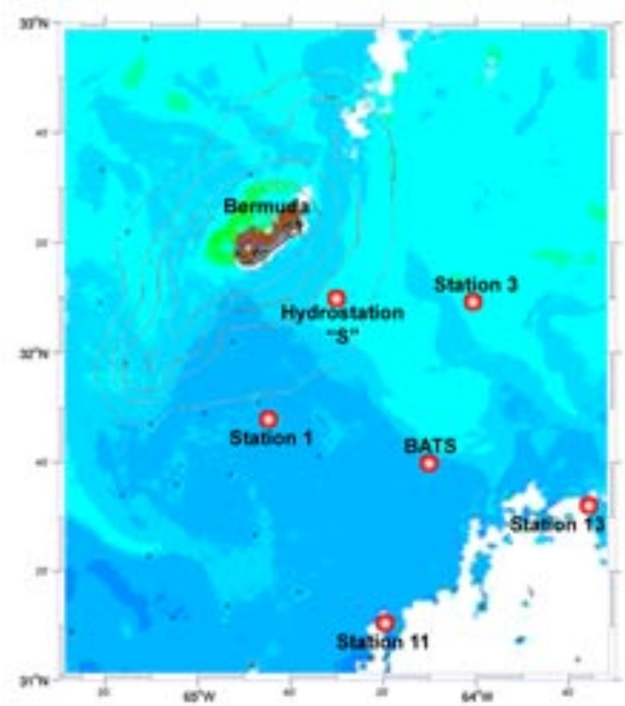
SEPP and TIPP for metagenomic analysis

Tandy Warnow
Department of Computer Science
University of Texas

Metagenomics:

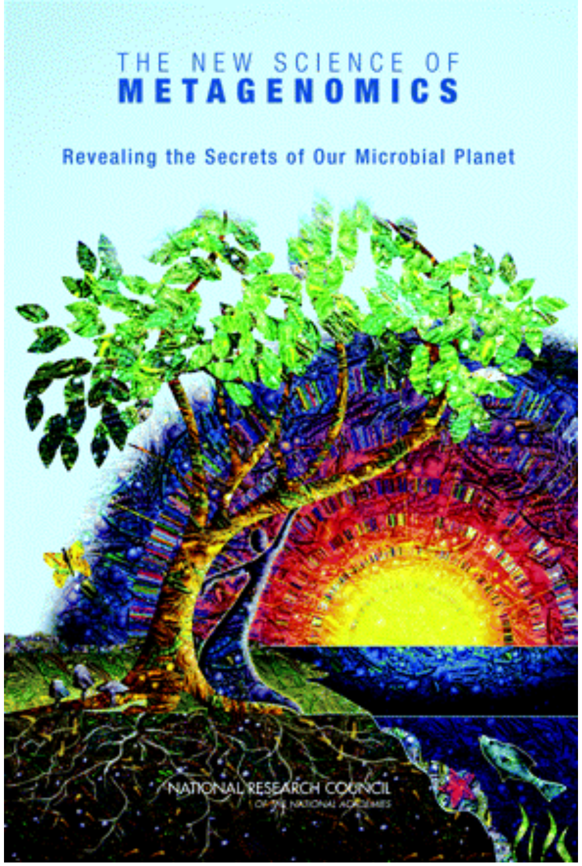
Venter et al., Exploring the Sargasso Sea:

Scientists Discover One Million New Genes in Ocean Microbes



Computational Phylogenetics and Metagenomics

Courtesy of the Tree of Life project



THE NEW SCIENCE OF
METAGENOMICS

Revealing the Secrets of Our Microbial Planet

NATIONAL RESEARCH COUNCIL
OF THE NATIONAL ACADEMIES

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 - TIPP+Metaphyler collaboration with Mihai Pop and Bo Liu)

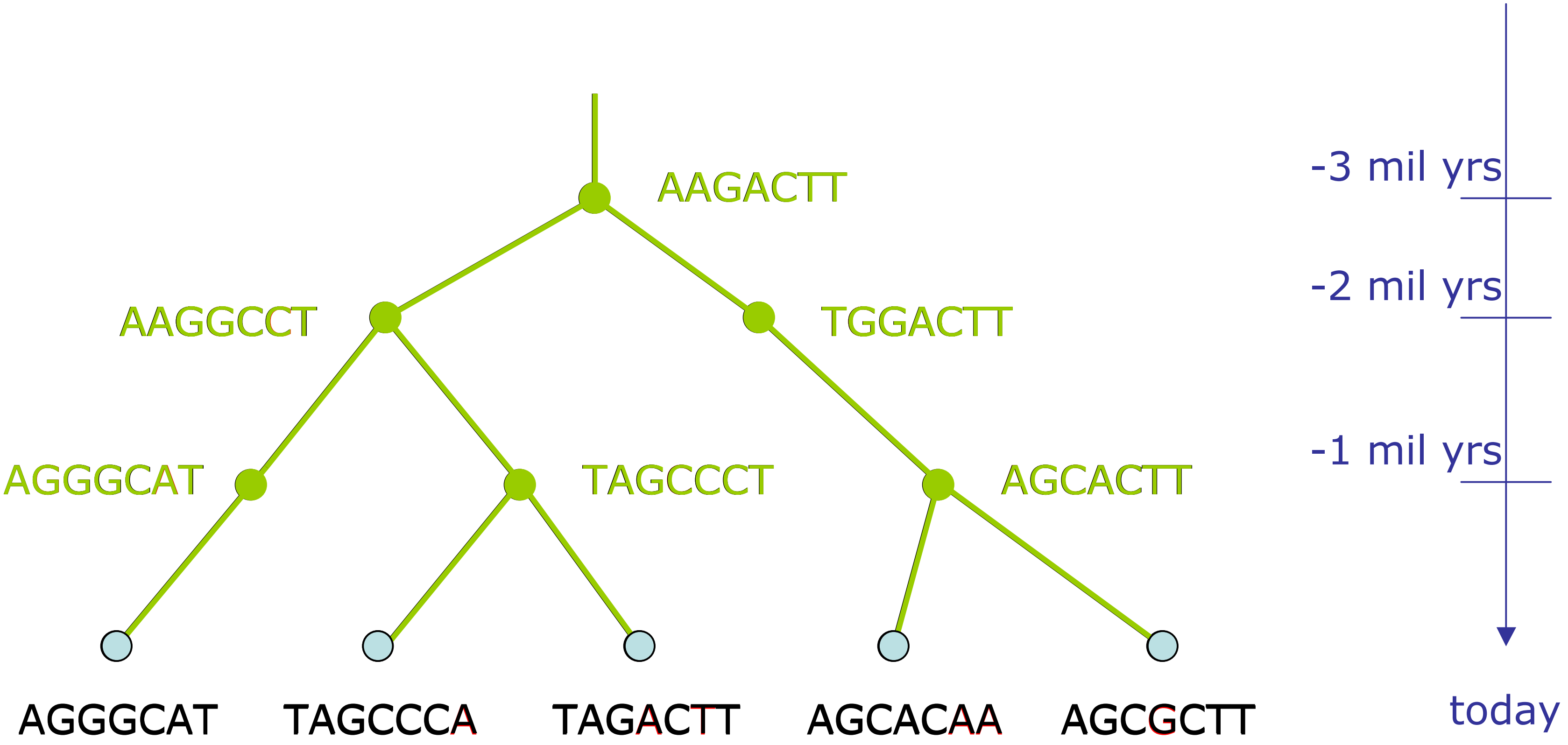
Part 1: SATé

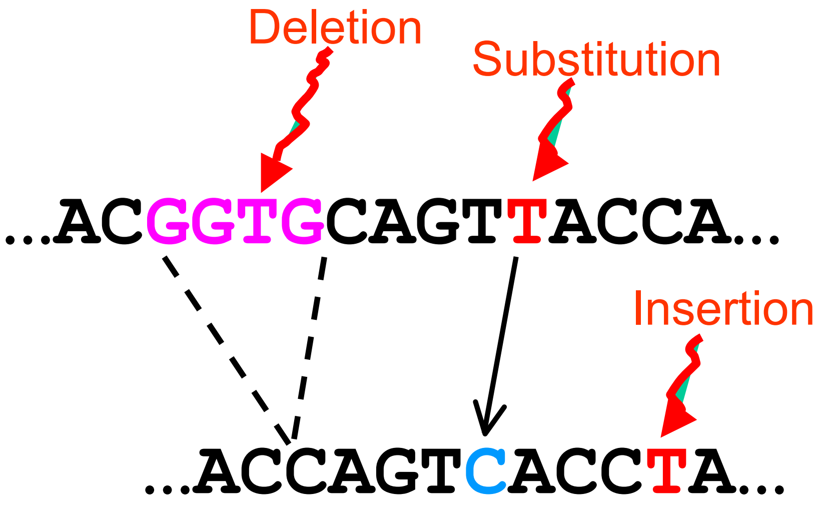
Liu, Nelesen, Raghavan, Linder, and Warnow, *Science*, 19 June 2009, pp. 1561-1564.

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

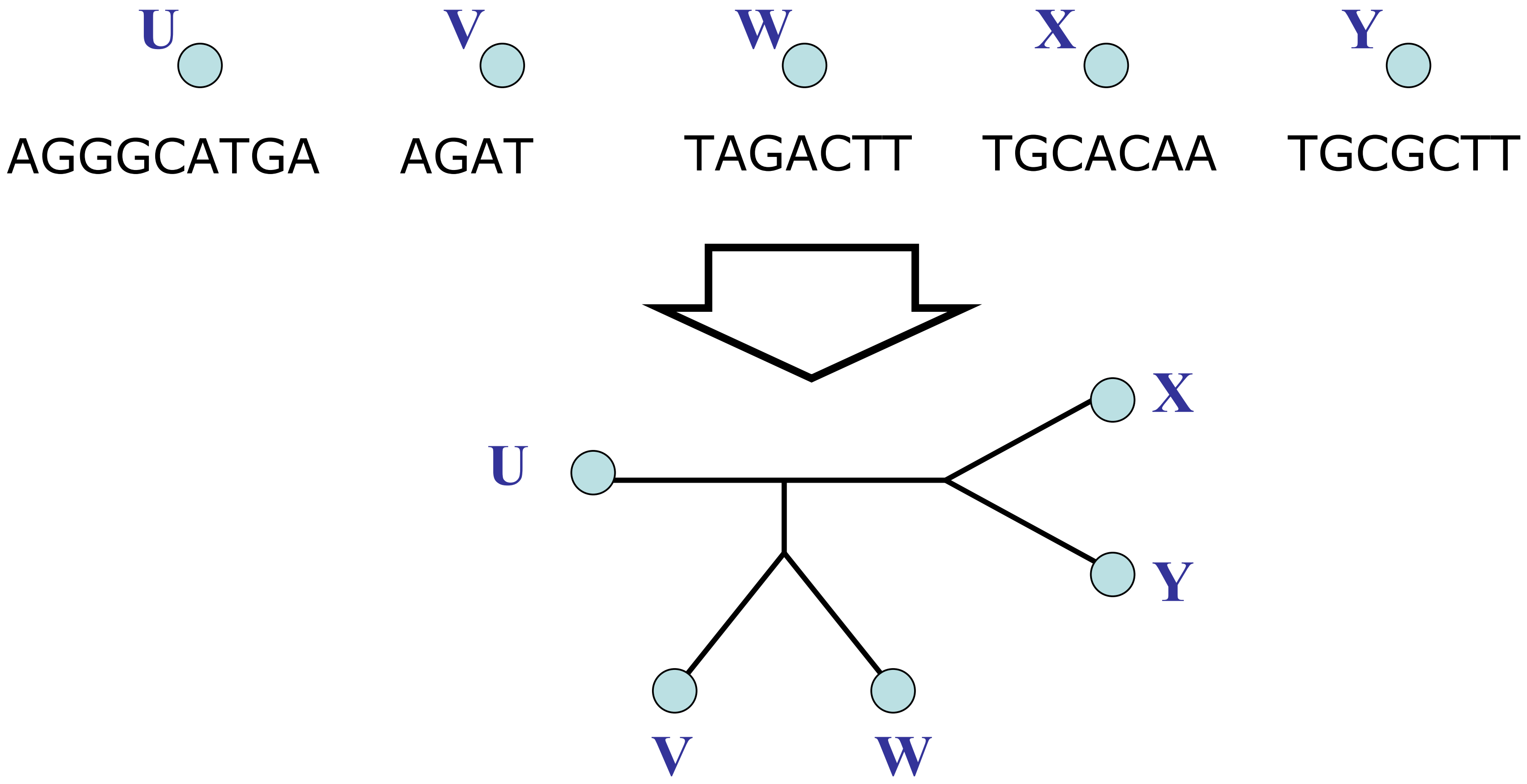




```
...ACGGTGCAGTTACC-A...
...AC----CAGTCACCTA...
```

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



AGGGCATGA AGAT TAGACTT TGCACAA TGCGCTT

Input: unaligned sequences

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

Phase 1: Multiple Sequence Alignment

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

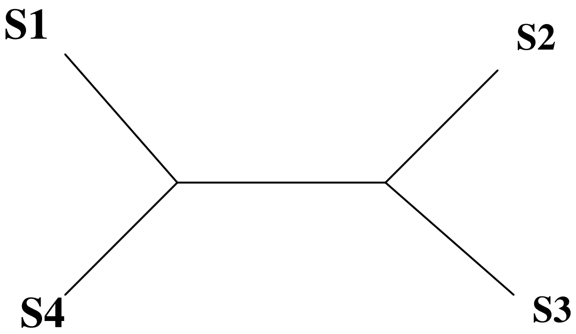
→

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

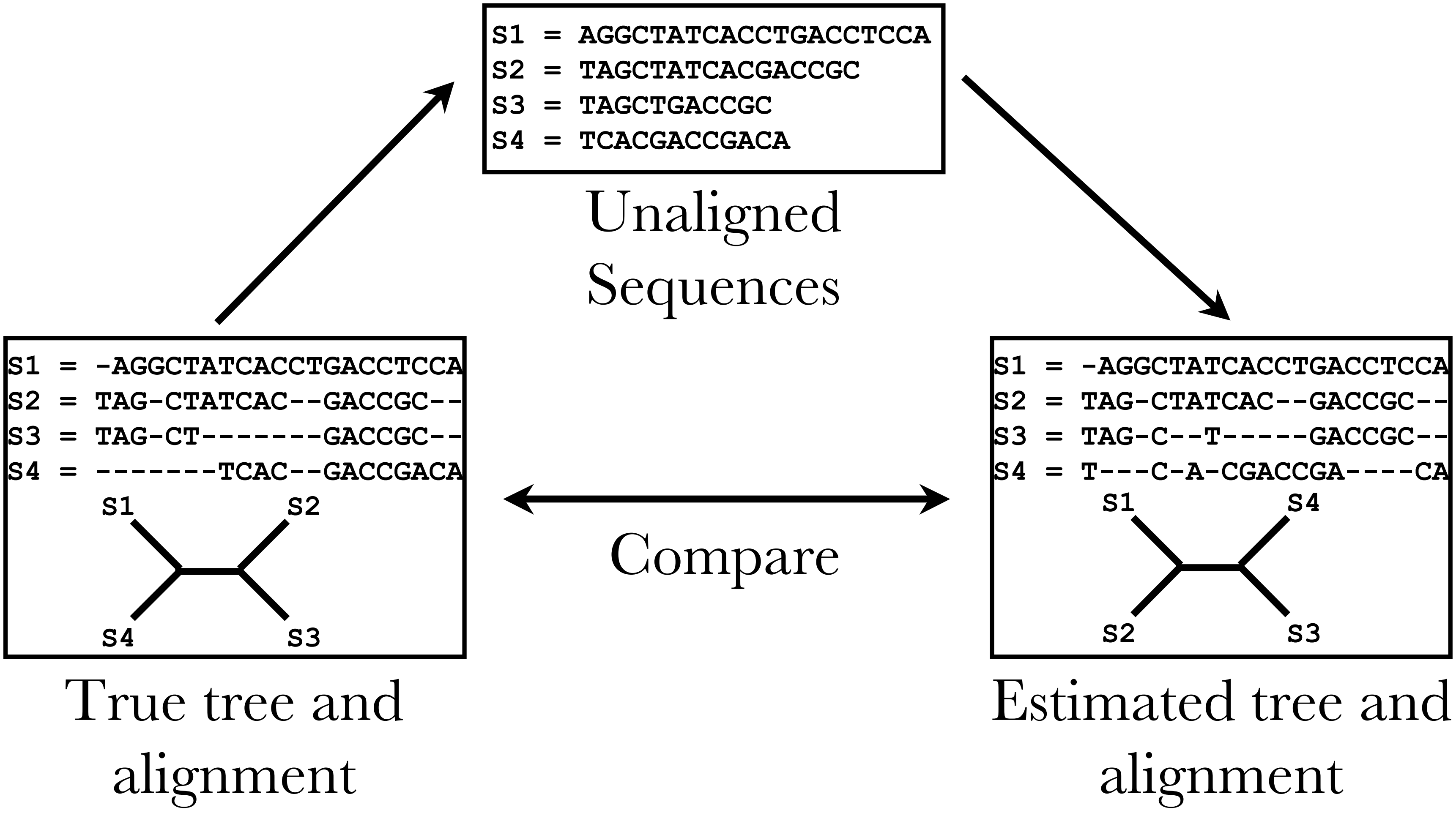
Phase 2: Construct tree

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

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



Simulation Studies



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

Unaligned Sequences

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

True tree and alignment

Compare

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

Estimated tree and alignment

Two-phase estimation

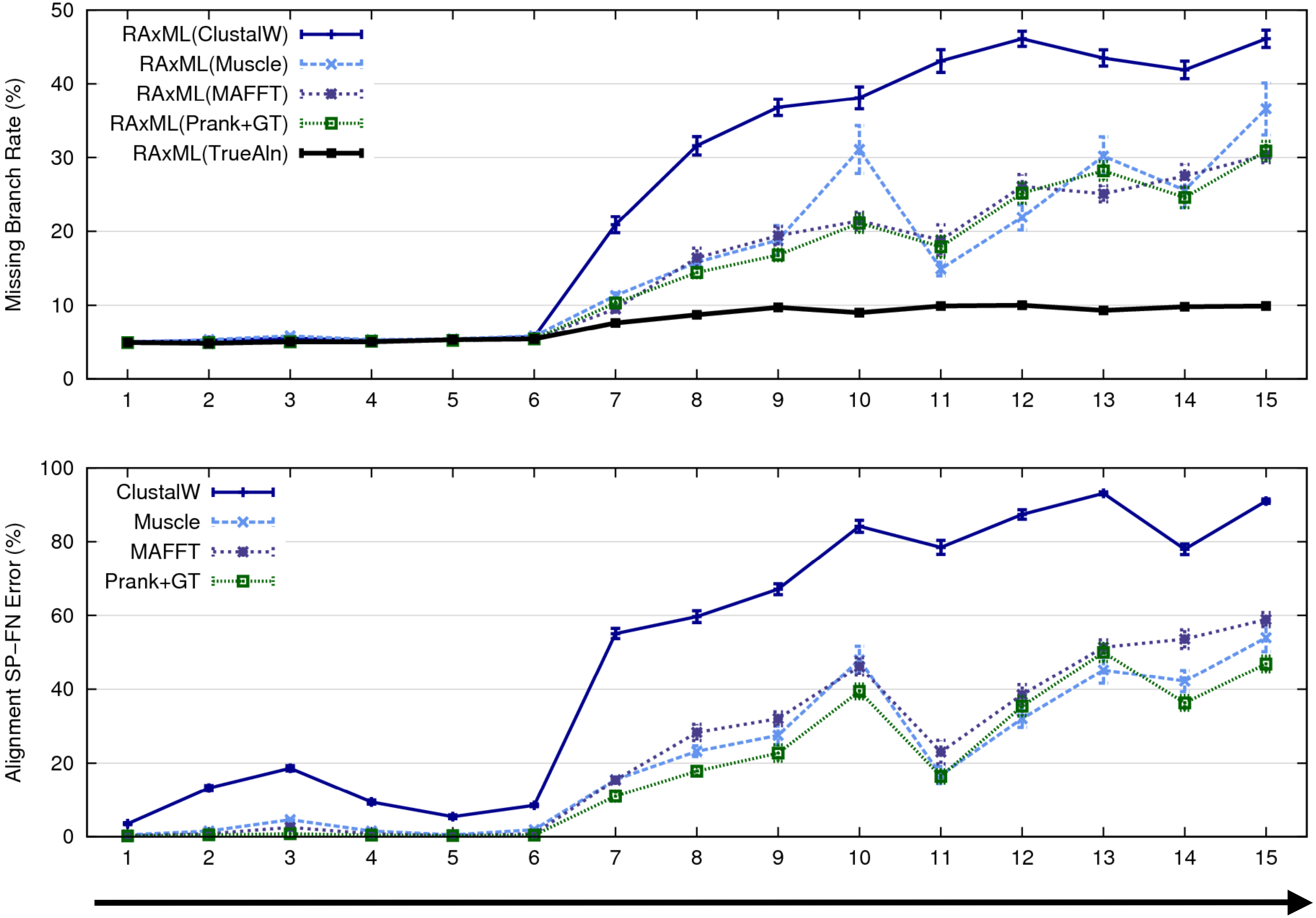
Alignment methods

- 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*



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)

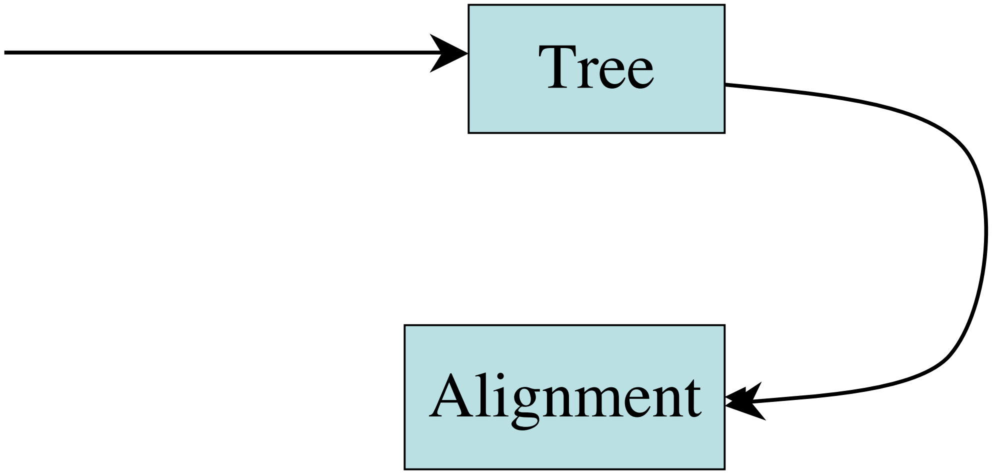
SATé Algorithm

Obtain initial alignment
and estimated ML tree

Tree

SATé Algorithm

Obtain initial alignment
and estimated ML tree

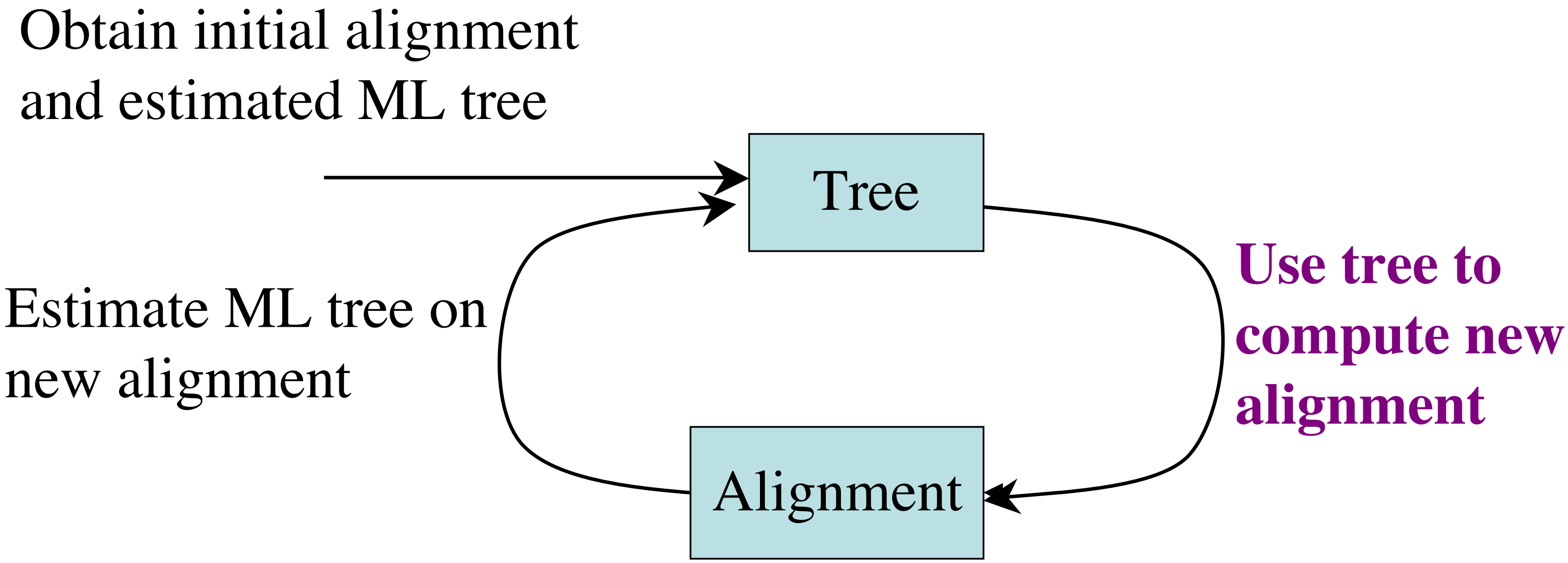


Tree

**Use tree to
compute new
alignment**

Alignment

SATé Algorithm



Obtain initial alignment and estimated ML tree

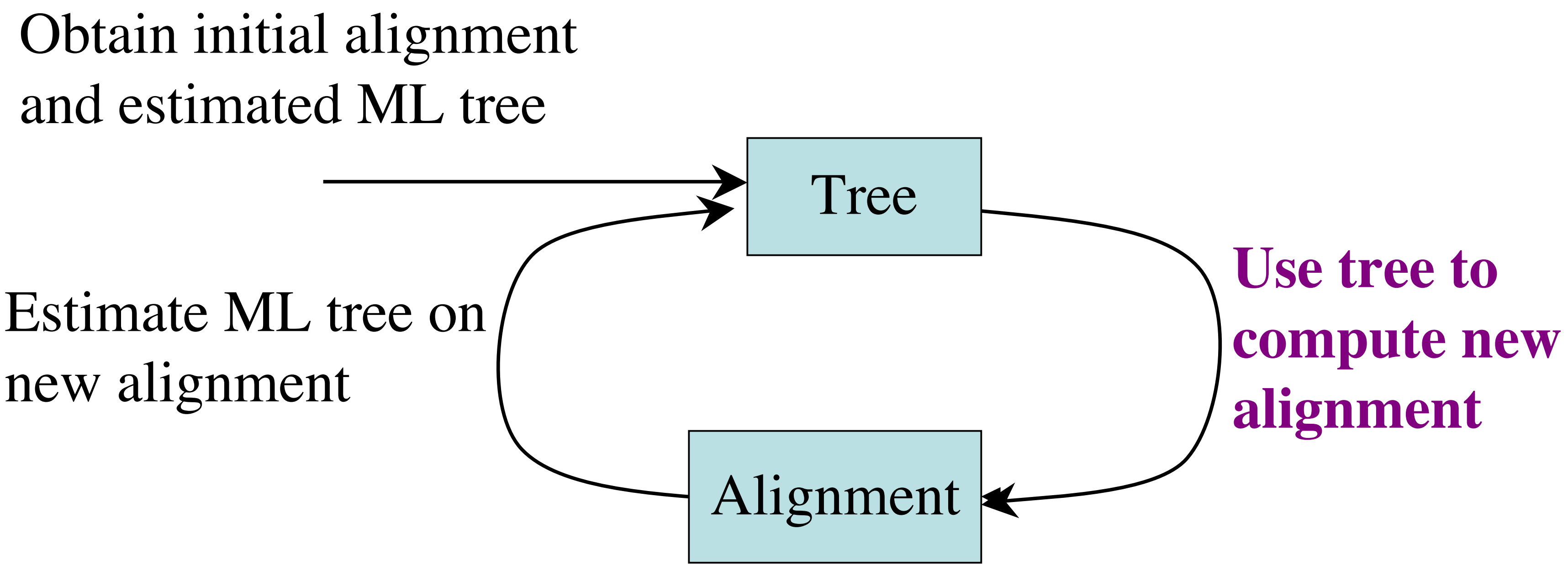
Tree

Use tree to compute new alignment

Alignment

Estimate ML tree on new alignment

SATé Algorithm



Obtain initial alignment and estimated ML tree

Tree

Use tree to compute new alignment

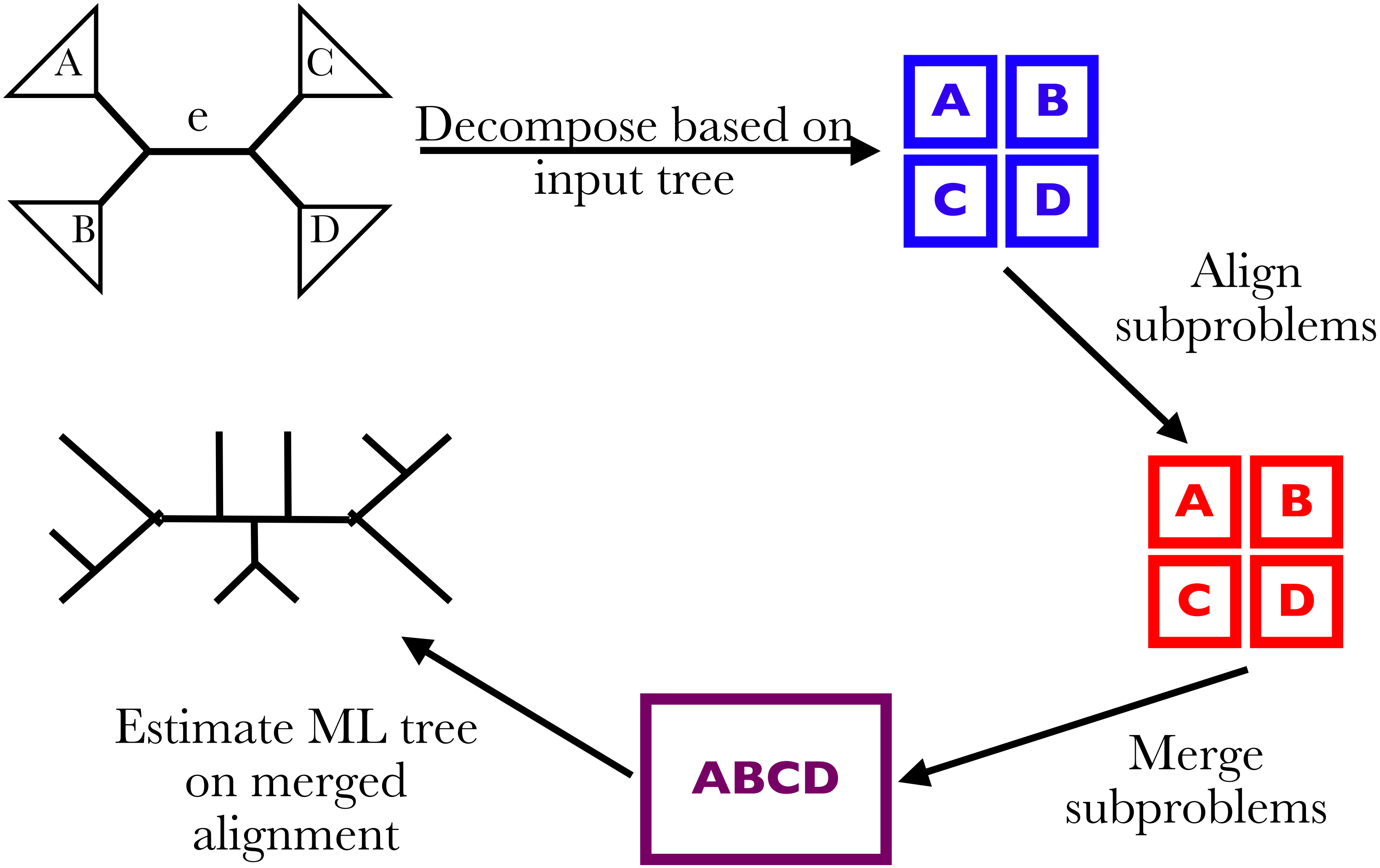
Alignment

Estimate ML tree on new alignment

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)

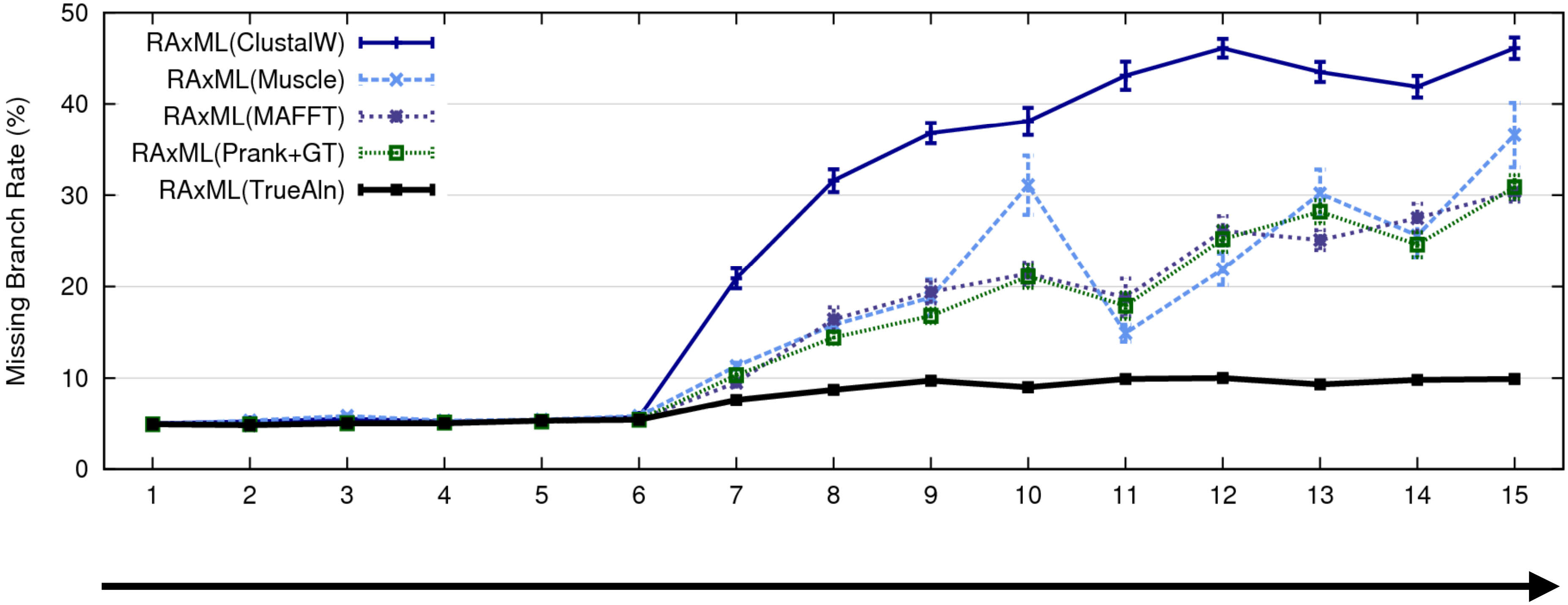


Decompose based on input tree

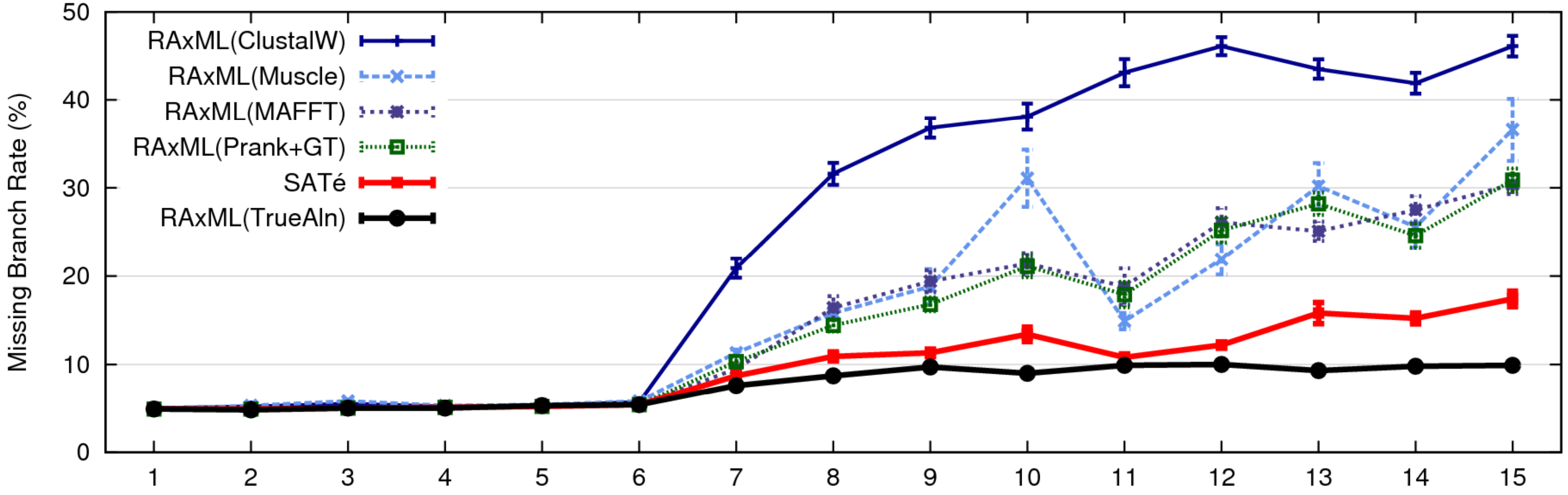
Align subproblems

Merge subproblems

Estimate ML tree on merged alignment



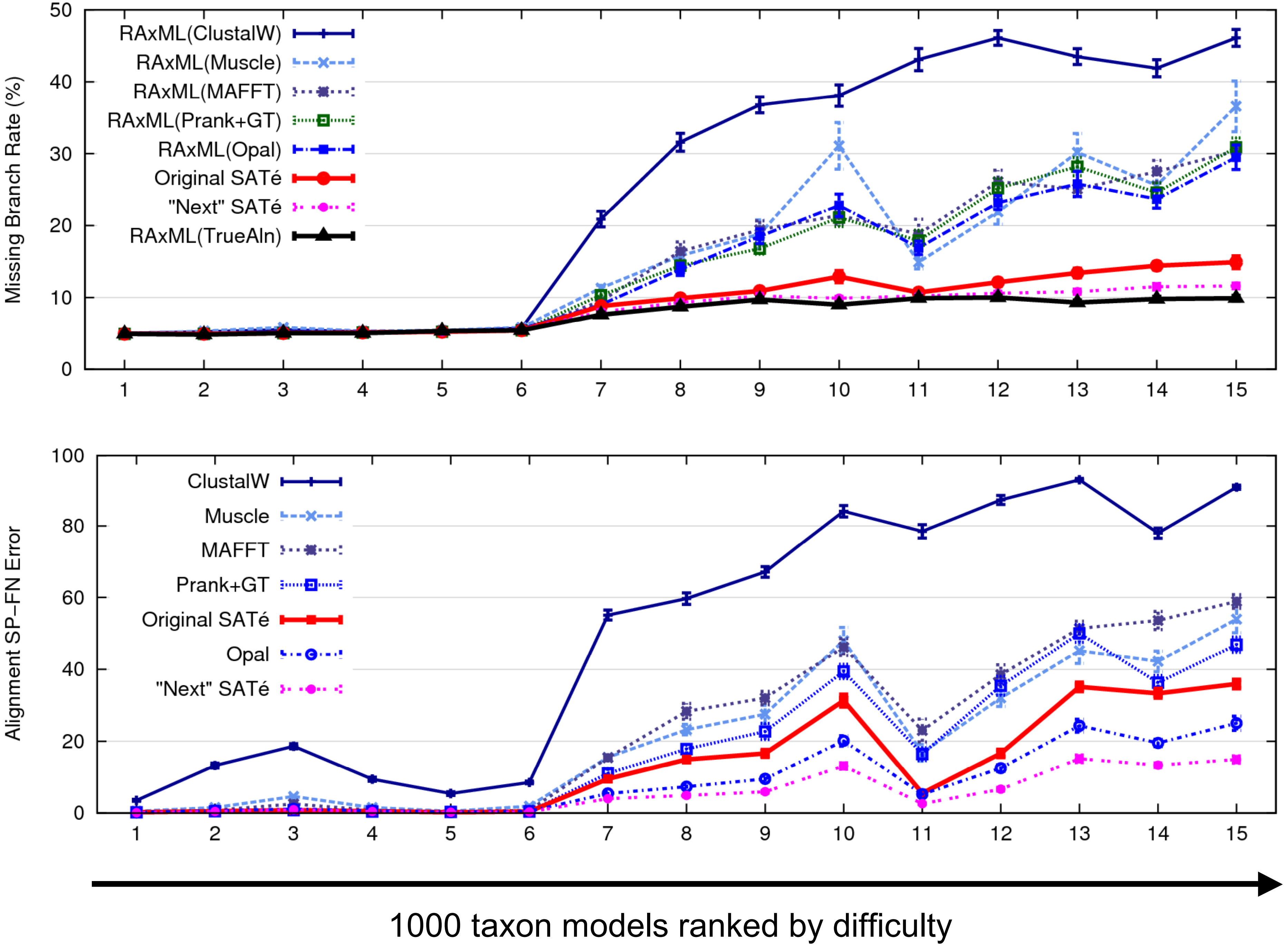
1000 taxon models, ordered by difficulty



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)
0
10
20
30
40
50
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Alignment SP–FN Error
ClustalW
Muscle
MAFFT
Prank+GT
Original SATé
Opal
"Next" SATé
0
20
40
60
80
100
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
1000 taxon models ranked by difficulty

Part II: SEPP

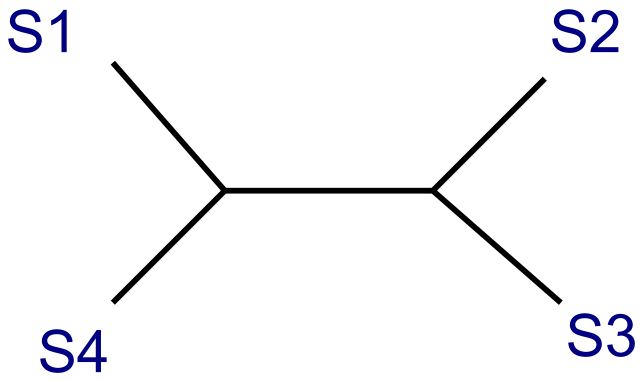
- SEPP: SATé-enabled Phylogenetic Placement, by Mirarab, Nguyen, and Warnow
- Pacific Symposium on Biocomputing, 2012 (special session on the Human Microbiome)

Phylogenetic Placement

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

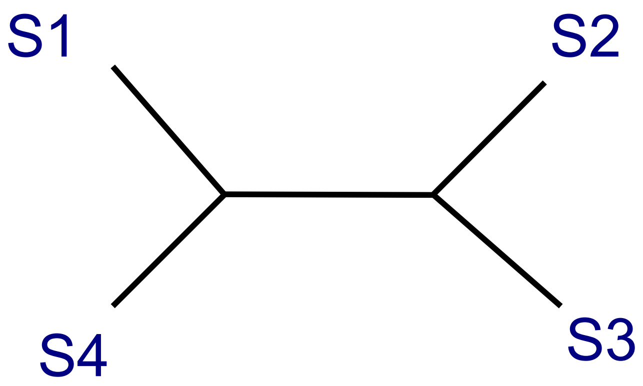
Align Sequence

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



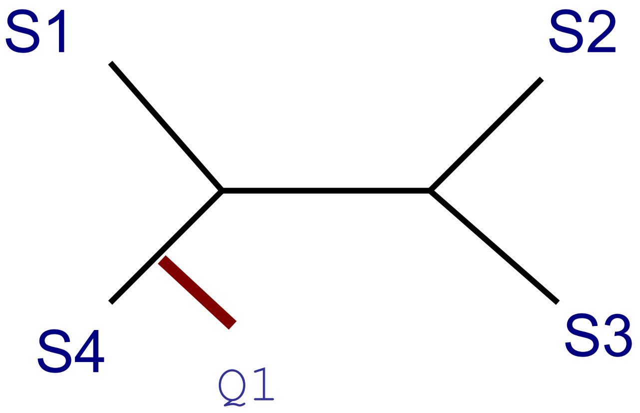
Align Sequence

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

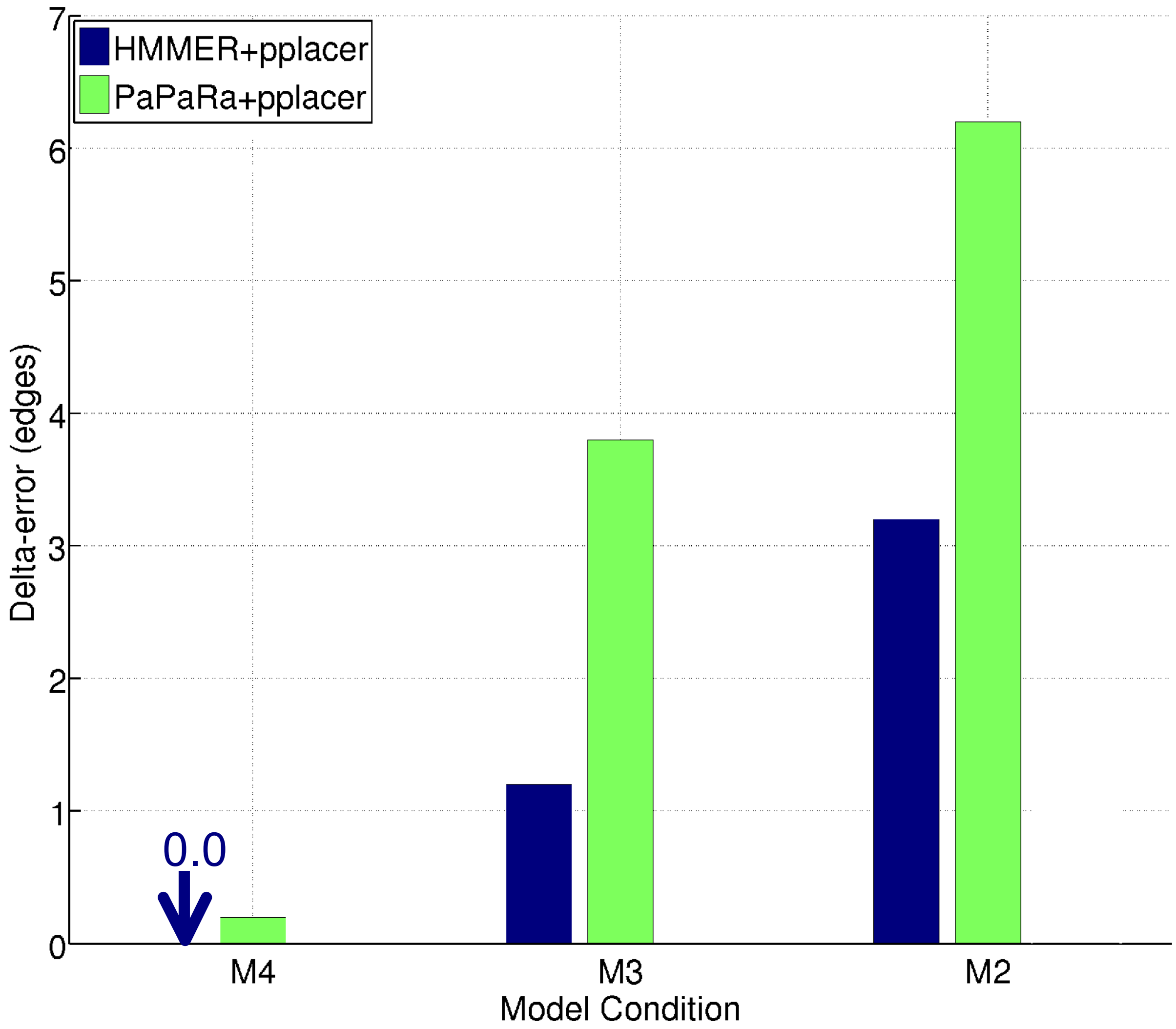


Phylogenetic Placement

- 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)

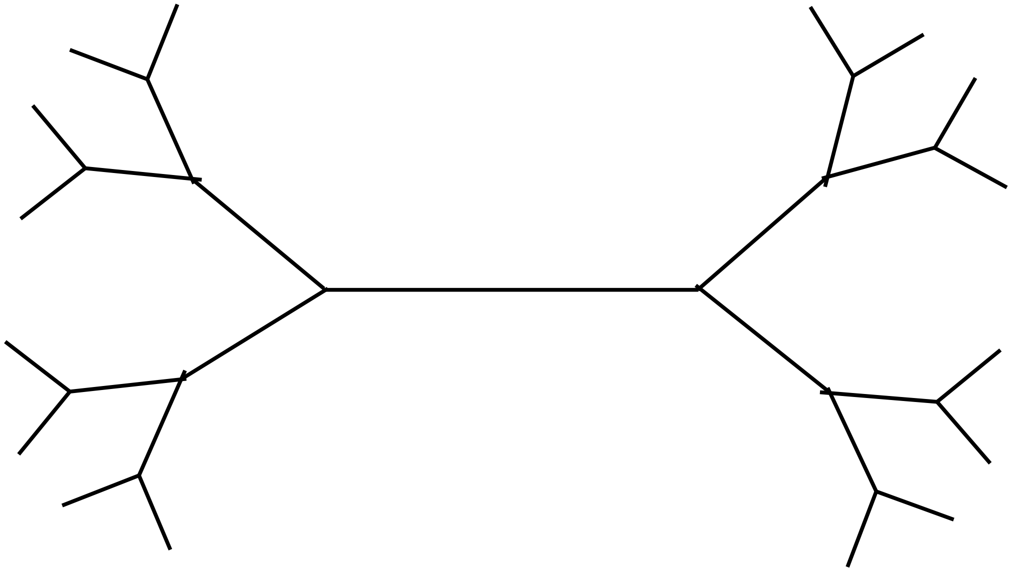
Note: pplacer and EPA use maximum likelihood

HMMER vs. PaPaRa

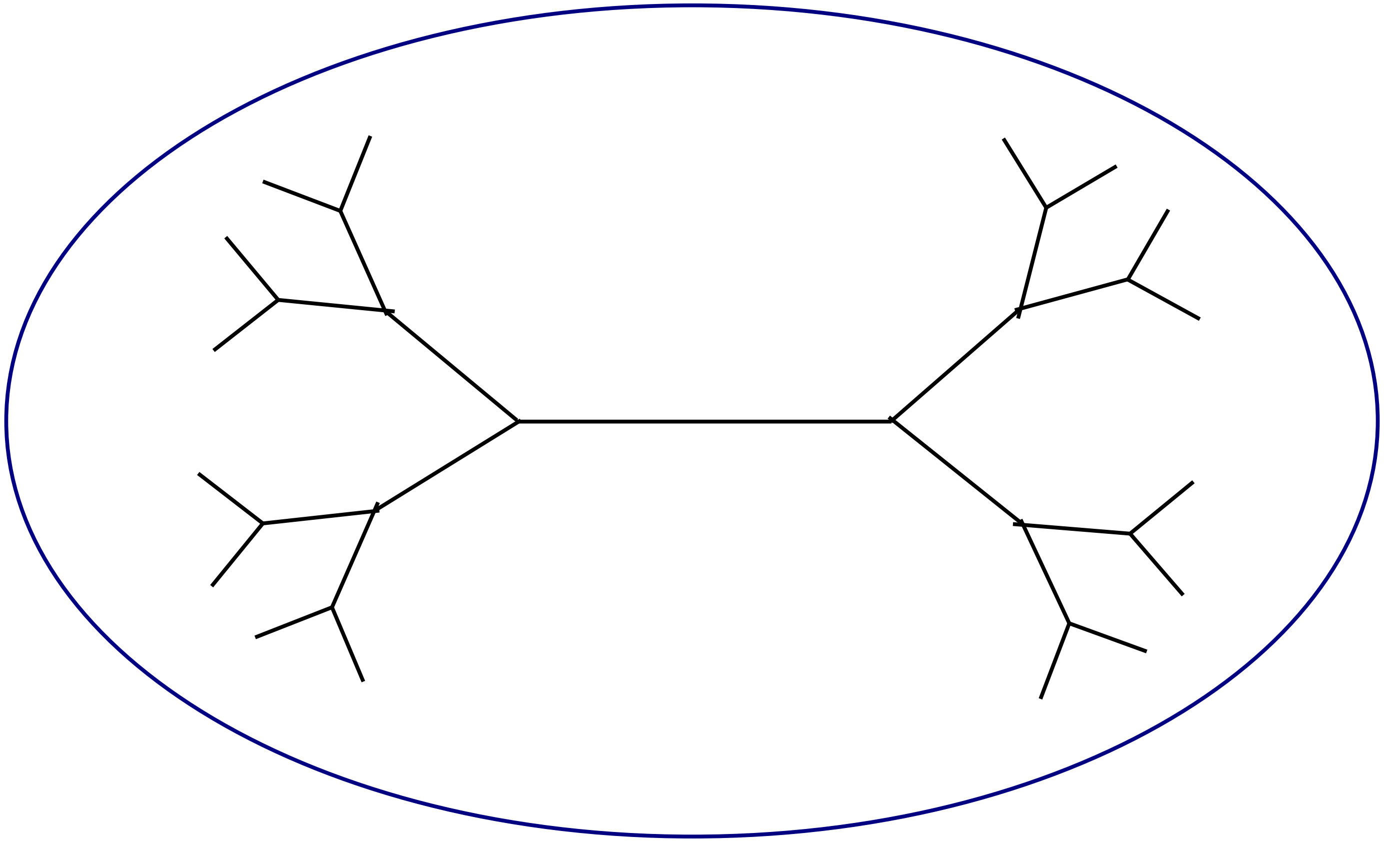


Increasing rate of evolution

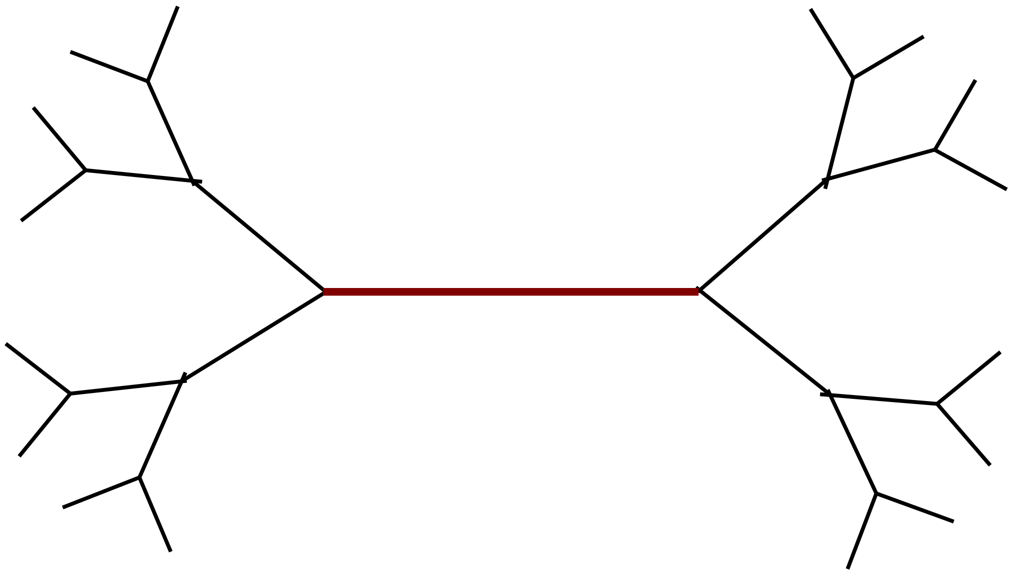
Insights from SATé



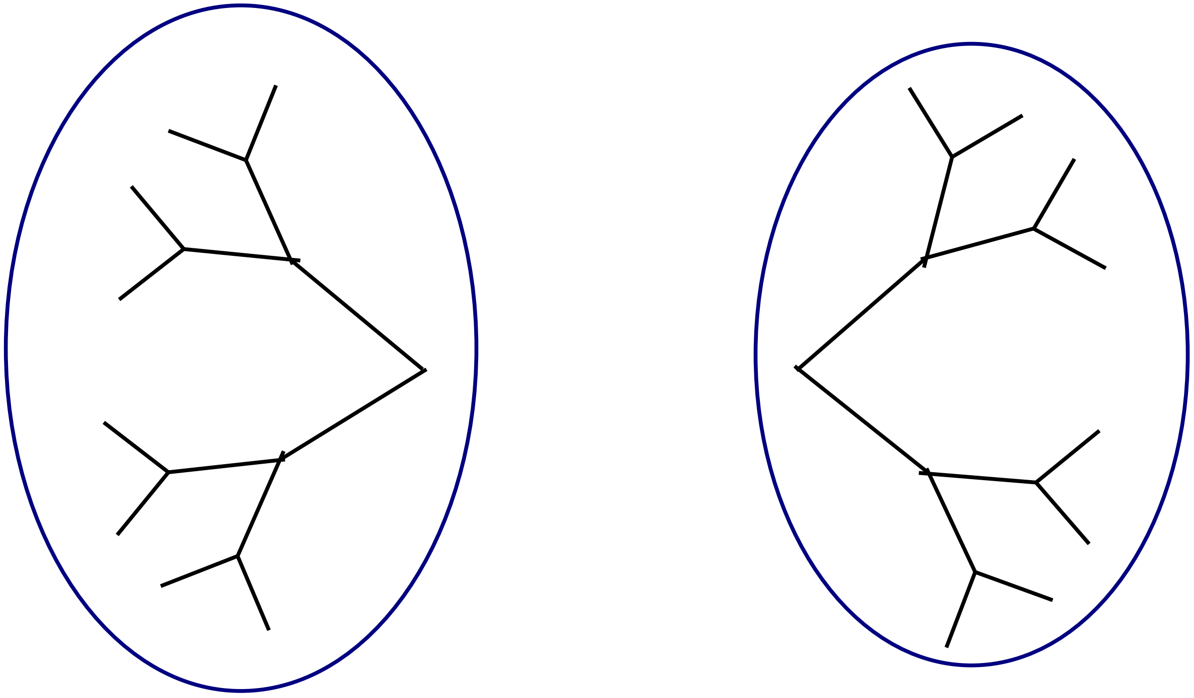
Insights from SATé



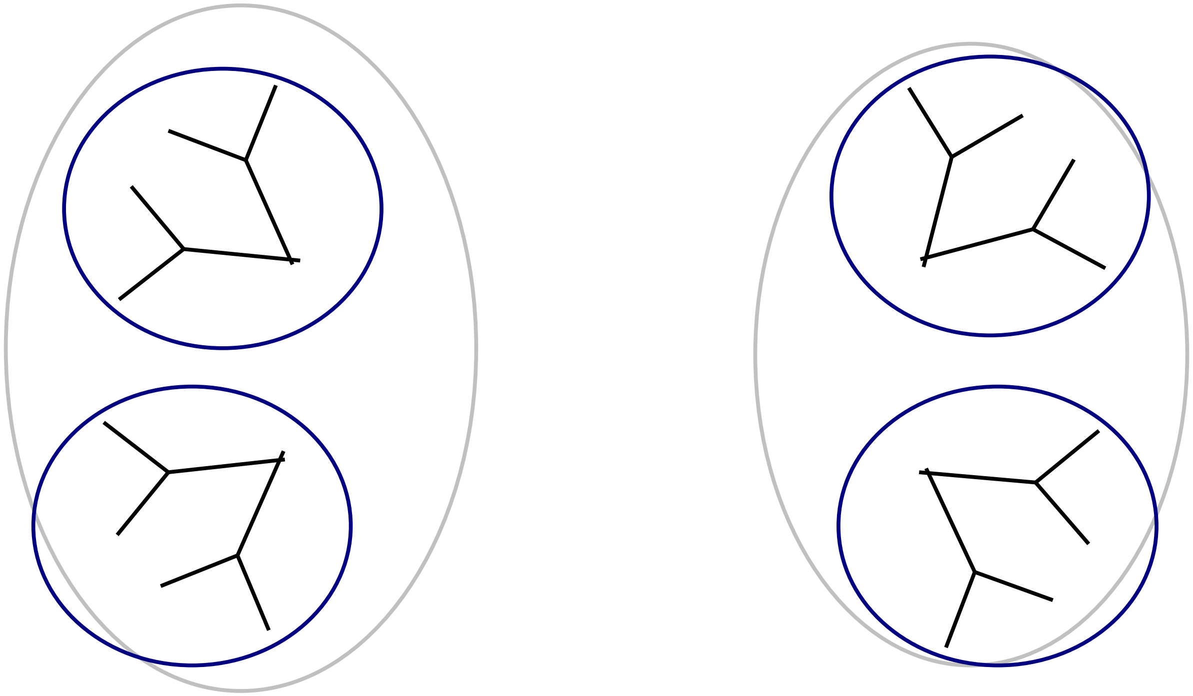
Insights from SATé



Insights from SATé



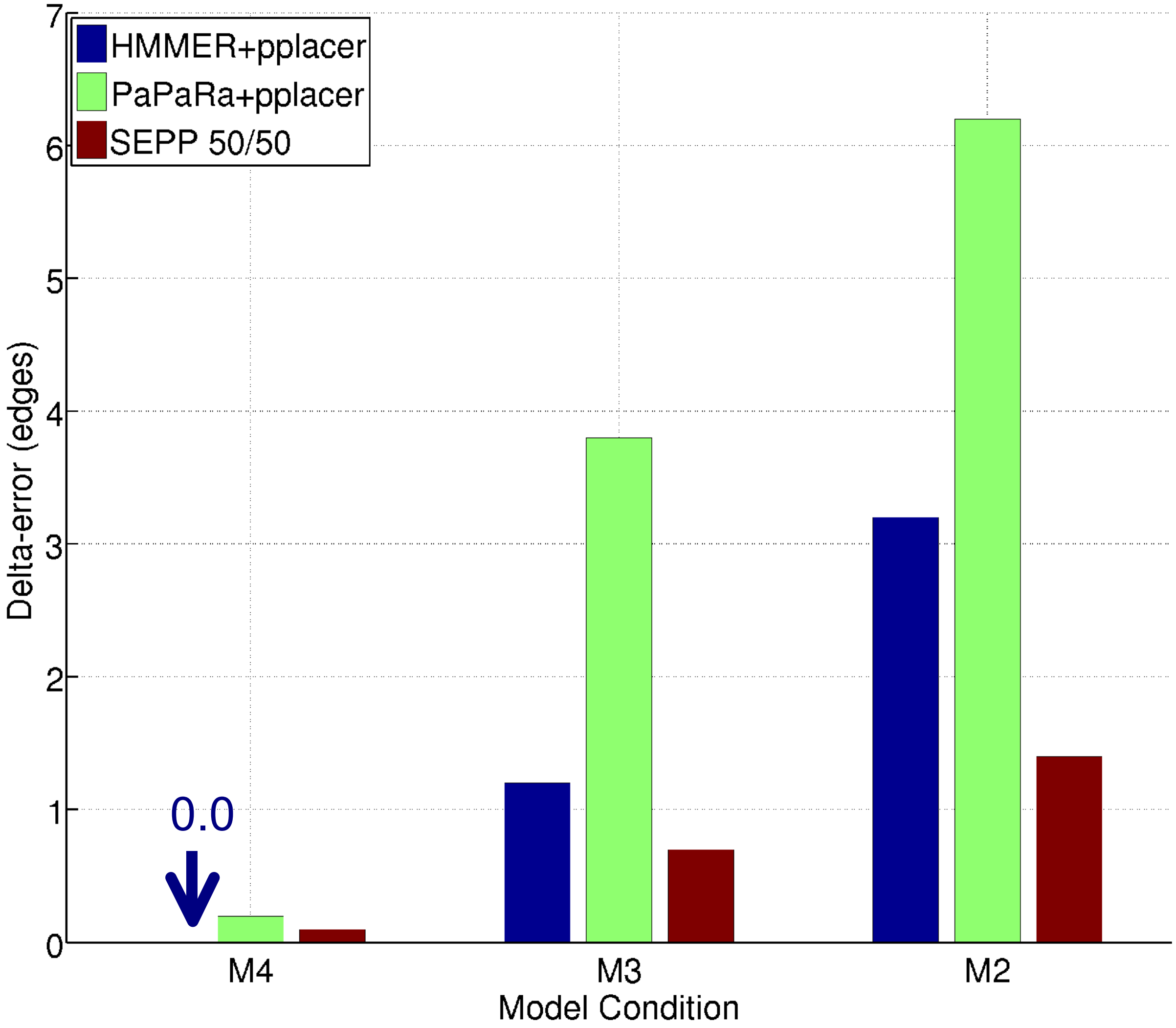
Insights from SATé



SEPP Parameter Exploration

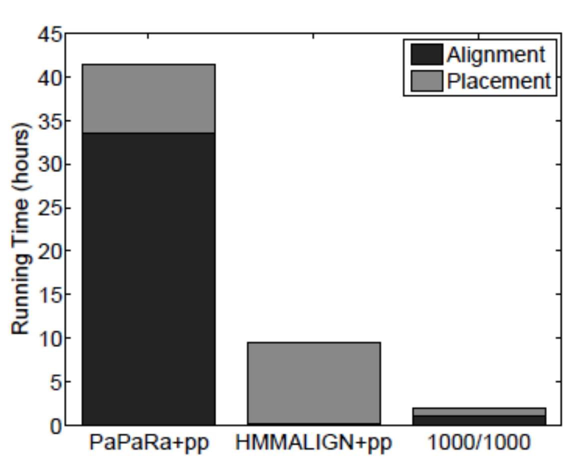
- Alignment subset size and placement subset size impact the accuracy, running time, and memory of SEPP
- 10% rule (subset sizes 10% of backbone) had best overall performance

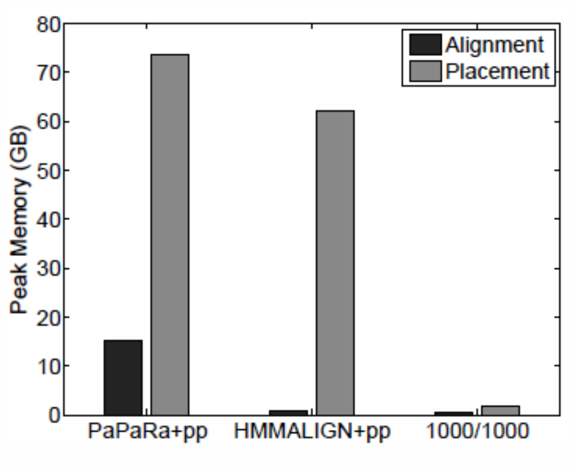
SEPP (10%-rule) on simulated data

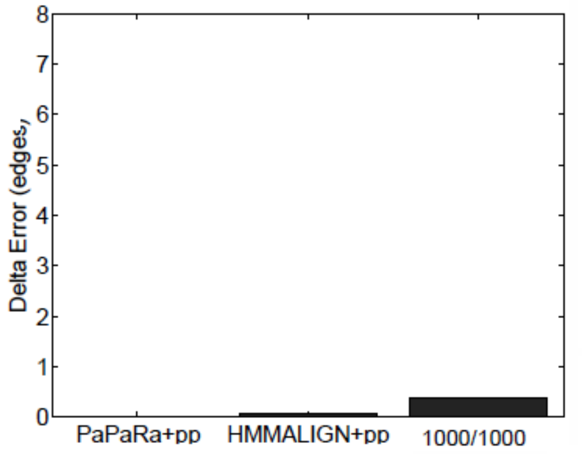


Increasing rate of evolution

SEPP (10%) on Biological Data

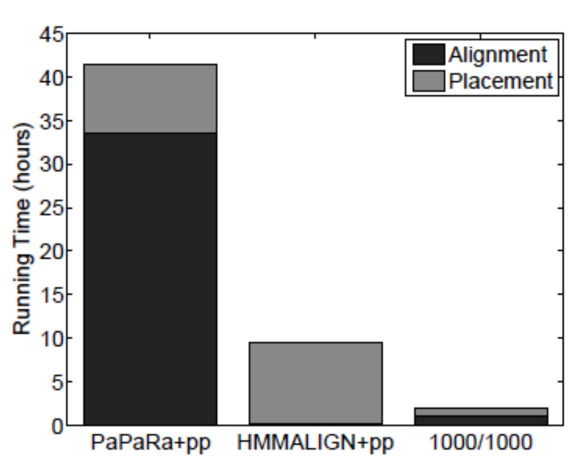


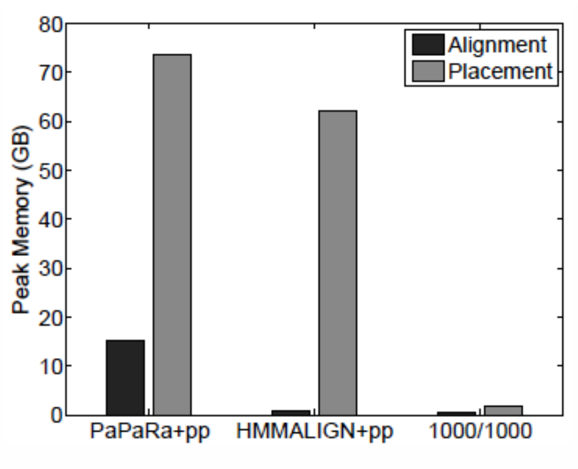


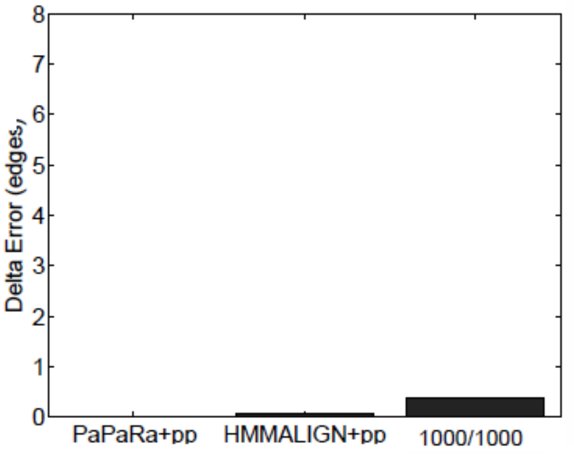


16S.B.ALL dataset, 13k curated backbone tree, 13k total fragments

SEPP (10%) on Biological Data







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

For 1 million fragments:

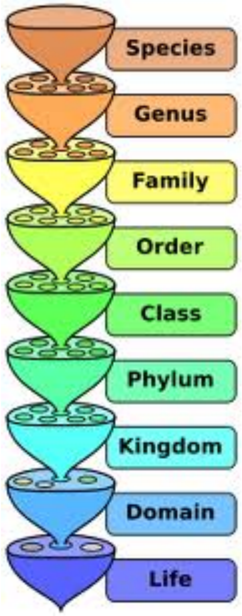
PaPaRa+pplacer: ~133 days

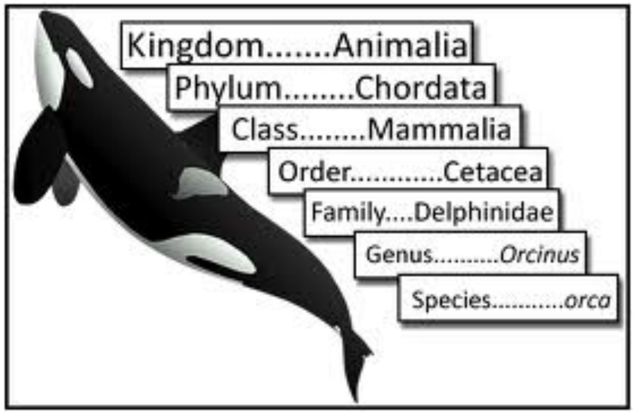
HMMALIGN+pplacer: ~30 days

SEPP 1000/1000: ~6 days

Part III: Taxon Identification

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 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)

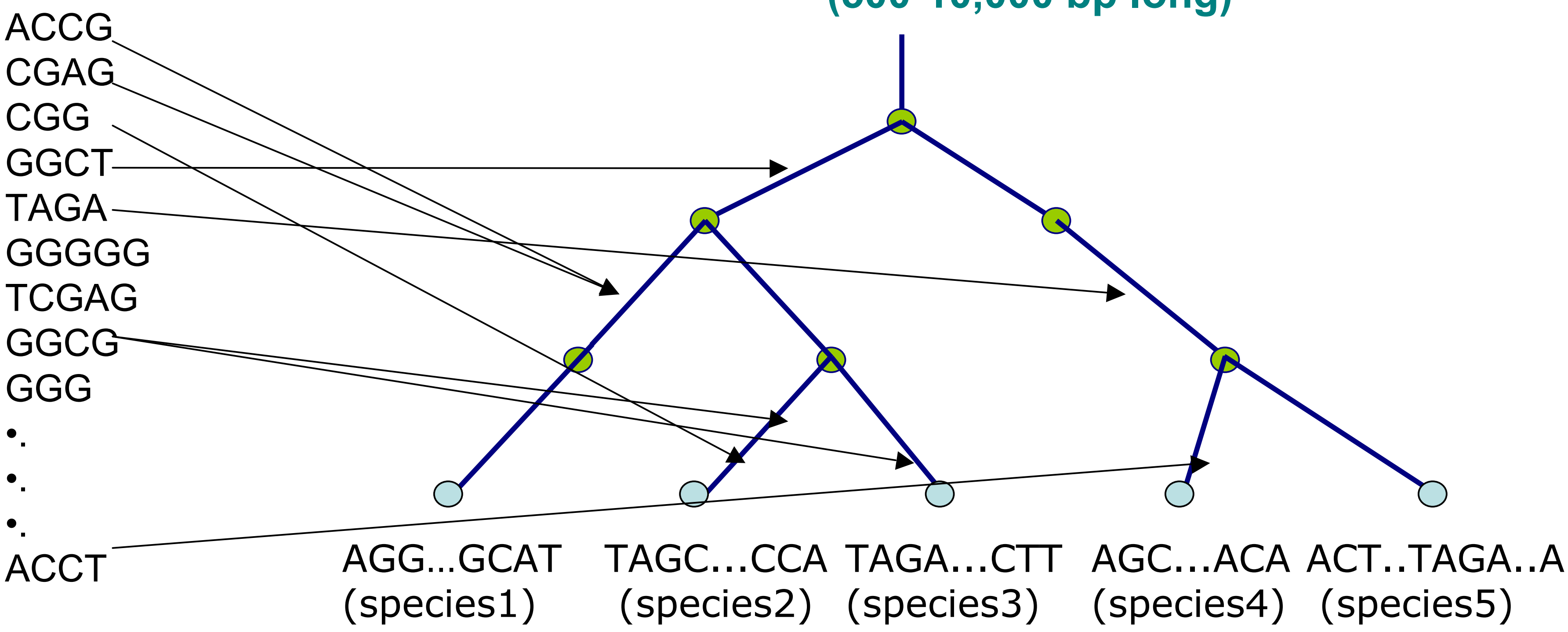
Thus, the best taxon identification methods have **high precision** (make accurate predictions), but **low sensitivity** (i.e., they **fail to classify** a large portion of reads) even at higher taxonomy levels.

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)



ACCG
CGAG
CGG
GGCT
TAGA
GGGGG
TCGAG
GGCG
GGG
•.
•.
•.
ACCT

AGG...GCAT
(species1)

TAGC...CCA
(species2)

TAGA...CTT
(species3)

AGC...ACA
(species4)

ACT..TAGA..A
(species5)

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)

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

- Find **2 or more** backbone alignment/tree pairs of full-length sequences
- For each backbone alignment/tree pair:
  - Compute **many** extended alignments (one for each alignment subset, but using statistical support computed using HMMER to select enough subsets to cover 95% of the probability).
  - For each extended alignment, use pplacer statistical support values to place fragment into taxonomy (each produces multiple placements). Collect all placements generated over all extended alignments, and find clade in taxonomy containing 99% of the probability.

- Classify each fragment at the **LCA** of all placements obtained for the fragment

TIPP version 2

- Find **2 or more** backbone alignment/tree pairs of full-length sequences
- For each backbone alignment/tree pair:
    - Compute **many** extended alignments (one for each alignment subset, but using statistical support computed using HMMER to select enough subsets to cover 95% of the probability).
    - For each extended alignment, use pplacer statistical support values to place fragment into taxonomy (each produces multiple placements). Collect all placements generated over all extended alignments, and find clade in taxonomy containing 99% of the probability.

- Classify each fragment at the **LCA** of all placements obtained for the fragment

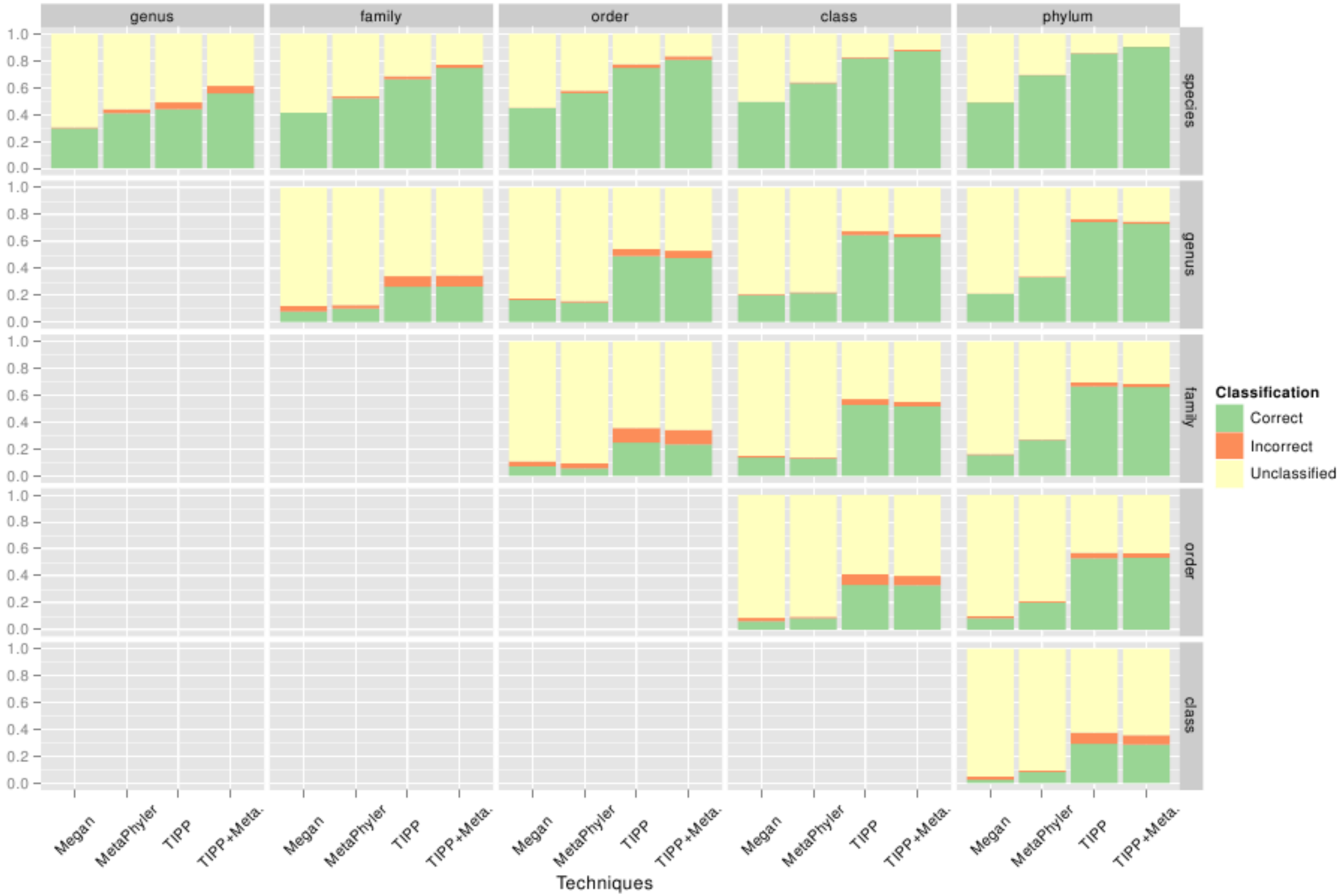
TIPP version 2 dramatically reduces false positive rate with small reduction in true positive rate by considering uncertainty, using statistical techniques.

TIPP+Metaphyler

- Use Metaphyler to perform initial placement of read into taxonomy

- Use TIPP to modify the placement, moving the read further into the clade identified by Metaphyler

Results on rpsB gene (60 bp)



Summary

- SATé gives better alignments and trees than standard alignment estimation methods
- SEPP can enable alignment of short (fragmentary) sequences into alignments of full-length sequences, and phylogenetic placement into gene trees or taxonomies
- TIPP enables taxon identification of short reads -- not limited to 31 marker genes, and no training is needed.

Overall message

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

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*

Phylogenetic “boosters” (meta-methods)

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

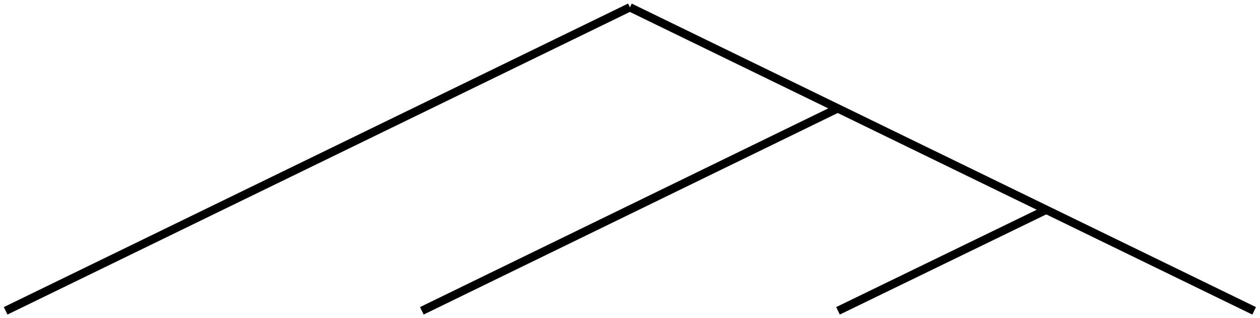
Examples:

- 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)

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
  - TIPP: Siavash Mirarab, Nam Nguyen, Bo Liu, and Mihai Pop

Phylogeny (evolutionary tree)

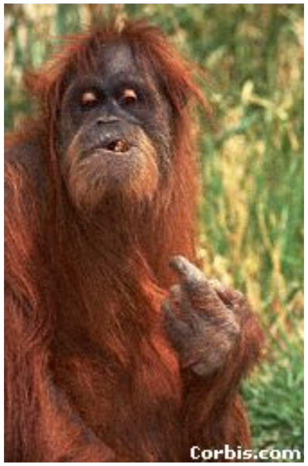


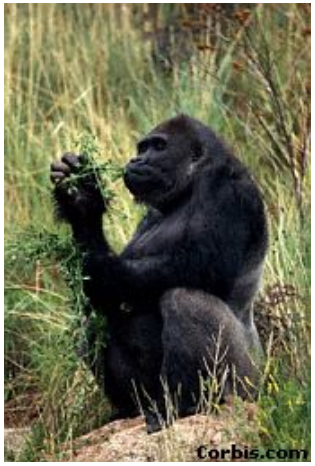
Orangutan

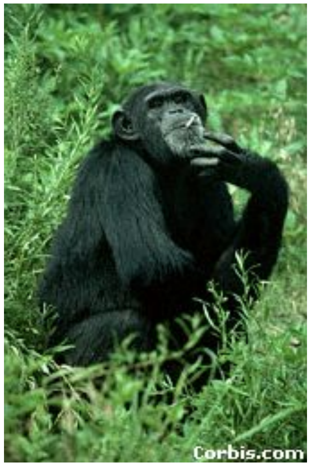
Gorilla

Chimpanzee

Human







From the Tree of the Life Website,
University of Arizona

How did life evolve on earth?

Courtesy of the Tree of Life project