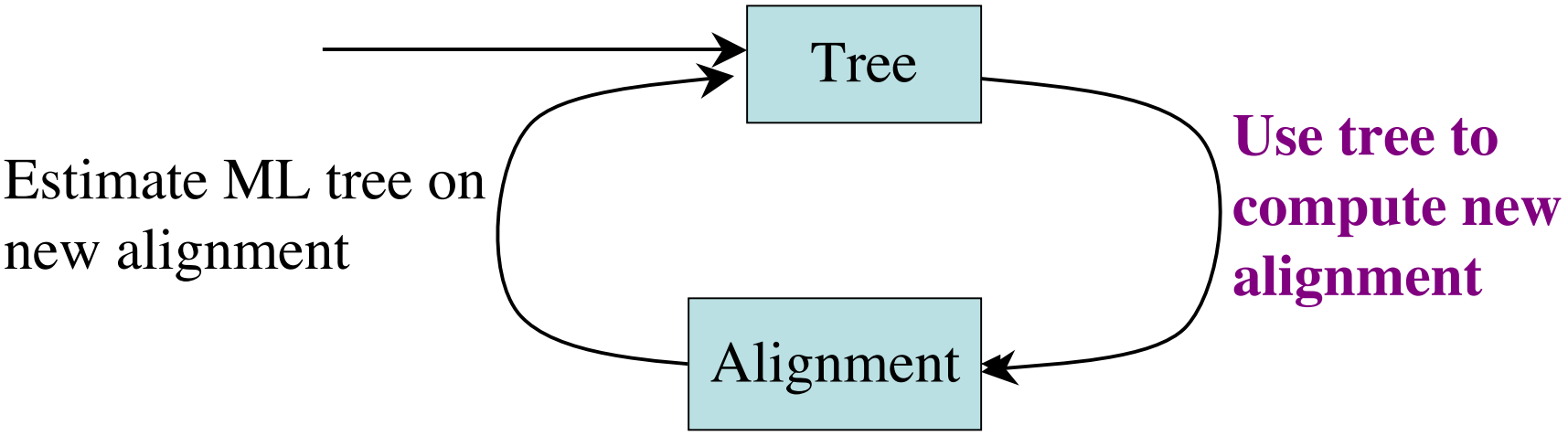


SATé Algorithm

Obtain initial alignment
and estimated ML tree



Obtain initial alignment
and estimated ML tree

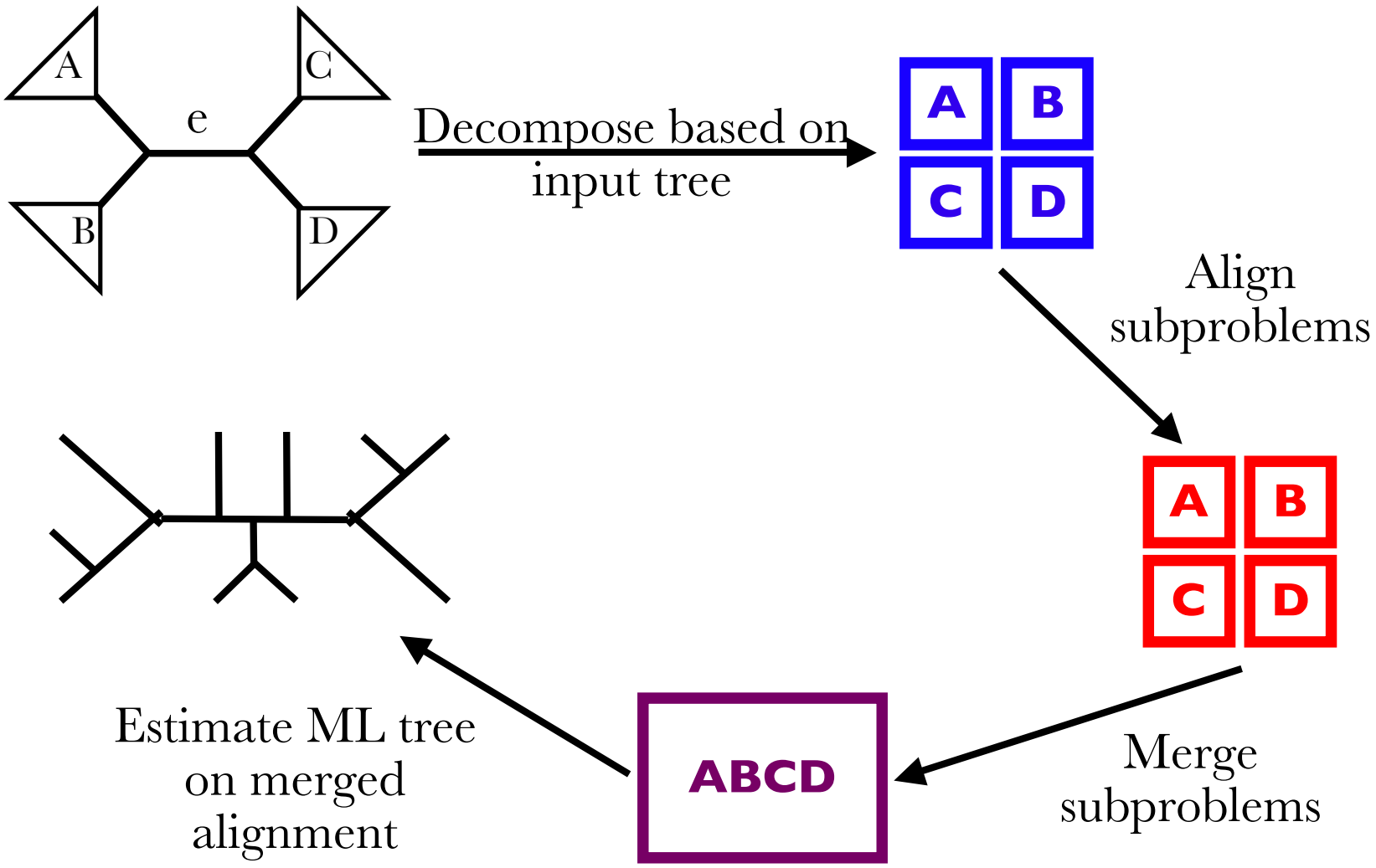


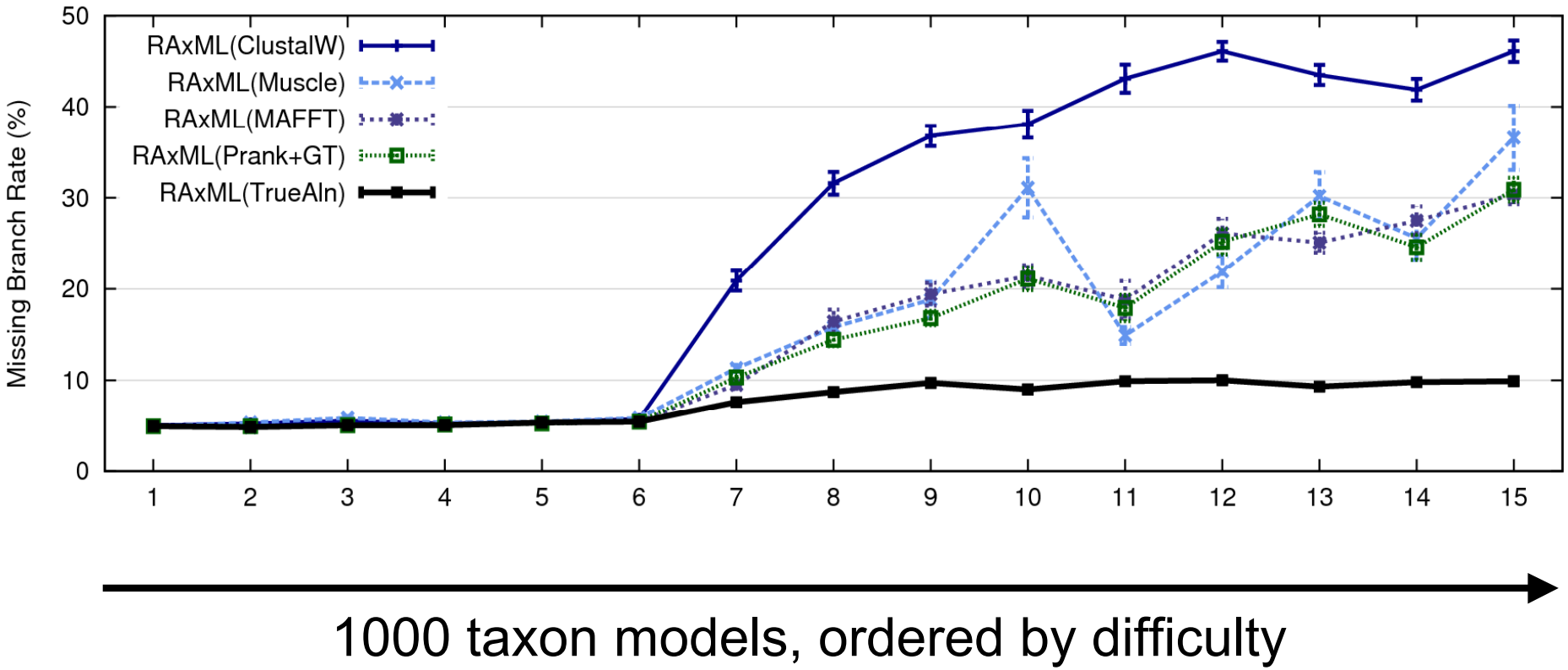
If new alignment/tree pair has worse ML score, realign using a different decomposition

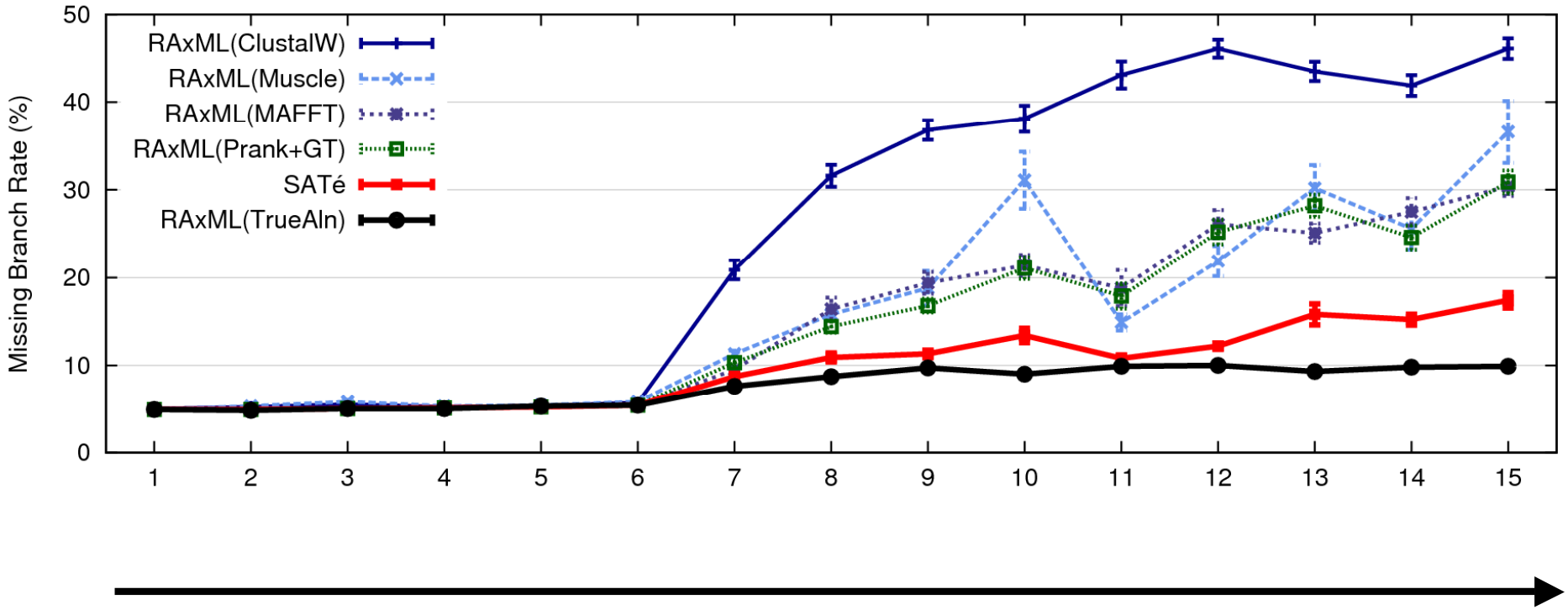
Repeat until termination condition (typically, 24 hours)

SATé Algorithm

One SATé iteration (really 32 subsets)



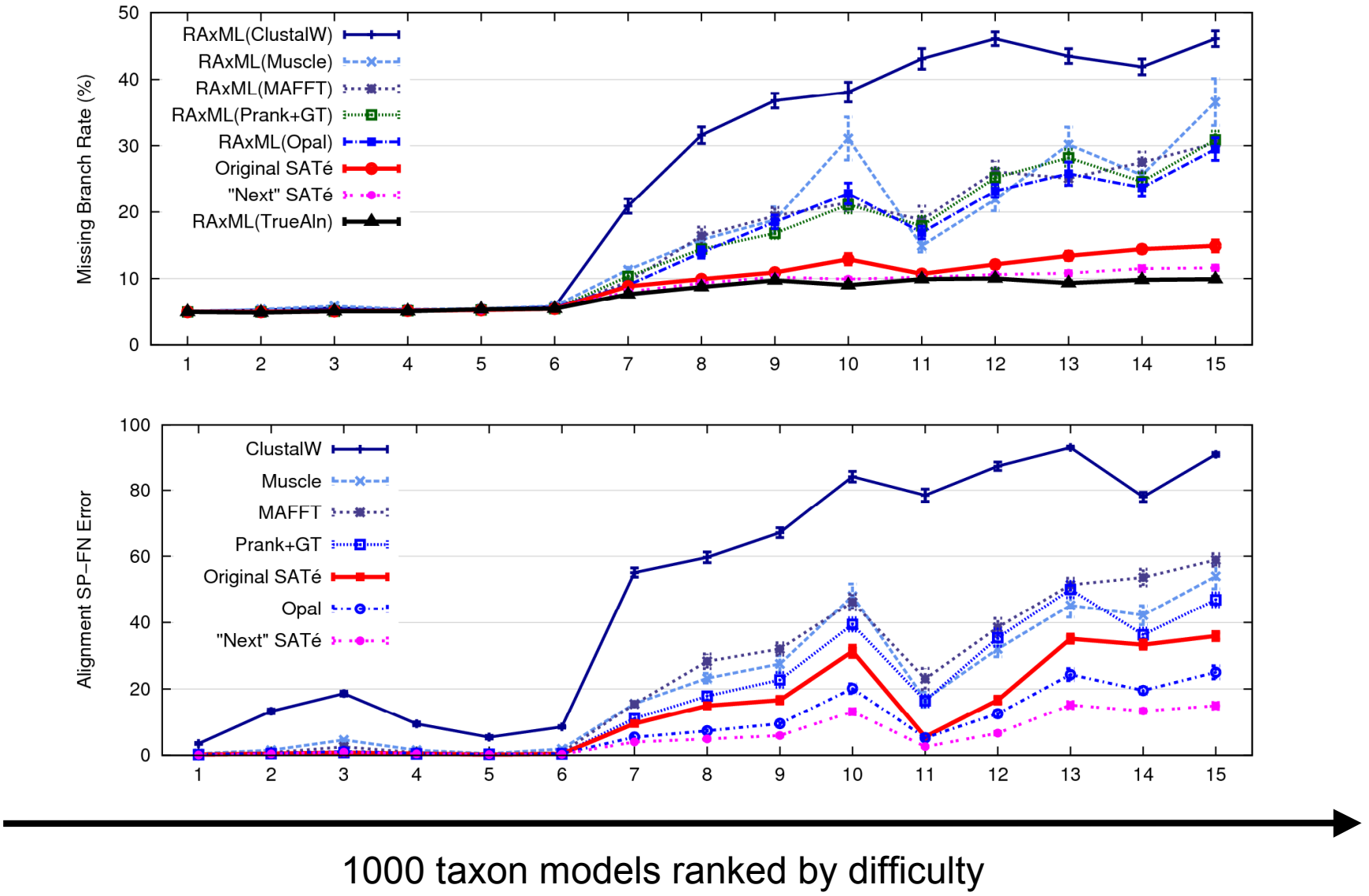

RAxML(ClustalW)
RAxML(Muscle)
RAxML(MAFFT)
RAxML(Prank+GT)
RAxML(TrueAln)
Missing Branch Rate (%)
0
10
20
30
40
50
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
1000 taxon models, ordered by difficulty



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

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

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 and other metagenomic projects

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.

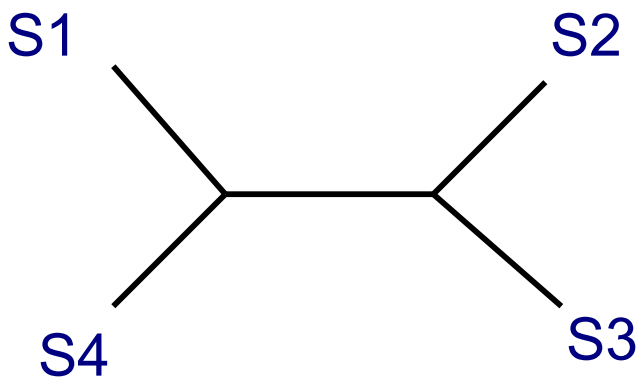
Phylogenetic Placement

- Align each query sequence to backbone alignment

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

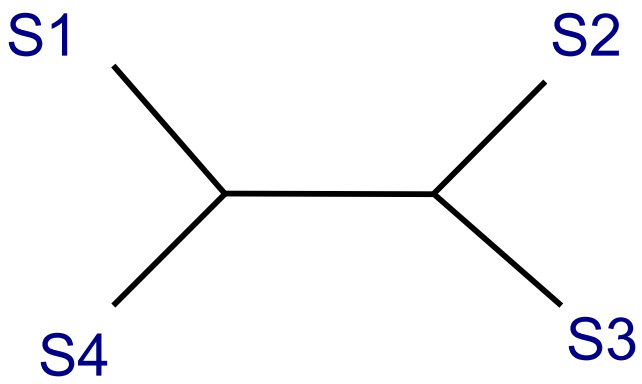
Align Sequence

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



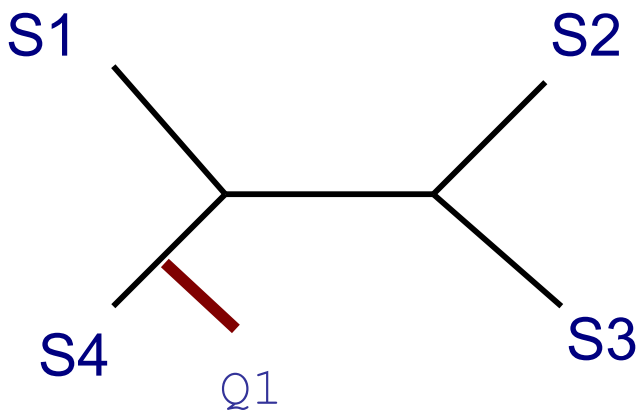
Align Sequence

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

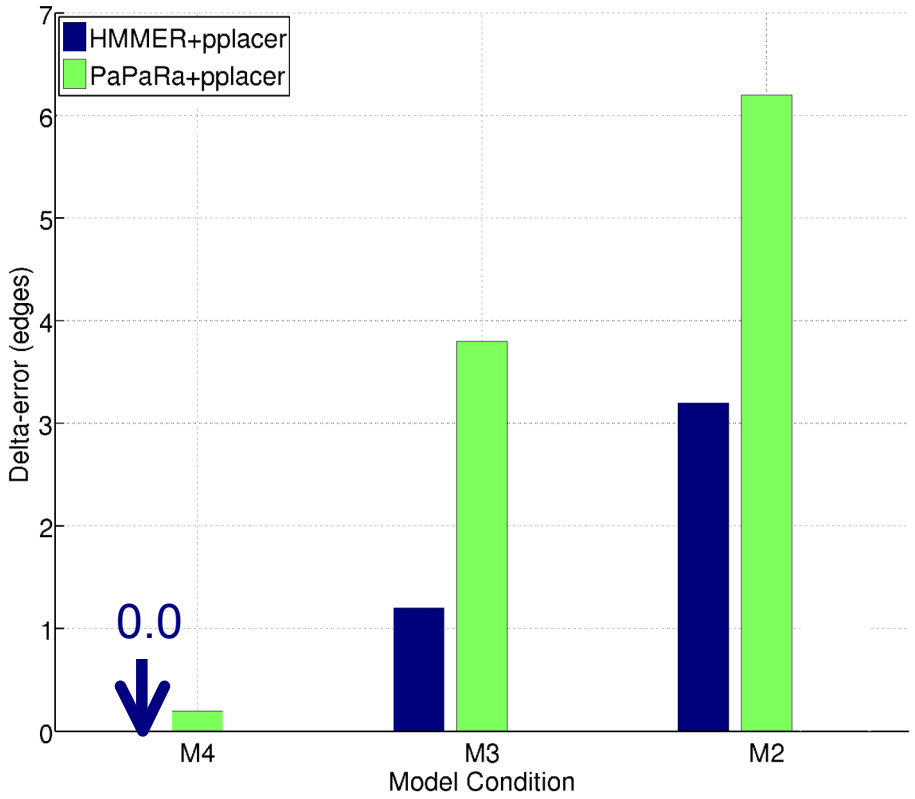


Phylogenetic Placement

- 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, 2010)
  - EPA (Berger and Stamatakis, Systematic Biology 2011)

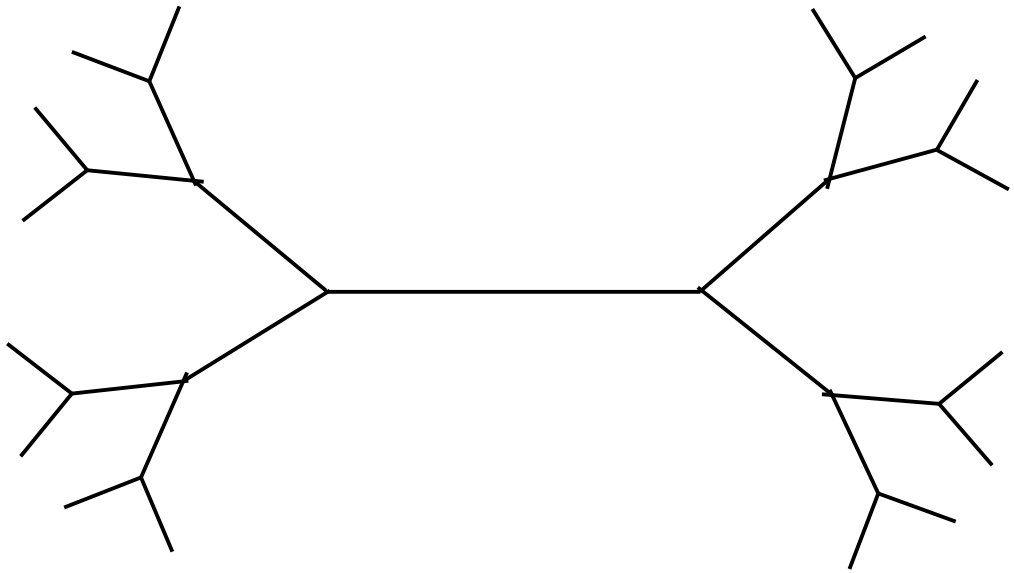
Note: pplacer and EPA use maximum likelihood

HMMER vs. PaPaRa

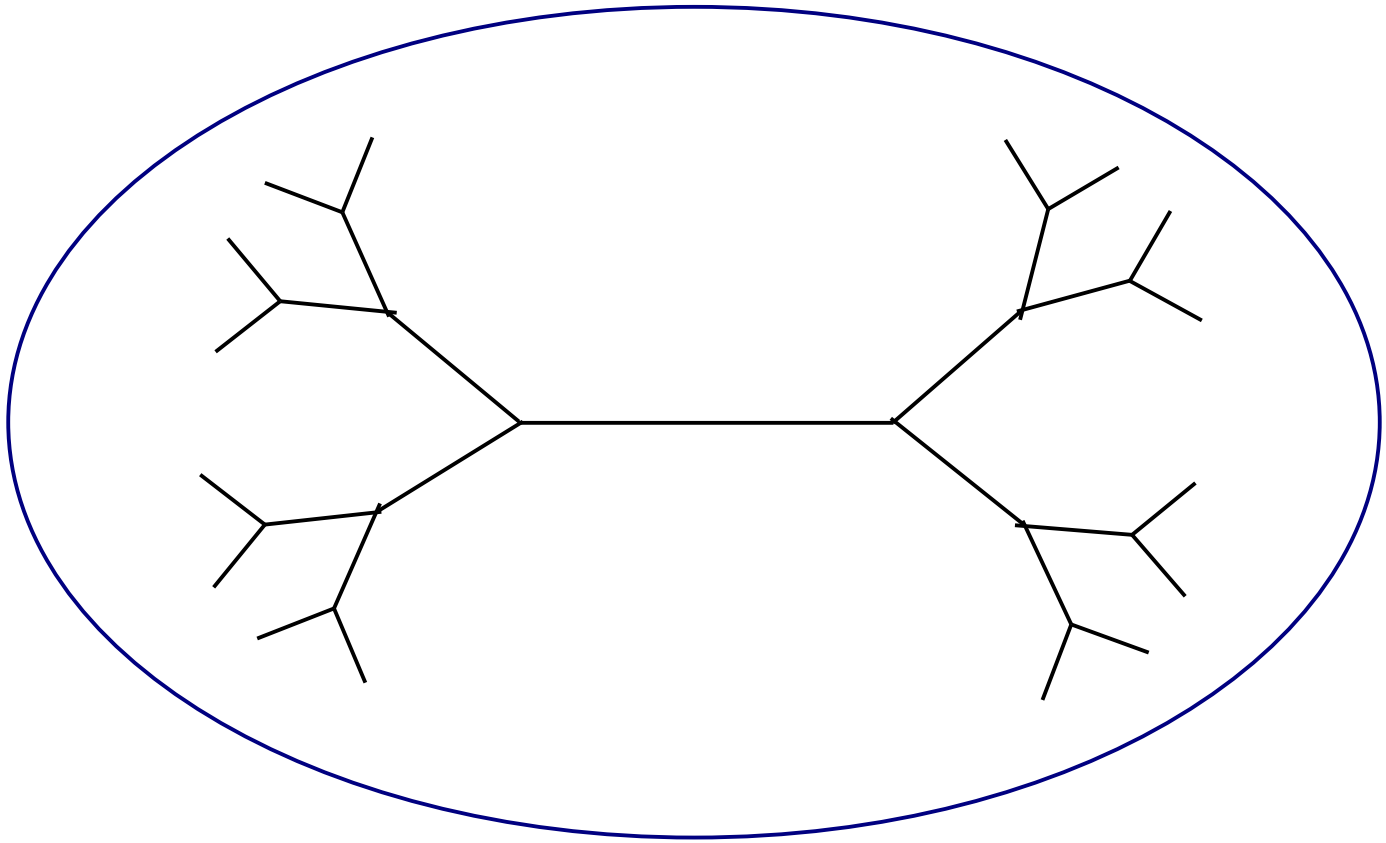


Increasing rate of evolution

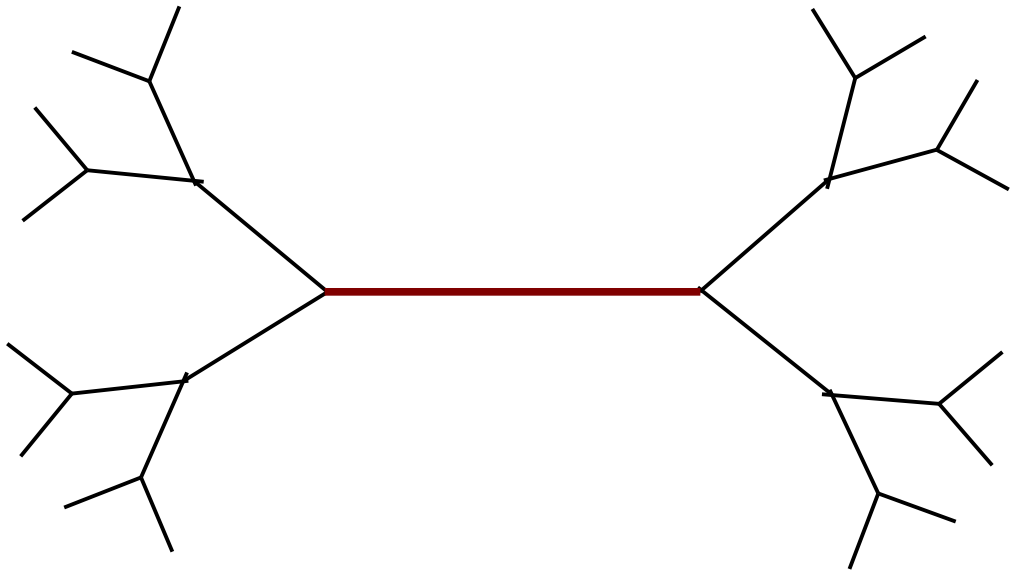
Insights from SATé



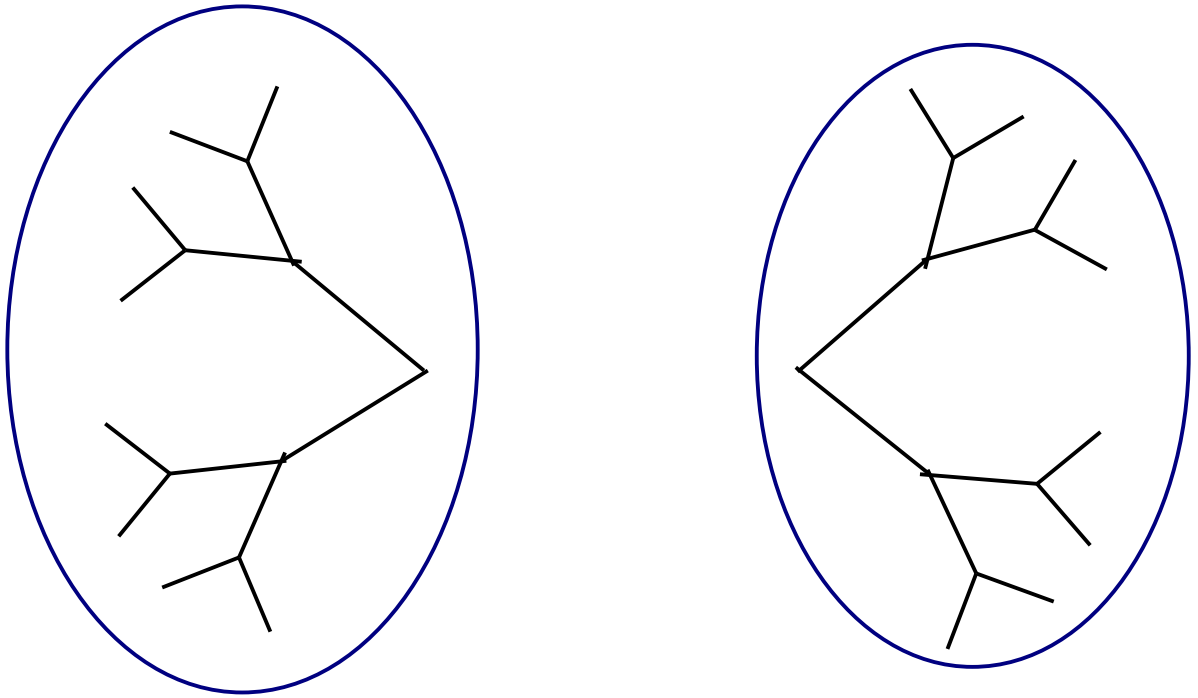
Insights from SATé



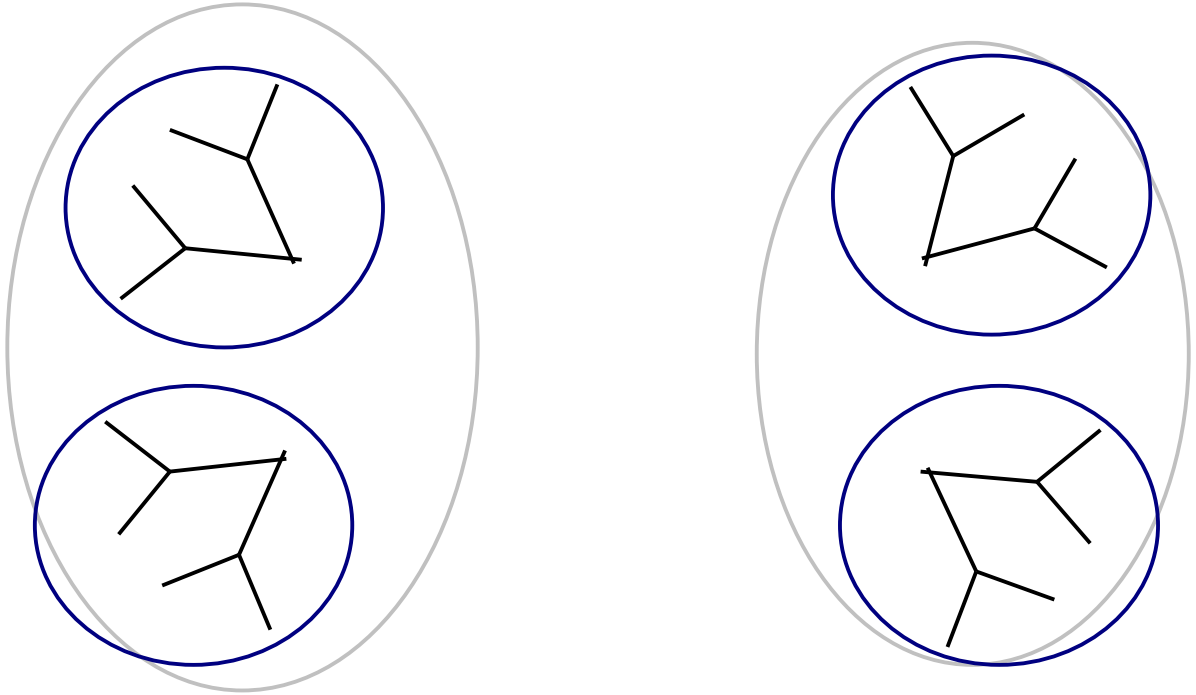
Insights from SATé



Insights from SATé



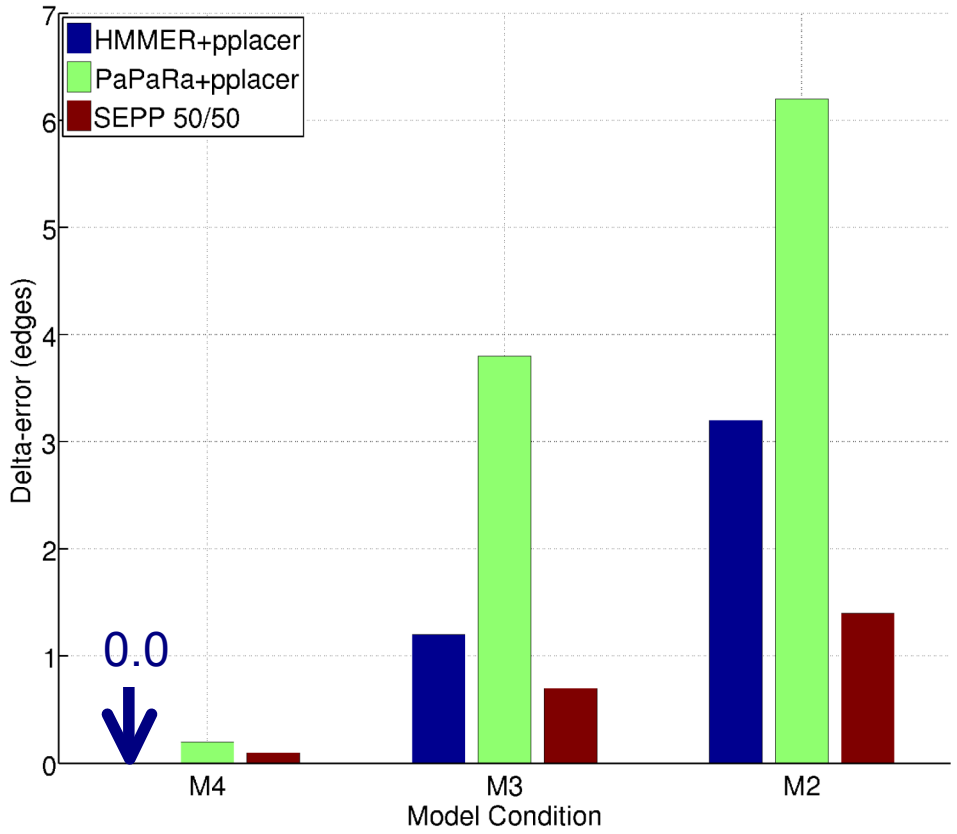
Insights from SATé



SEPP Parameter Exploration

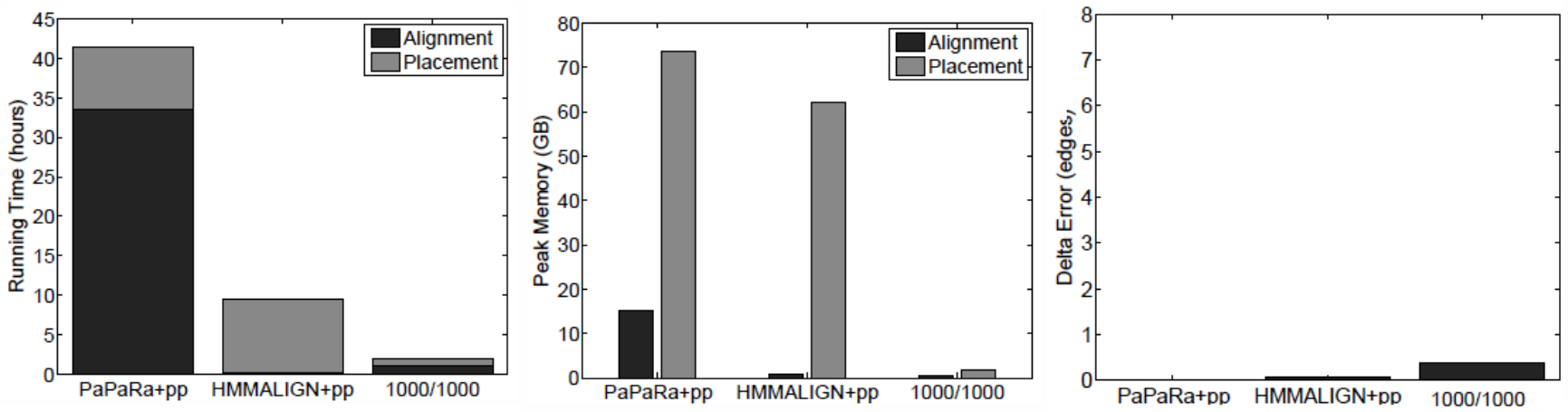
- Alignment subset size and placement subset size impact the accuracy, running time, and memory of SEPP
- 10% rule (subset sizes 10% of backbone) had best overall performance

SEPP (10%-rule) on simulated data



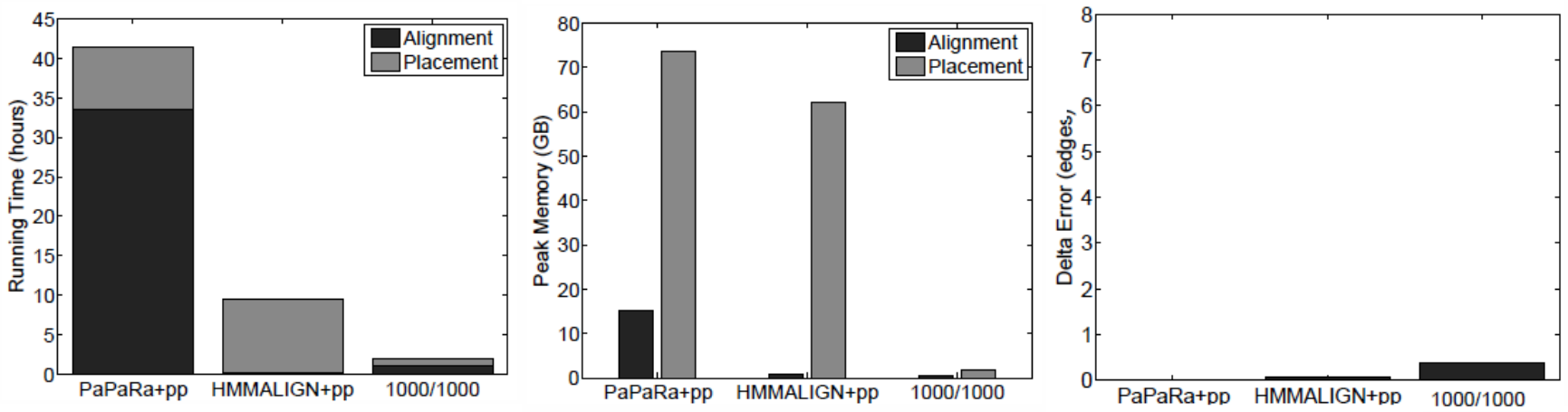
Increasing rate of evolution

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

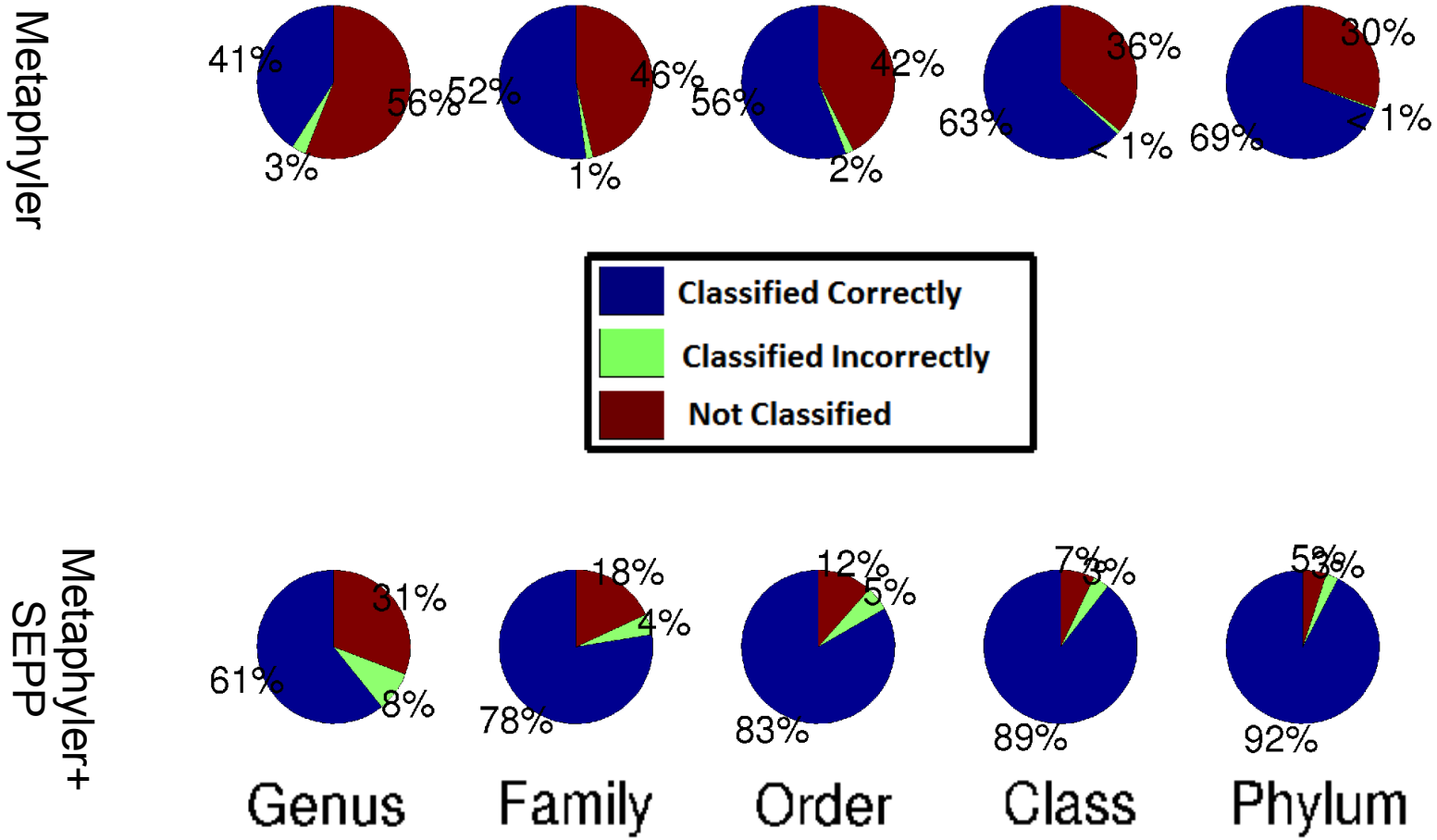
SEPP 1000/1000: ~6 days

Part III: Taxon Identification

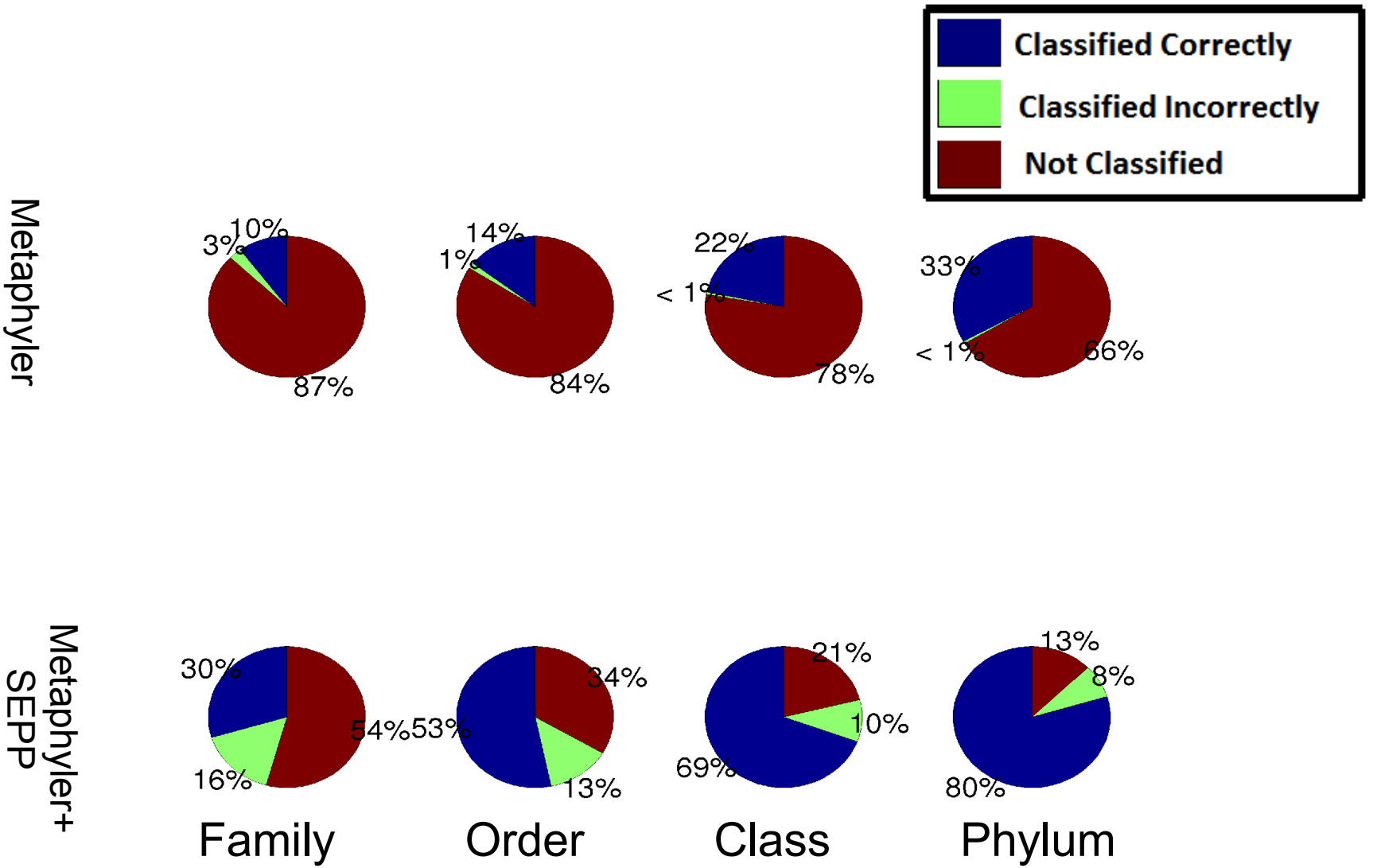
Taxon Identification

- Objective: identify species, genus, etc., for each short read
- Leading methods: Metaphyler (Univ Maryland), Phylopythia, PhymmBL, Megan
- The best of these methods are very precise, but fail to classify many short reads.

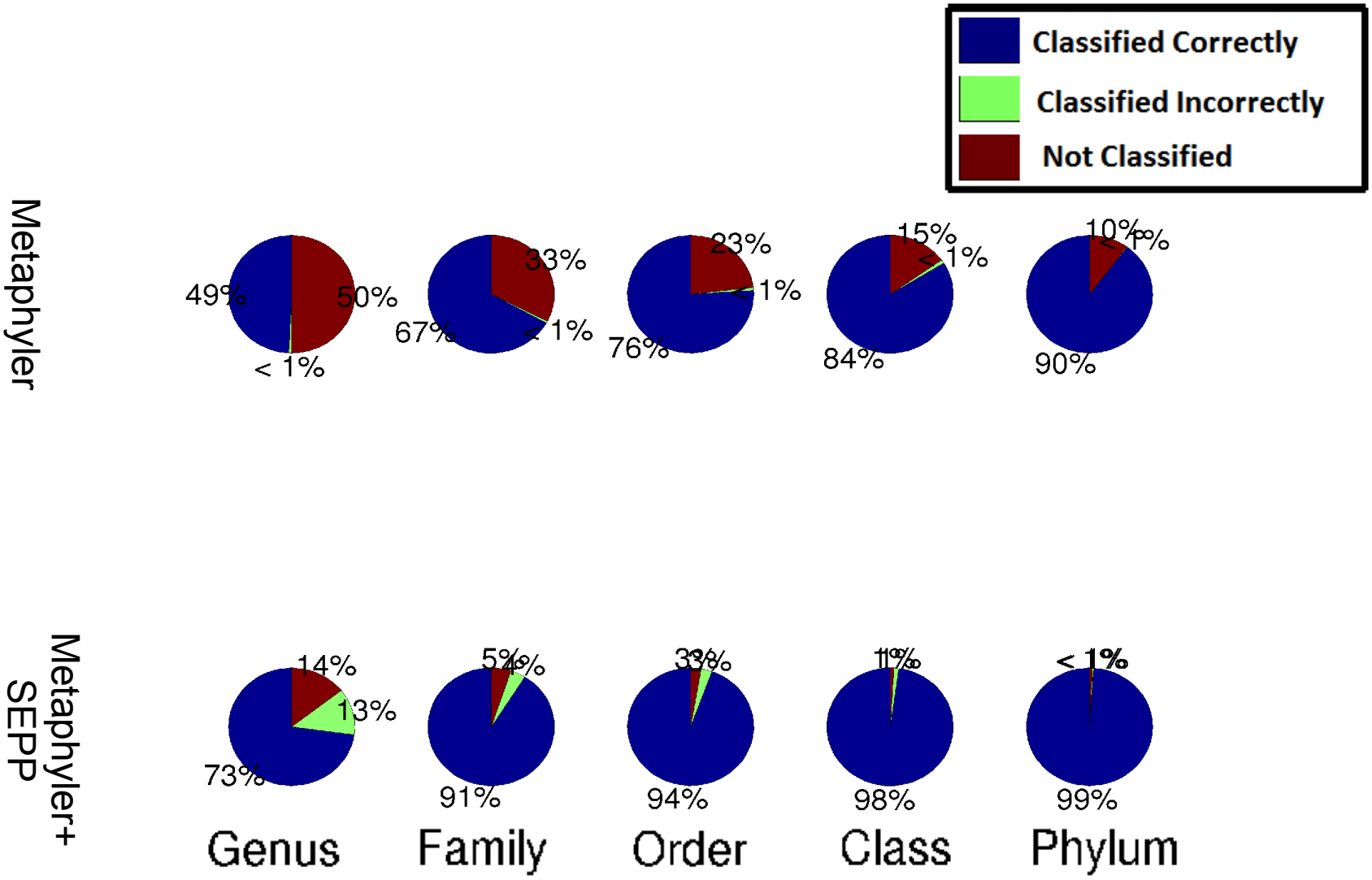
rpsB 60bp leave out species



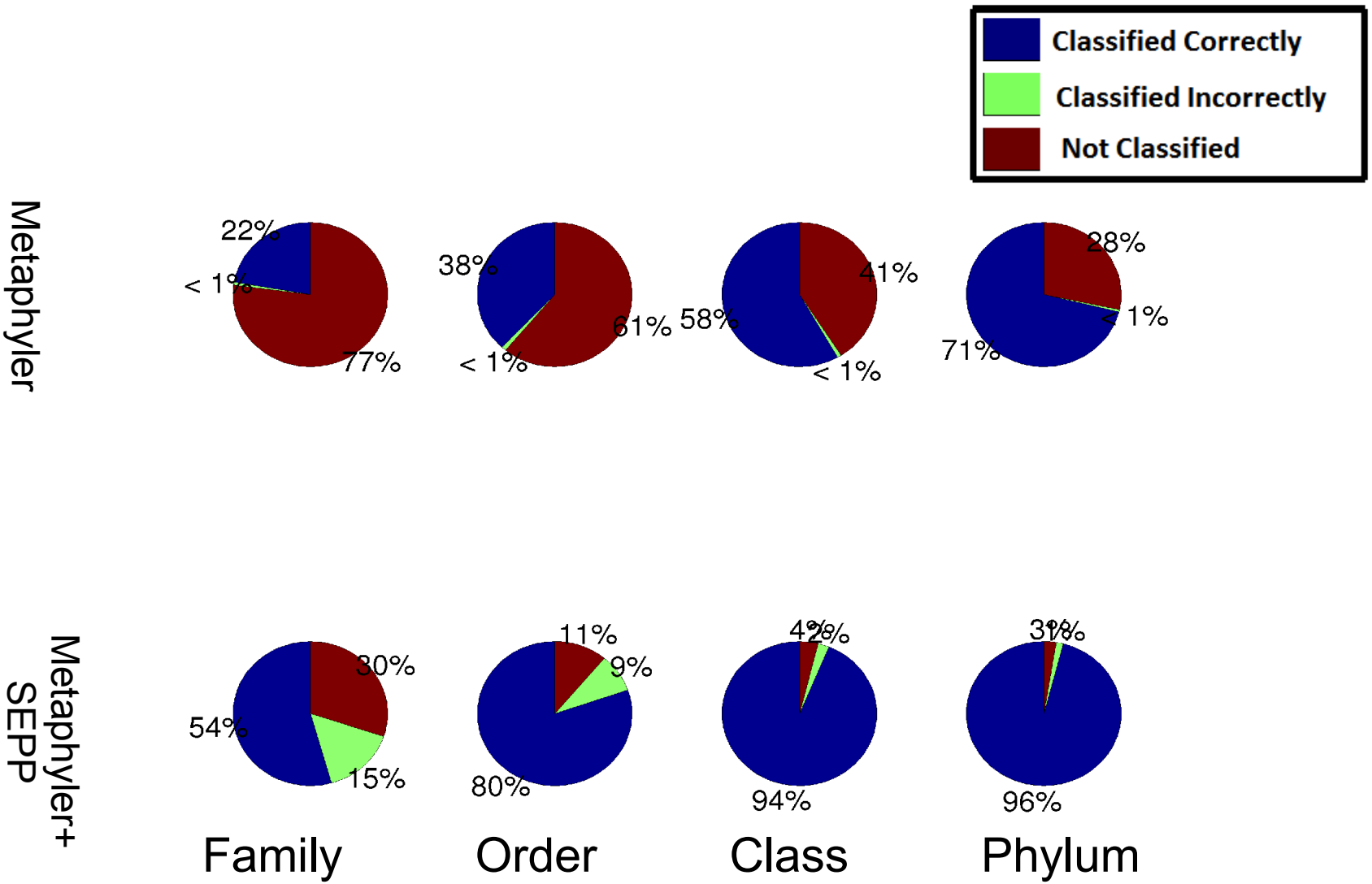
rpsB 60bp leave out genus



rpsB 300bp leave out species



rpsB 300bp leave out genus



Phylogenetic "Boosters"

- SATé: co-estimation of alignments and trees
- SEPP: phylogenetic analysis of fragmentary data
- Metaphyler+SEPP: taxonomic identification of short reads

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

Relevant Publications

- SATé: Liu et al. 2009, Science, vol. 324, no. 5934, pp. 1561-1564, and Systematic Biology (in press); software available at http://phylo.bio.ku.edu/software/sate/sate.html
- SEPP: Mirarab, Nguyen, and Warnow, Proceedings of the Pacific Symposium on Biocomputing (PSB) 2012
- Metaphyler: Liu et al., BMC Genomics 2011 (Suppl 2): S4
- Metaphyler+SEPP: joint with Mihai Pop, Bo Liu, Siavash Mirarab, and Nam Nguyen, in preparation

See http://www.cs.utexas.edu/users/tandy/papers.html

Summary

- Standard alignment and phylogeny estimation methods do not provide adequate accuracy on large datasets, and NGS data present novel challenges
- When markers tend to yield poor alignments and trees, 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: Siavash Mirarab and Nam Nguyen
  - Metaphyler+SEPP: Siavash Mirarab, Nam Nguyen, Bo Liu, and Mihai Pop