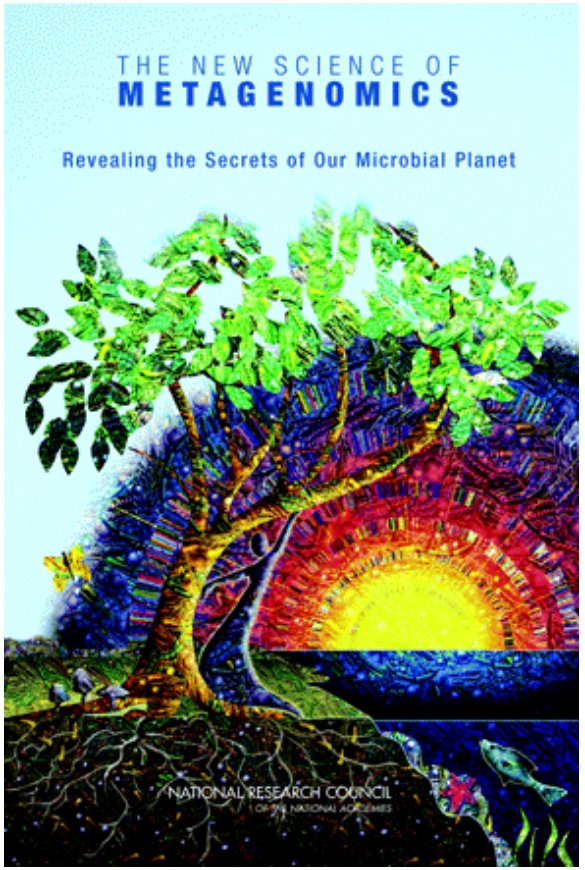
Challenge and novel aproaches for multiple sequence alignment and phylogenetic estimation

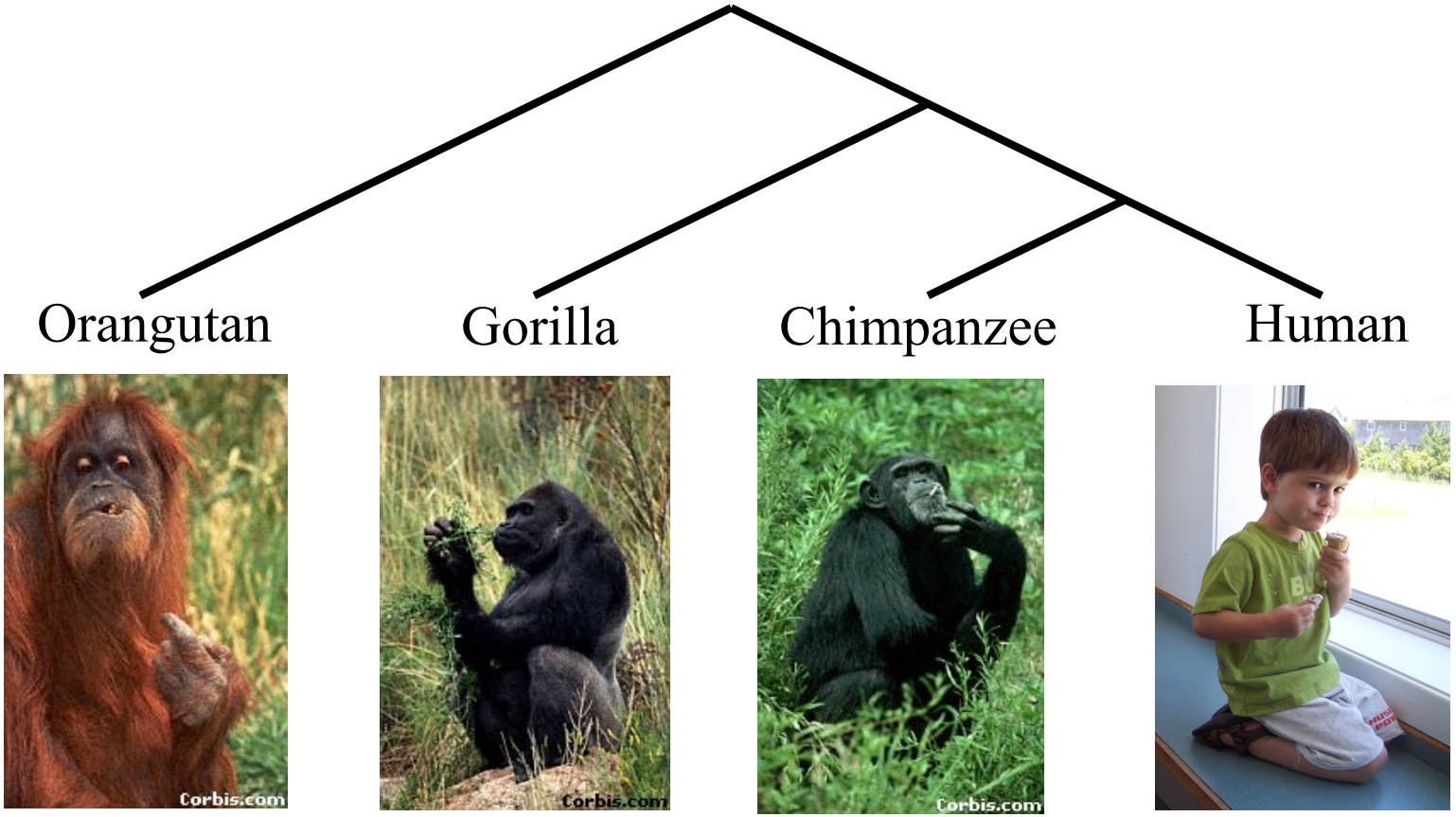
Tandy Warnow
Department of Computer Science
The University of Texas at Austin

Computational Phylogenetics and Metagenomics



Courtesy of the Tree of Life project

Phylogeny (evolutionary tree)



From the Tree of the Life Website,
University of Arizona

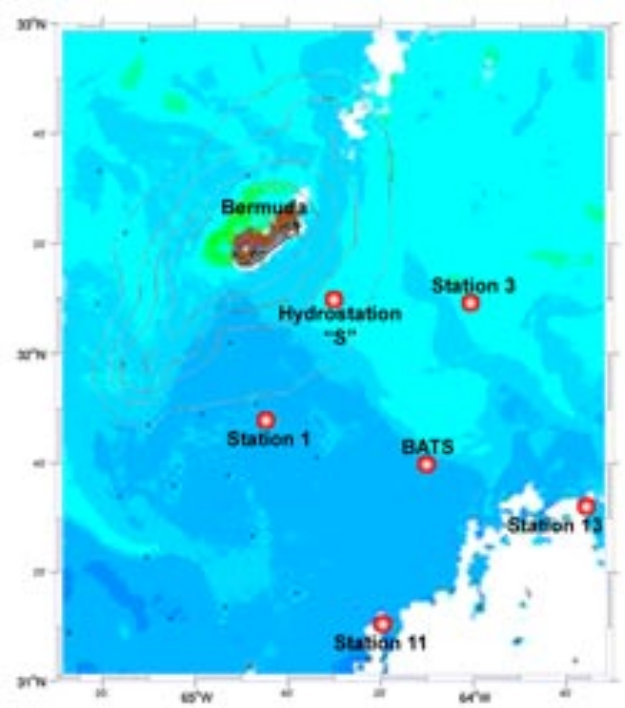
How did life evolve on earth?

Courtesy of the Tree of Life project

Metagenomics:

Venter et al., Exploring the Sargasso Sea:

Scientists Discover One Million New Genes in Ocean Microbes



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).

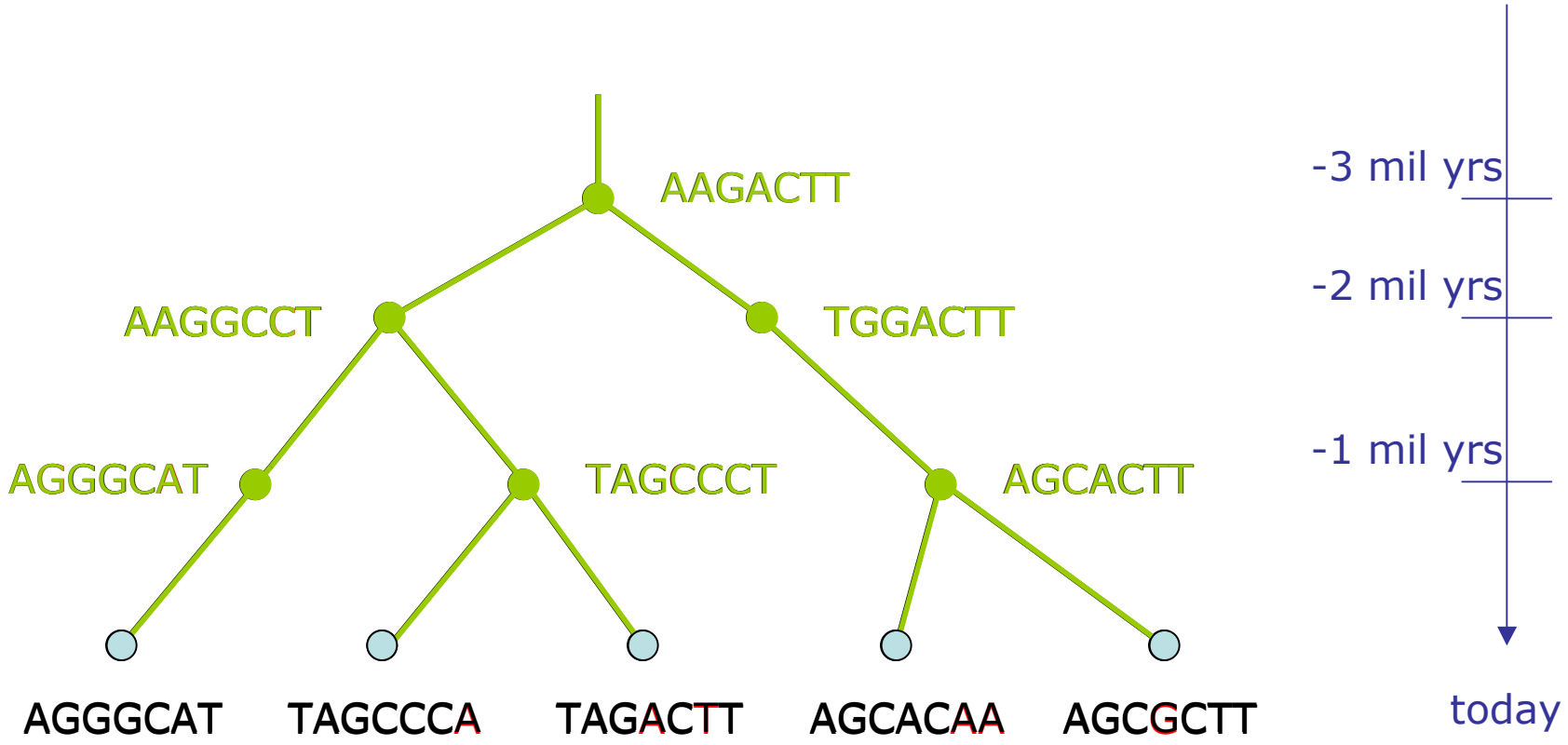
Phylogenetic “boosters” (meta-methods)

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

Examples:

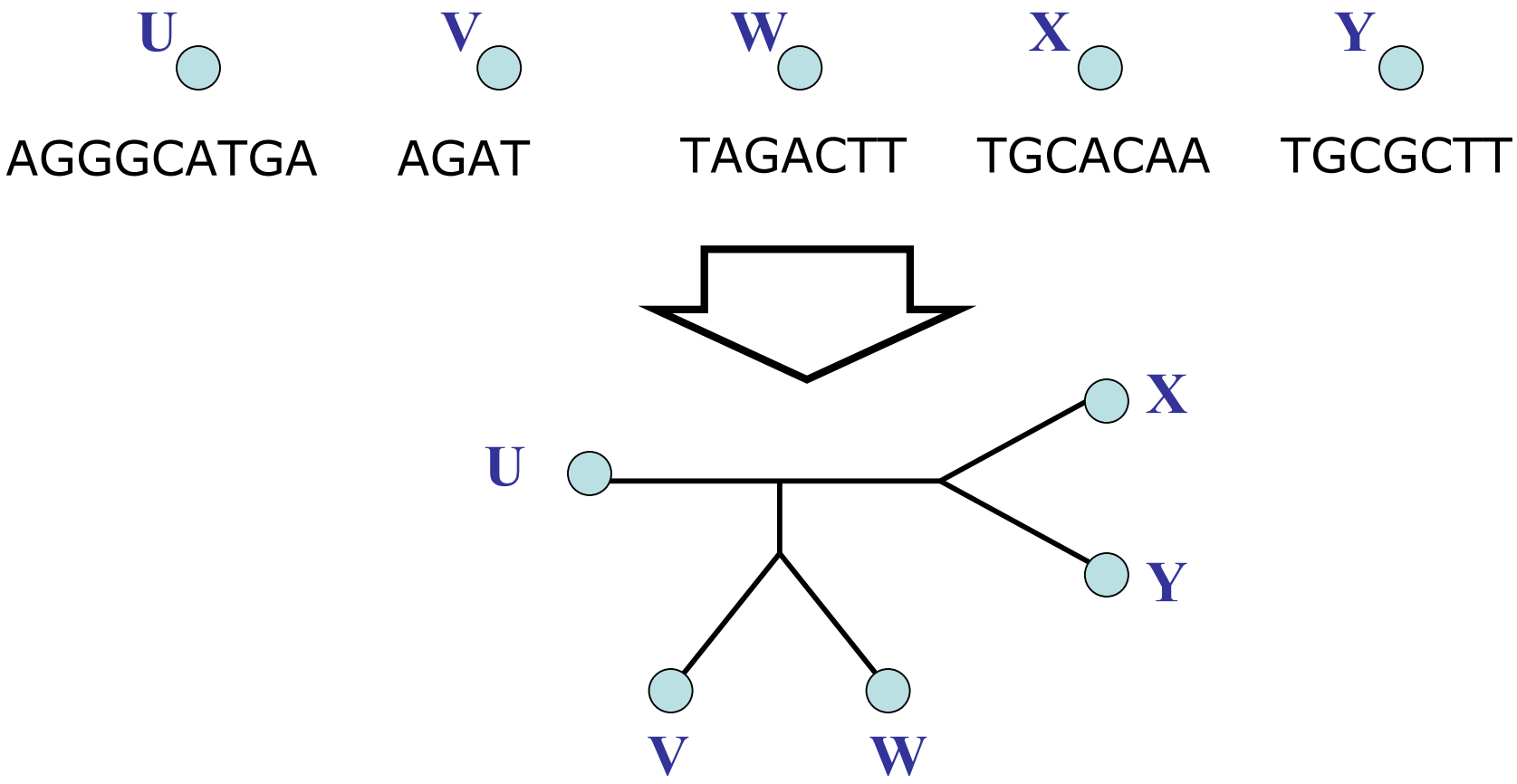
- DCM-boosting for distance-based methods (1999)
- DCM-boosting for heuristics for NP-hard problems (1999)
- SATé-boosting for alignment methods (2009)
- SuperFine-boosting for supertree methods (2011)
- DACTAL-boosting: almost alignment-free phylogeny estimation methods (2011)
- SEPP-boosting for phylogenetic placement of short sequences (2012)
- TIPP-boosting for metagenomic taxon identification (2013)

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

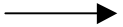
```
S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC
S3 = TAGCTGACCGC
S4 = TCACGACCGACA
```



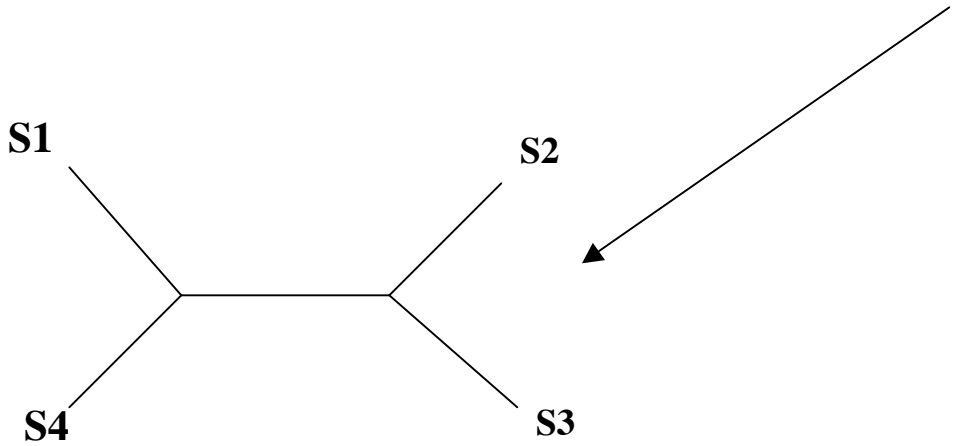
```
S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAC--GACCGC--
S3 = TAG-CT-------GACCGC--
S4 = --------TCAC--GACCGACA
```

Phase 2: Construct tree

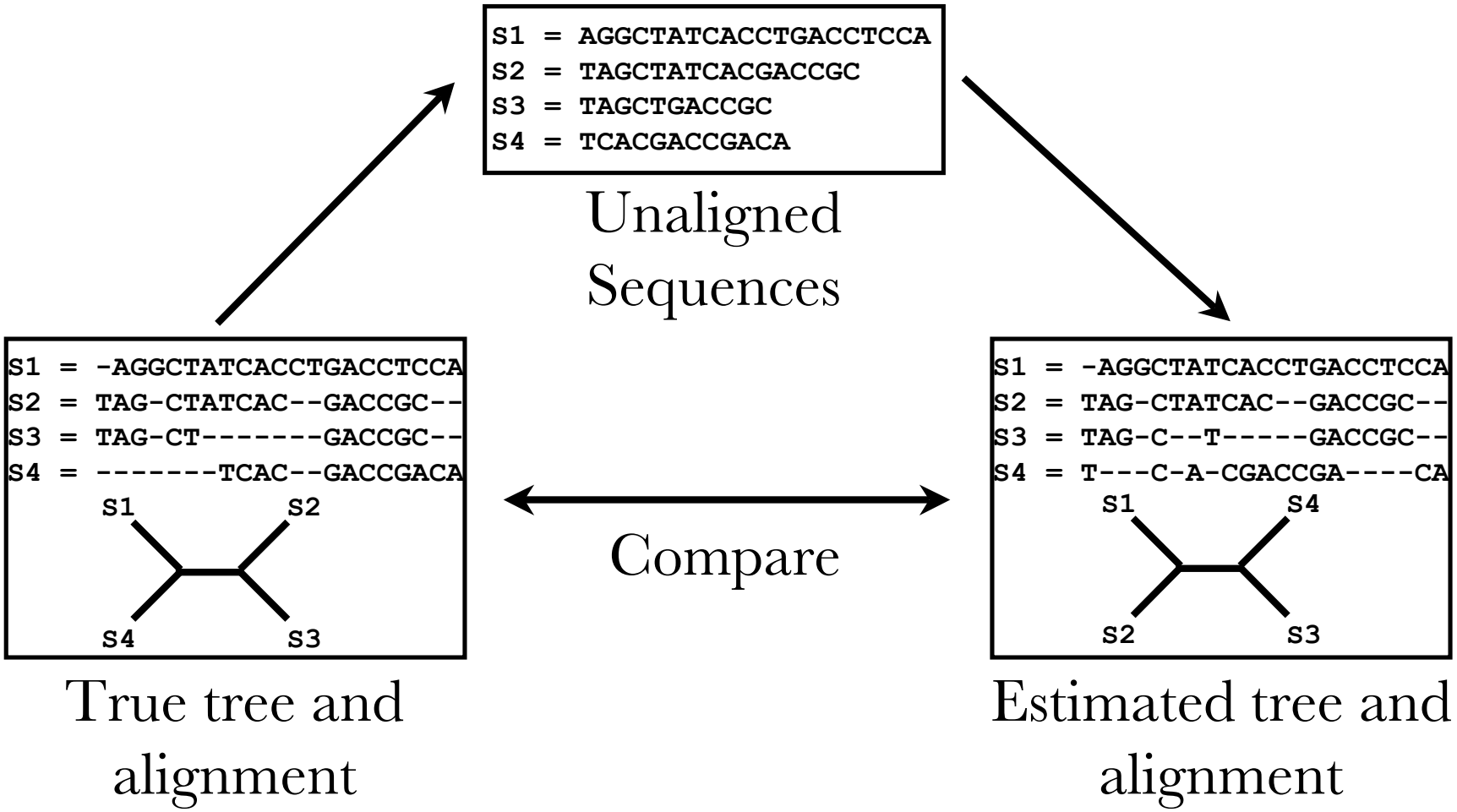
```
S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC
S3 = TAGCTGACCGC
S4 = TCACGACCGACA
```



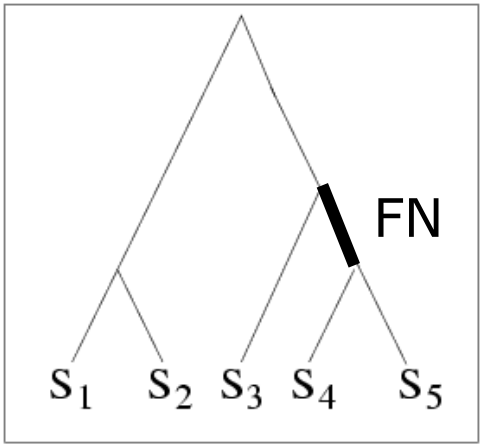
```
S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAC--GACCGC--
S3 = TAG-CT-------GACCGC--
S4 = -------TCAC--GACCGACA
```



Simulation Studies



Quantifying Error



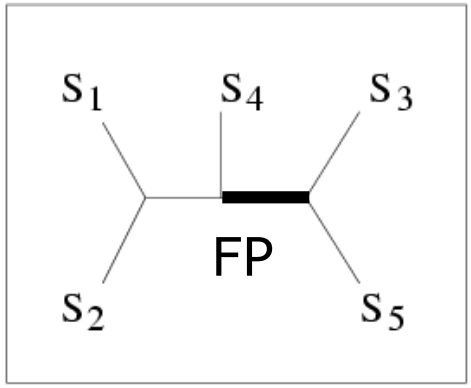
TRUE TREE

S_1	ACAATTAGAAC
S_2	ACCCTTAGAAC
S_3	ACCATTCCAAC
S_4	ACCAGACCAAC
S_5	ACCAGACCGGA

DNA SEQUENCES

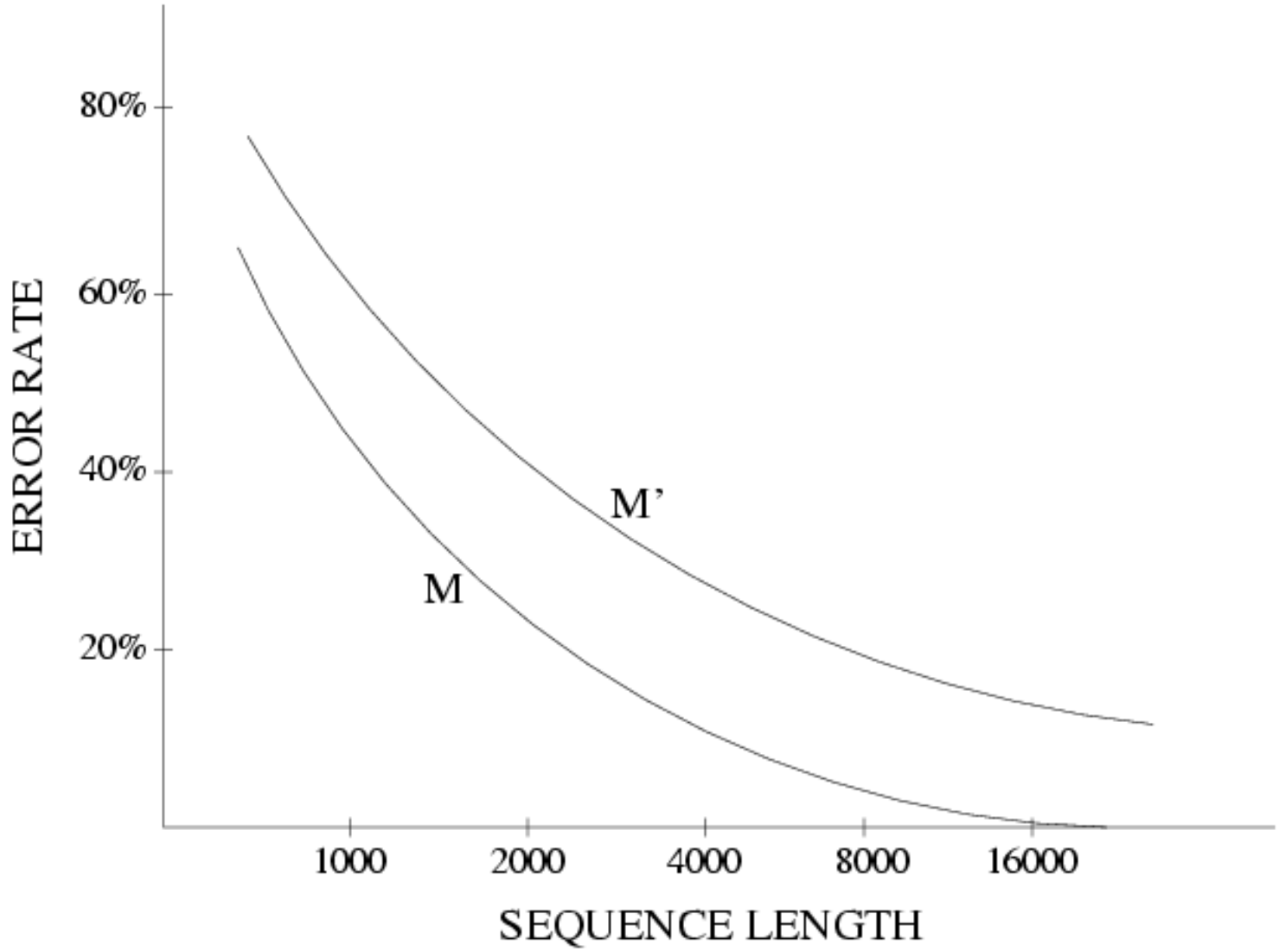
FN: false negative
(missing edge)
FP: false positive
(incorrect edge)

50% error rate



INFERRED TREE

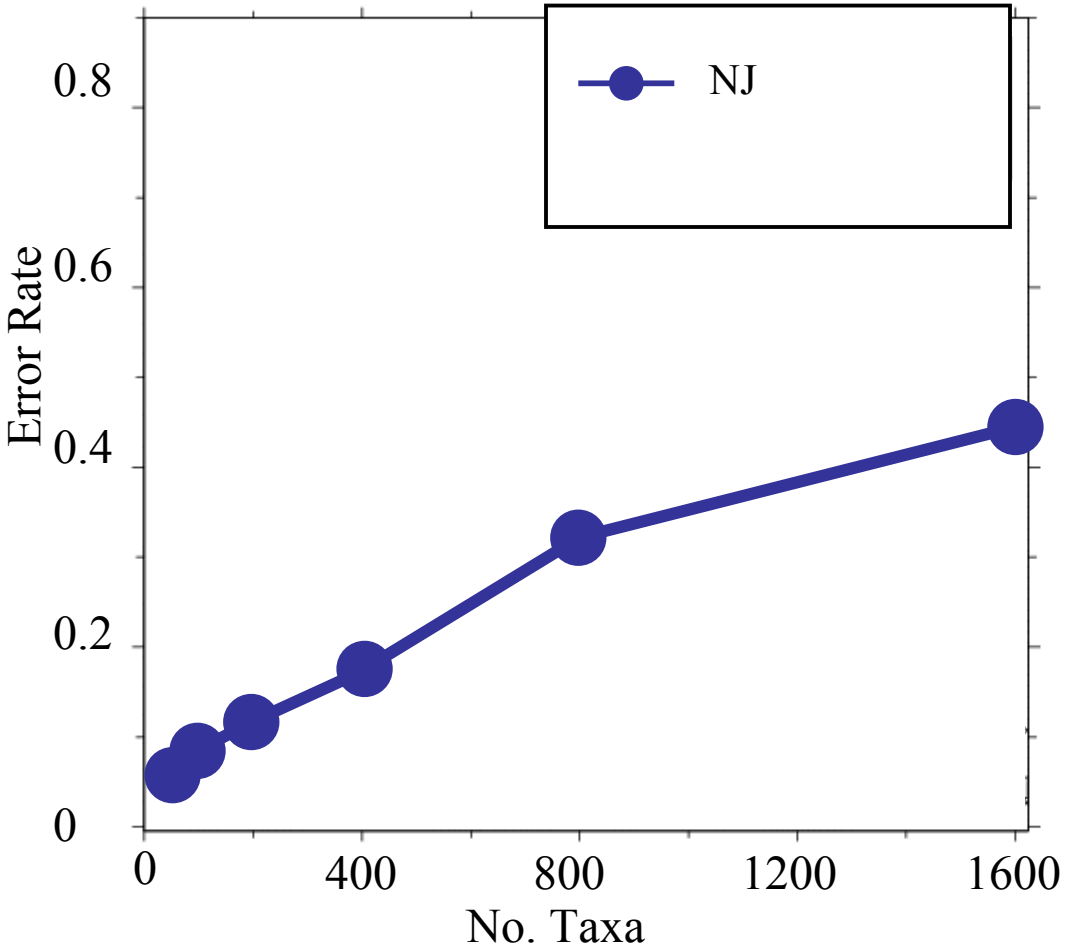
Statistical consistency and convergence rates



Part I: “Fast-Converging Methods”

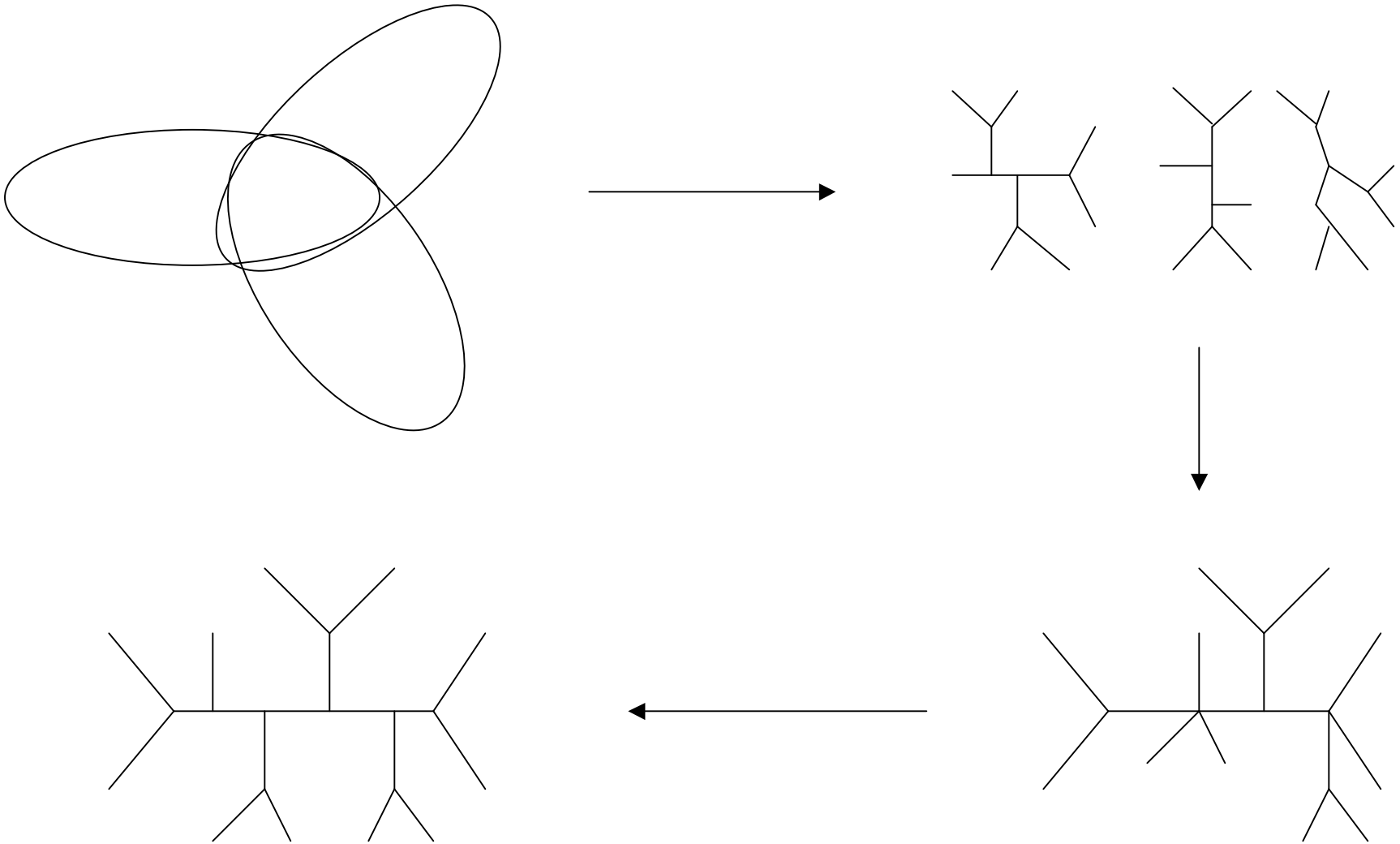
- Basic question: how much data does a phylogeny estimation method need to produce the true tree with high probability?

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



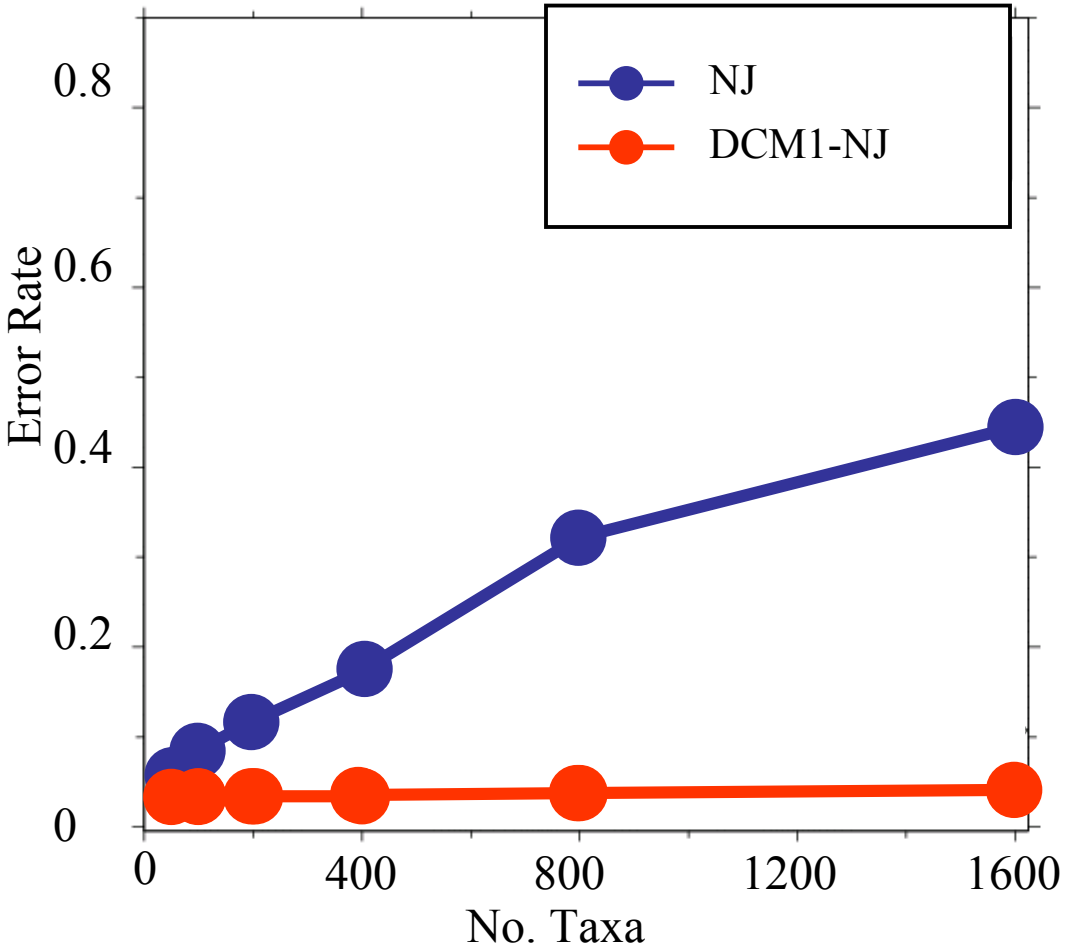
Theorem (Atteson): Exponential sequence length requirement for Neighbor Joining!

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

Part II: SATé

Simultaneous Alignment and Tree Estimation

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

Liu et al., Systematic Biology 2012

Public software distribution (open source) through the Mark Holder's group at the University of Kansas

Two-phase estimation

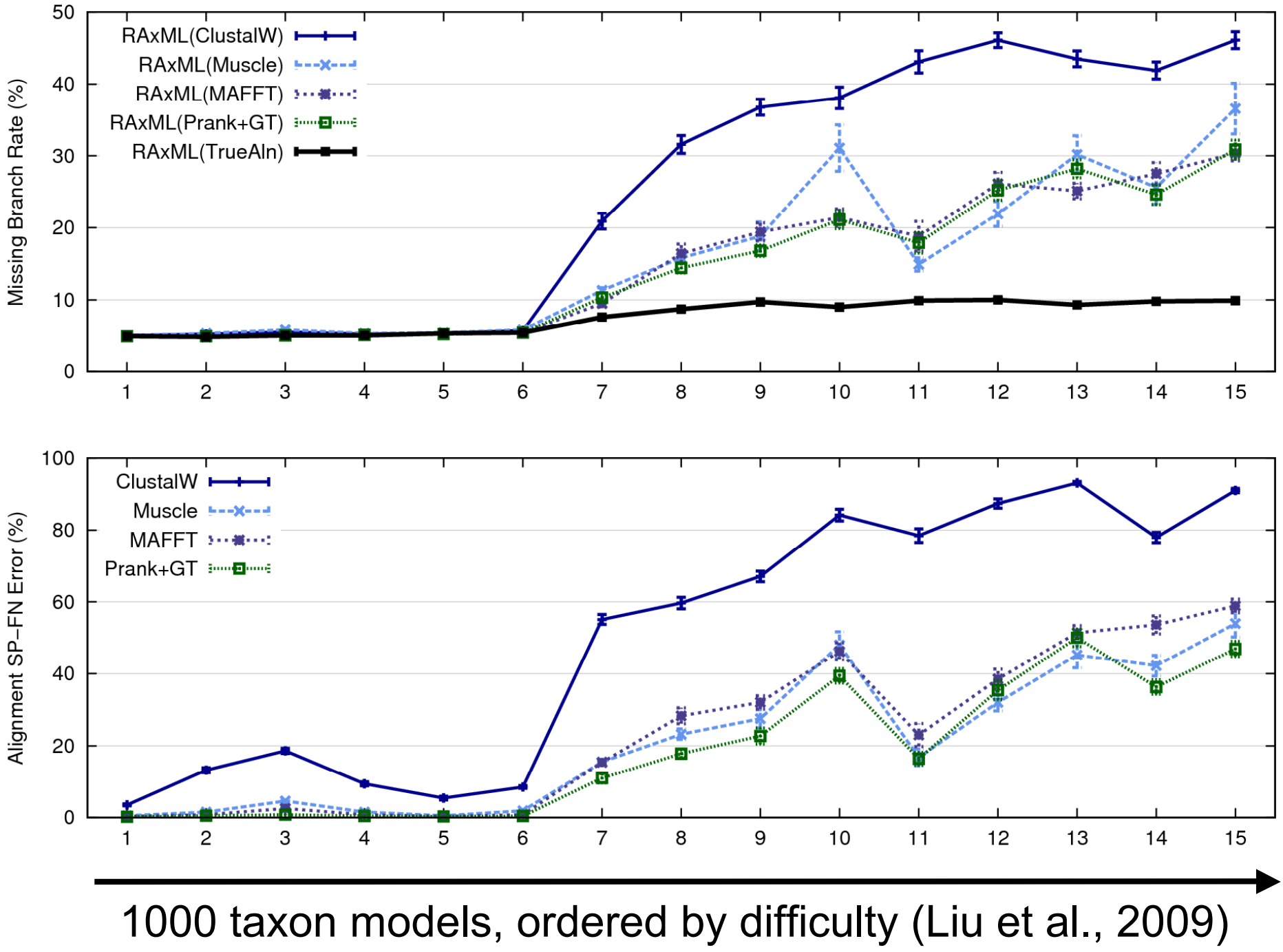
Alignment methods

- 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*


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)

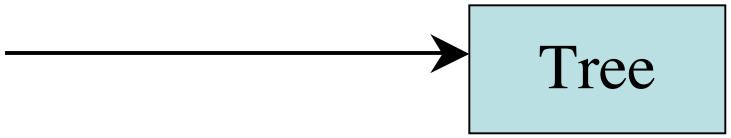
Problems

- 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)

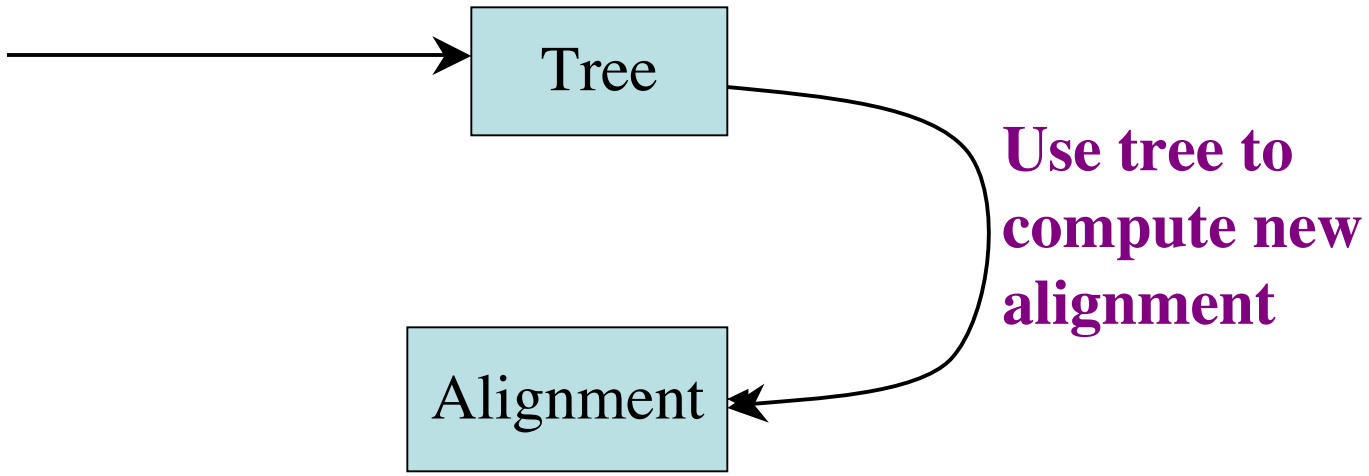
SATé Algorithm

Obtain initial alignment
and estimated ML tree



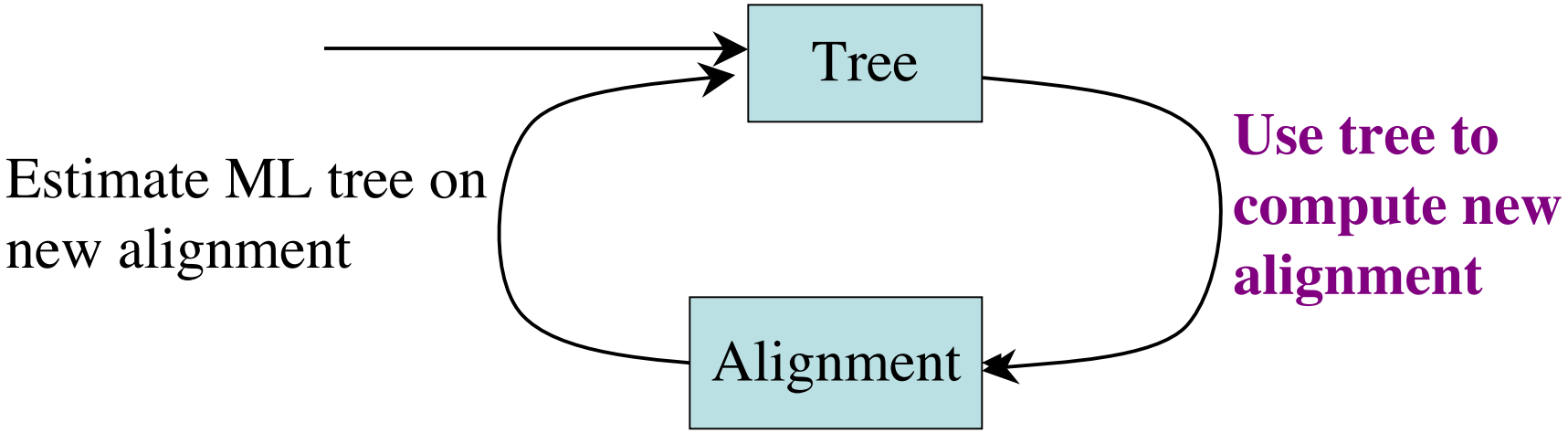
SATé Algorithm

Obtain initial alignment
and estimated ML tree

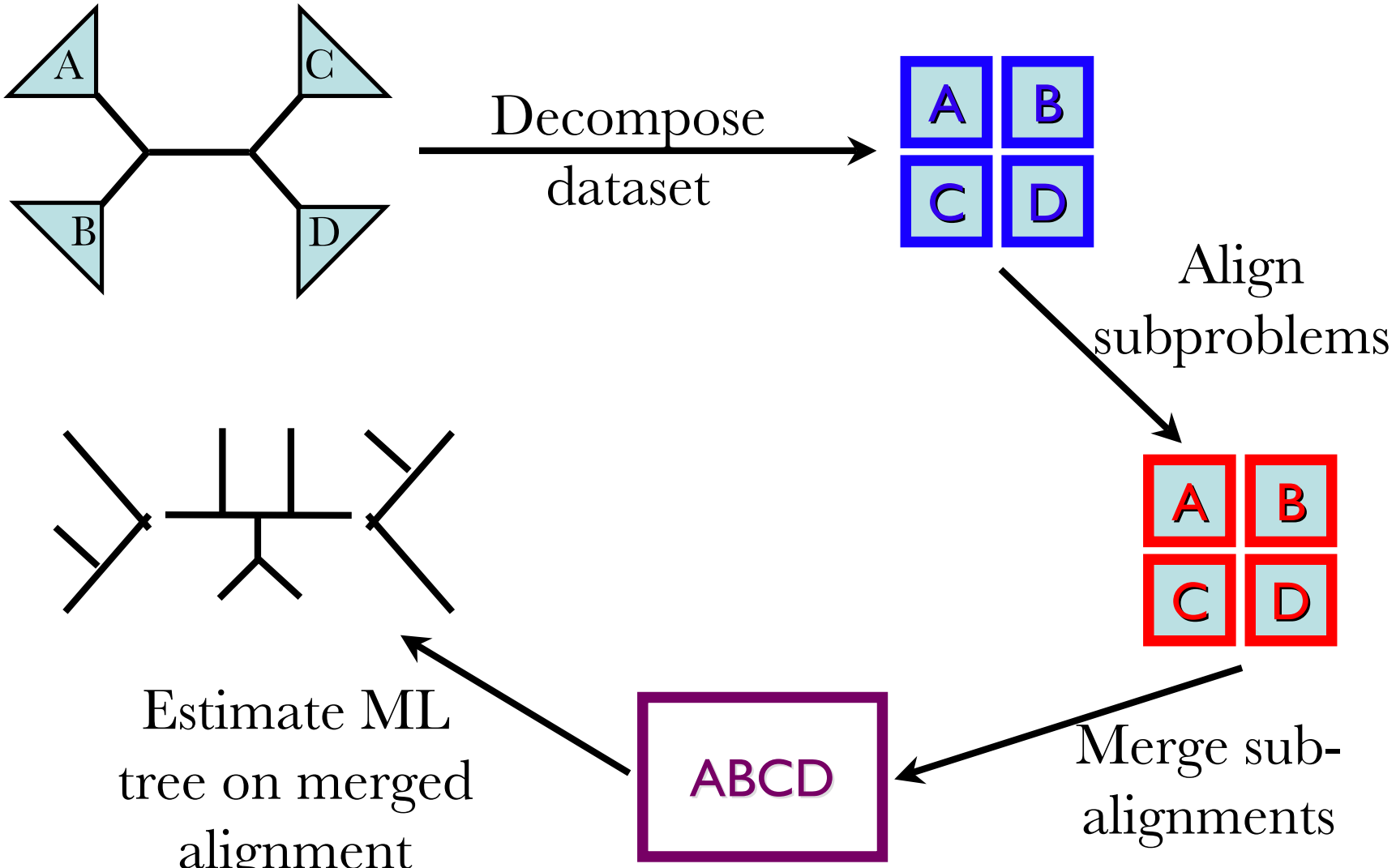


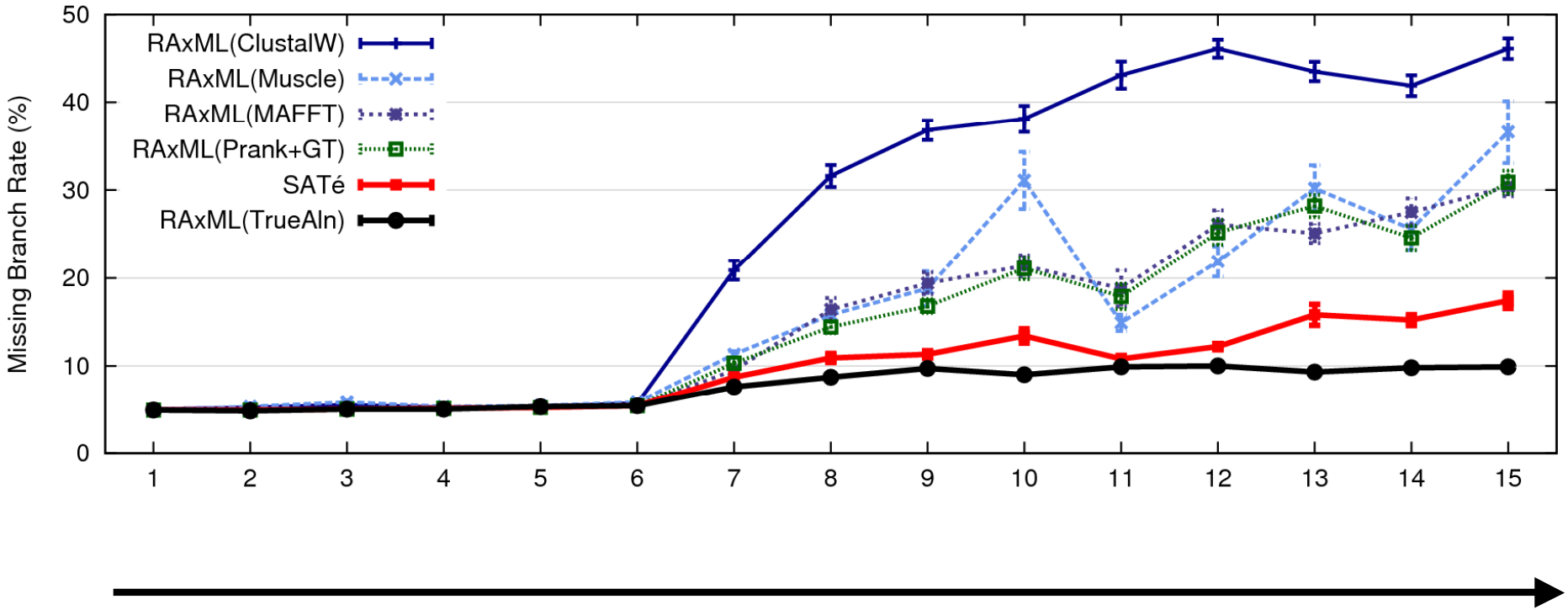
SATé Algorithm

Obtain initial alignment
and estimated ML tree



Re-aligning on a tree

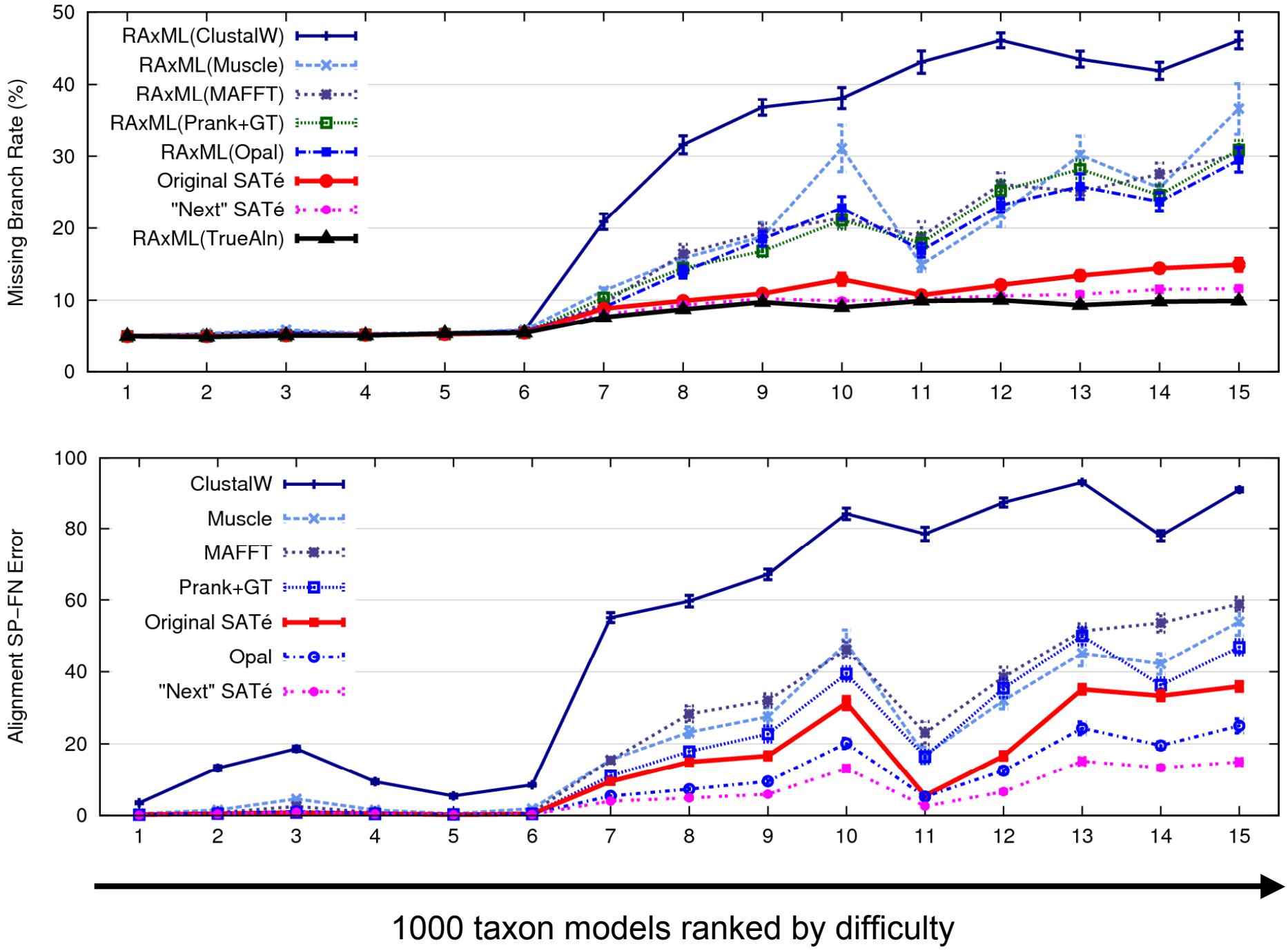




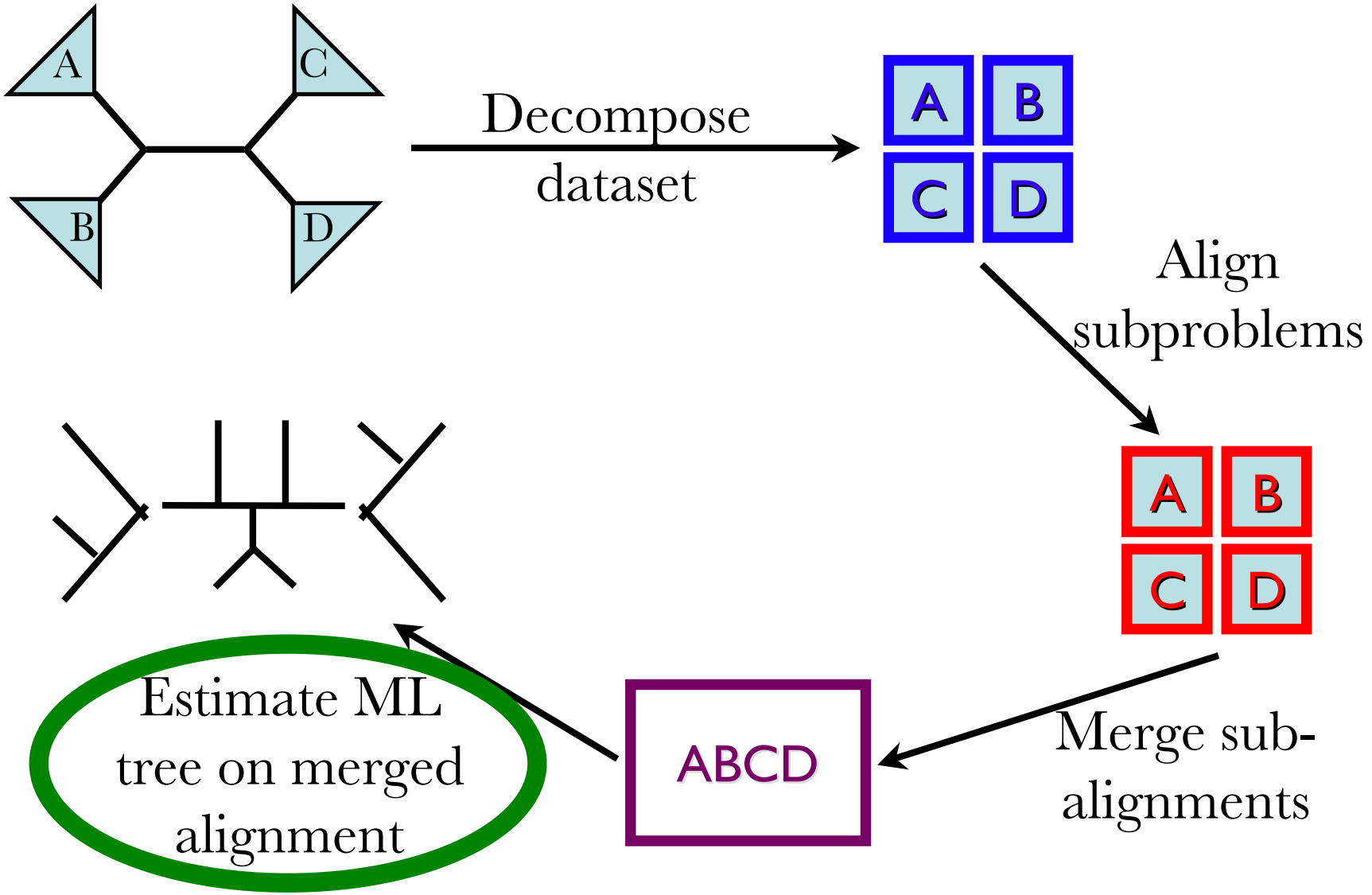
1000 taxon models, ordered by difficulty

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

Limitations

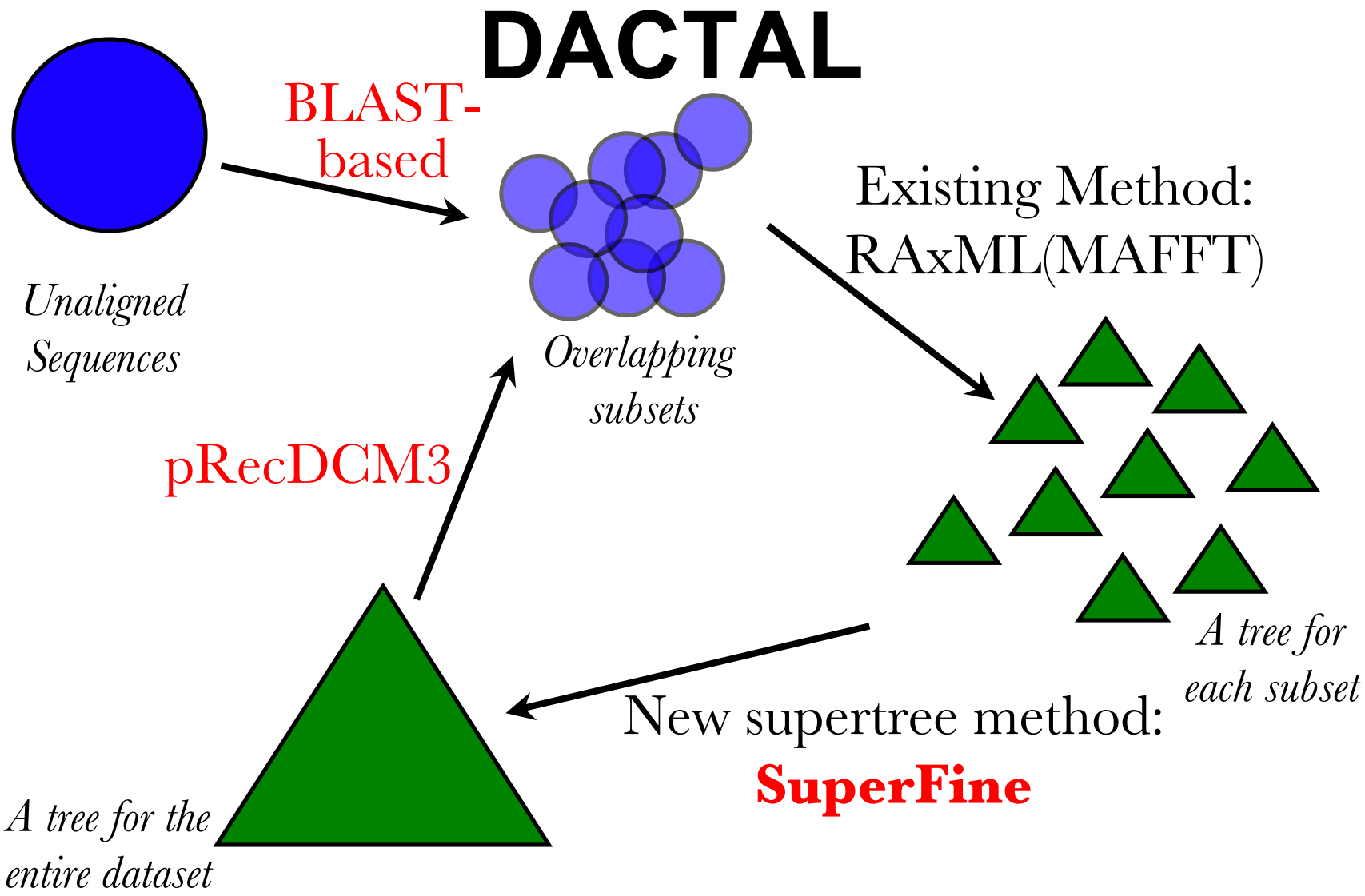


Part III: DACTAL

(Divide-And-Conquer Trees (Almost) without alignments)

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

Nelesen, Liu, Wang, Linder, and Warnow, ISMB 2012 and Bioinformatics 2012


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,

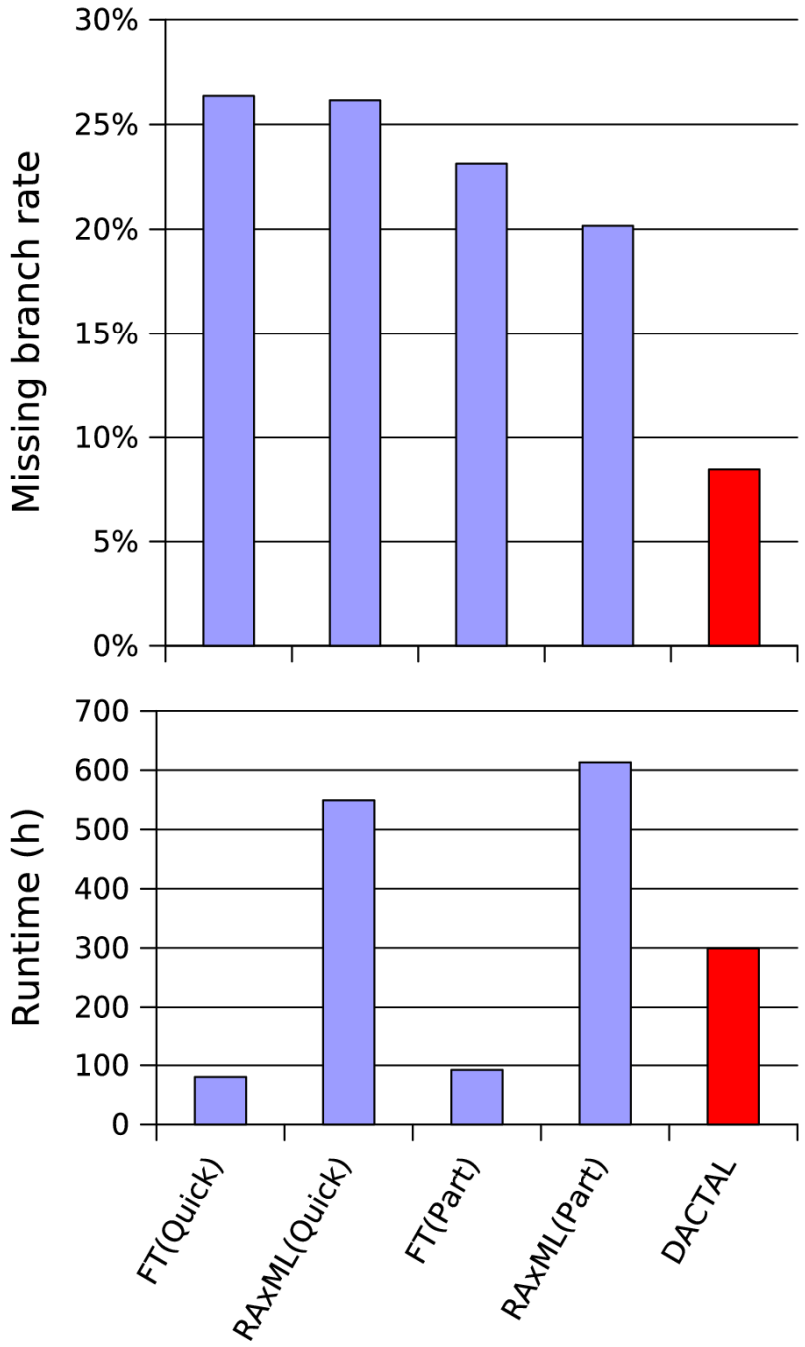
Three 16S datasets with **6,323** to **27,643** sequences

Reference alignments based on secondary structure

Reference trees are 75% RAxML bootstrap trees

DACTAL (shown in red) run for 5 iterations starting from FT(Part)

FastTree (FT) and RAxML are ML methods



Part III: SEPP

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

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

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.

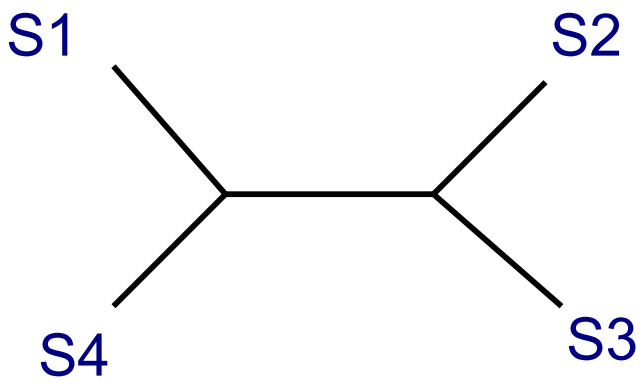
Phylogenetic Placement

- Align each query sequence to backbone alignment

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

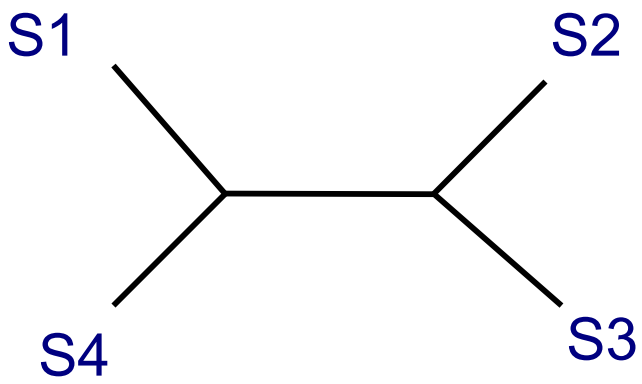
Align Sequence

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



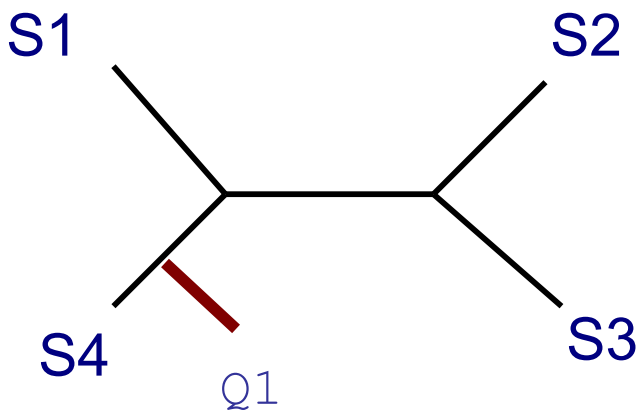
Align Sequence

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

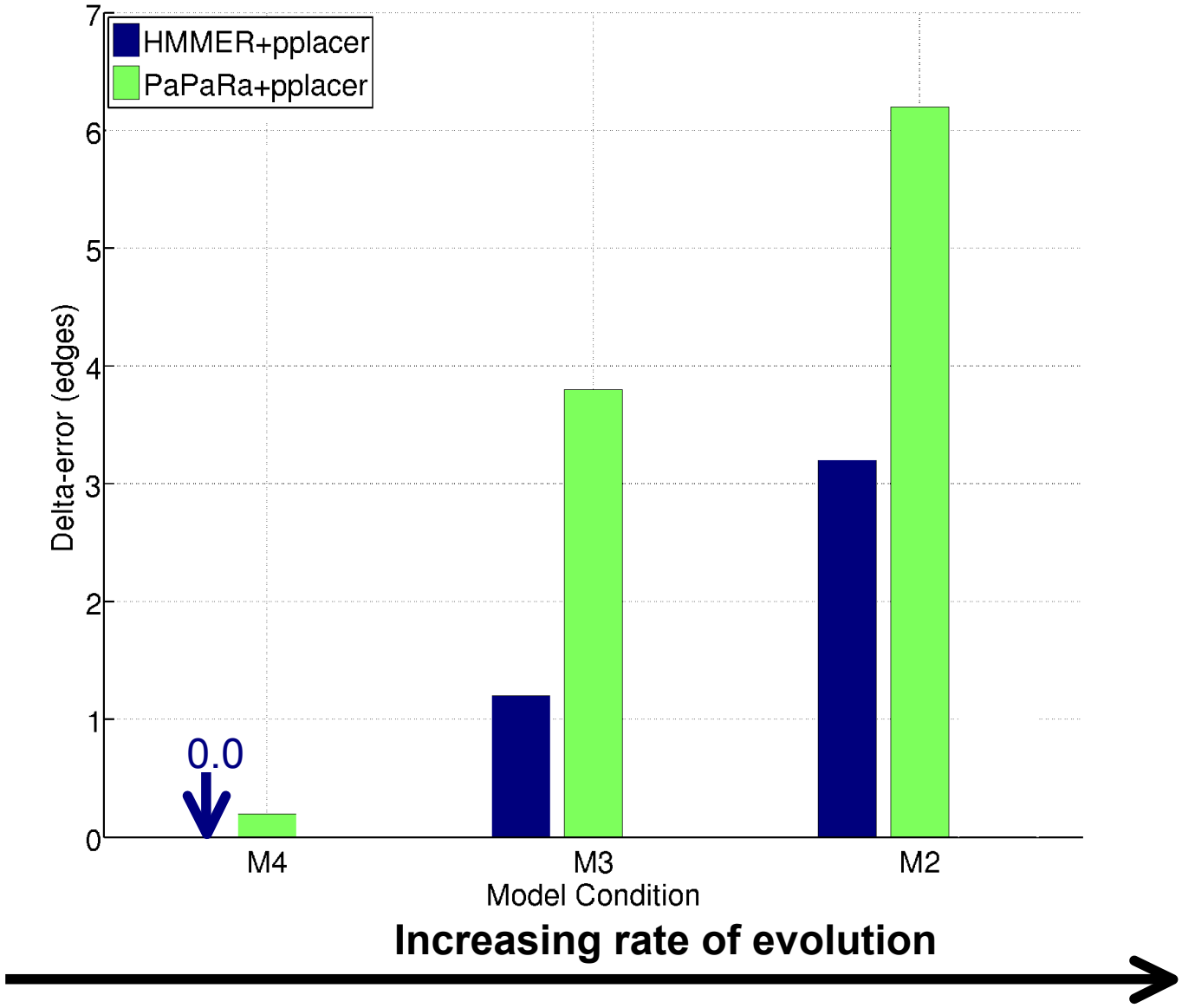


Phylogenetic Placement

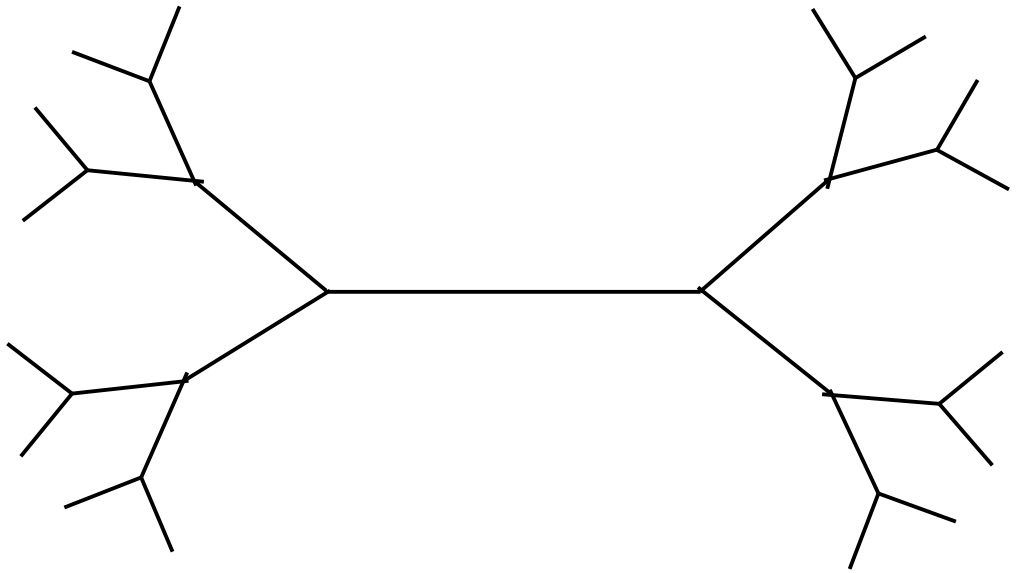
- Align each query sequence to backbone alignment
  - HMMALIGN (Eddy, Bioinformatics 1998)
  - PaPaRa (Berger and Stamatakis, Bioinformatics 2011)
- Place each query sequence into backbone tree
  - Pplacer (Matsen et al., BMC Bioinformatics, 2011)
  - EPA (Berger and Stamatakis, Systematic Biology 2011)

Note: pplacer and EPA use maximum likelihood

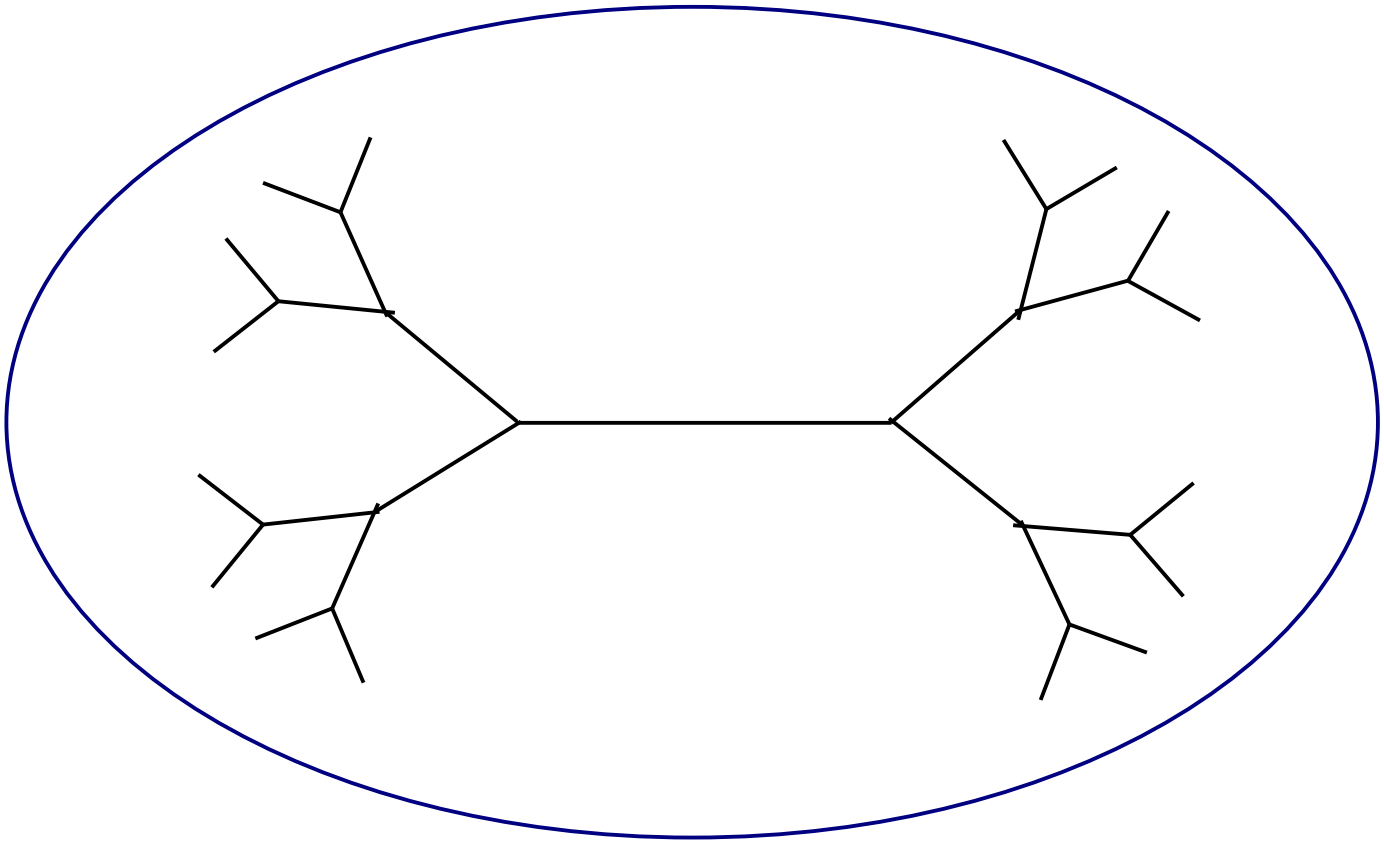
HMMER vs. PaPaRa



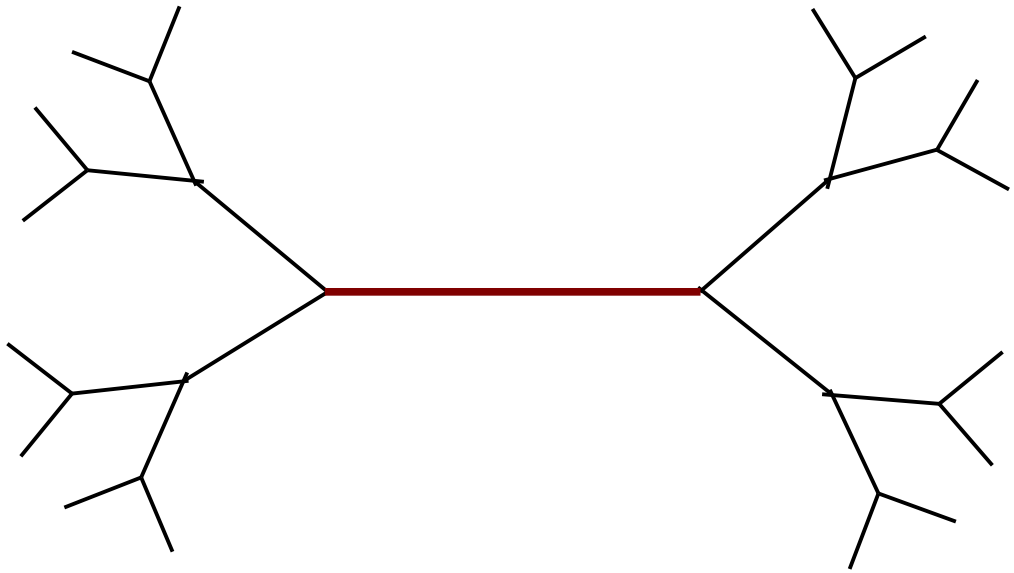
Insights from SATé



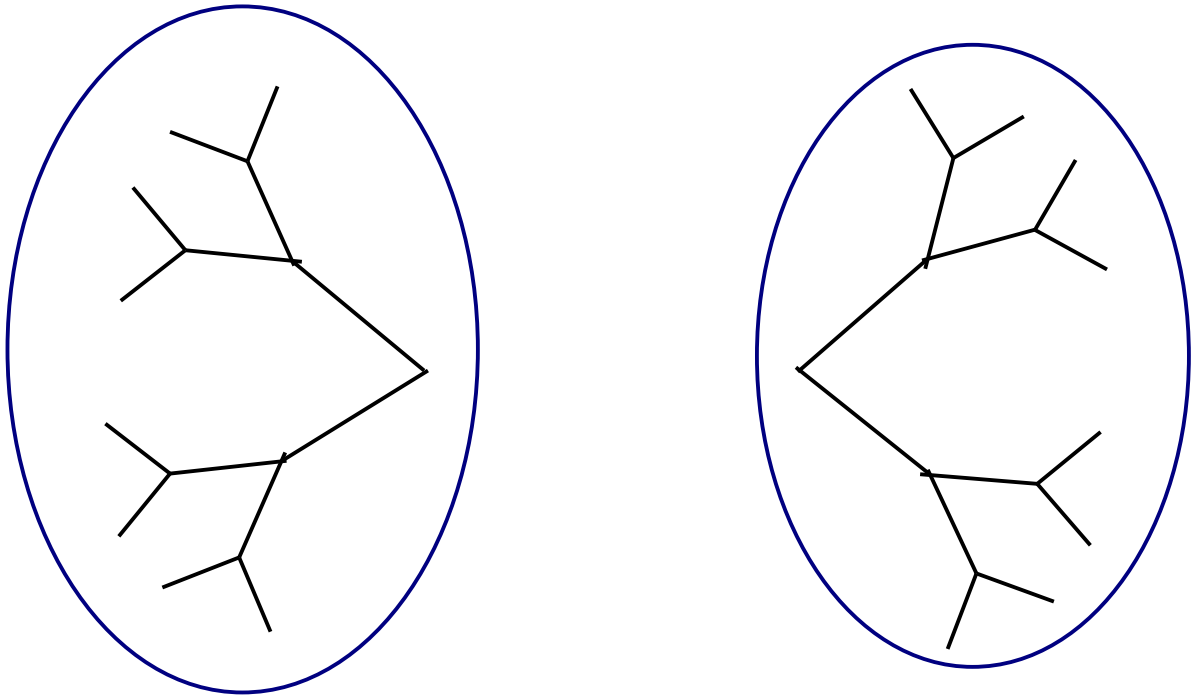
Insights from SATé



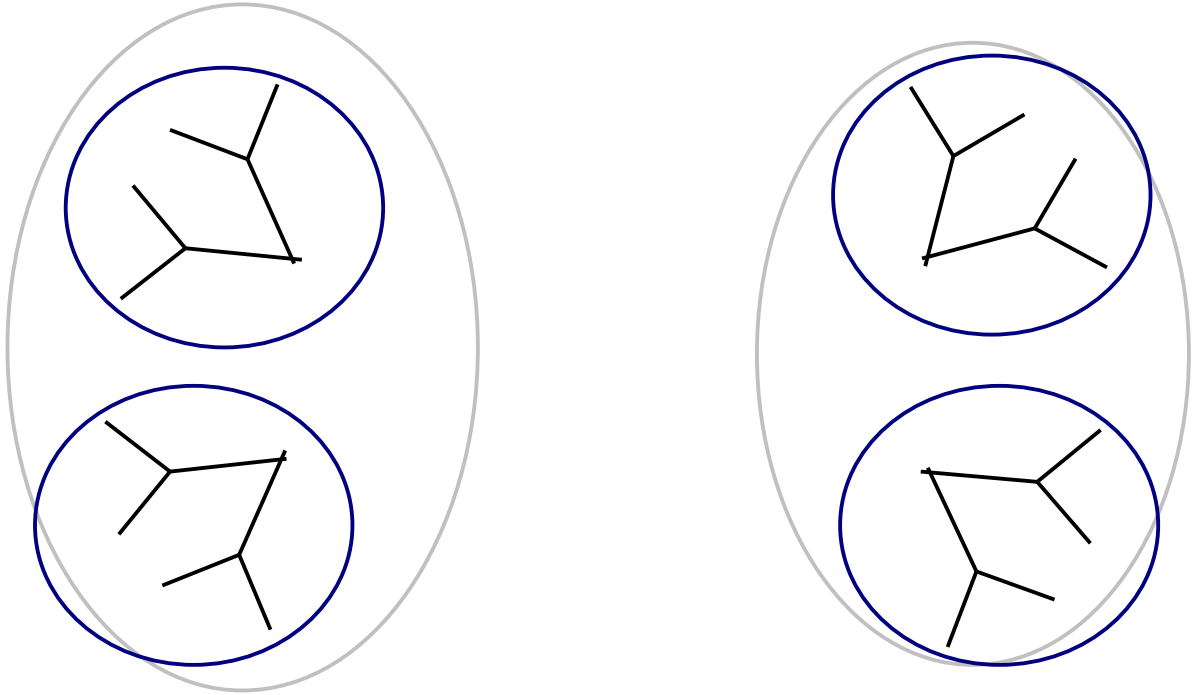
Insights from SATé



Insights from SATé



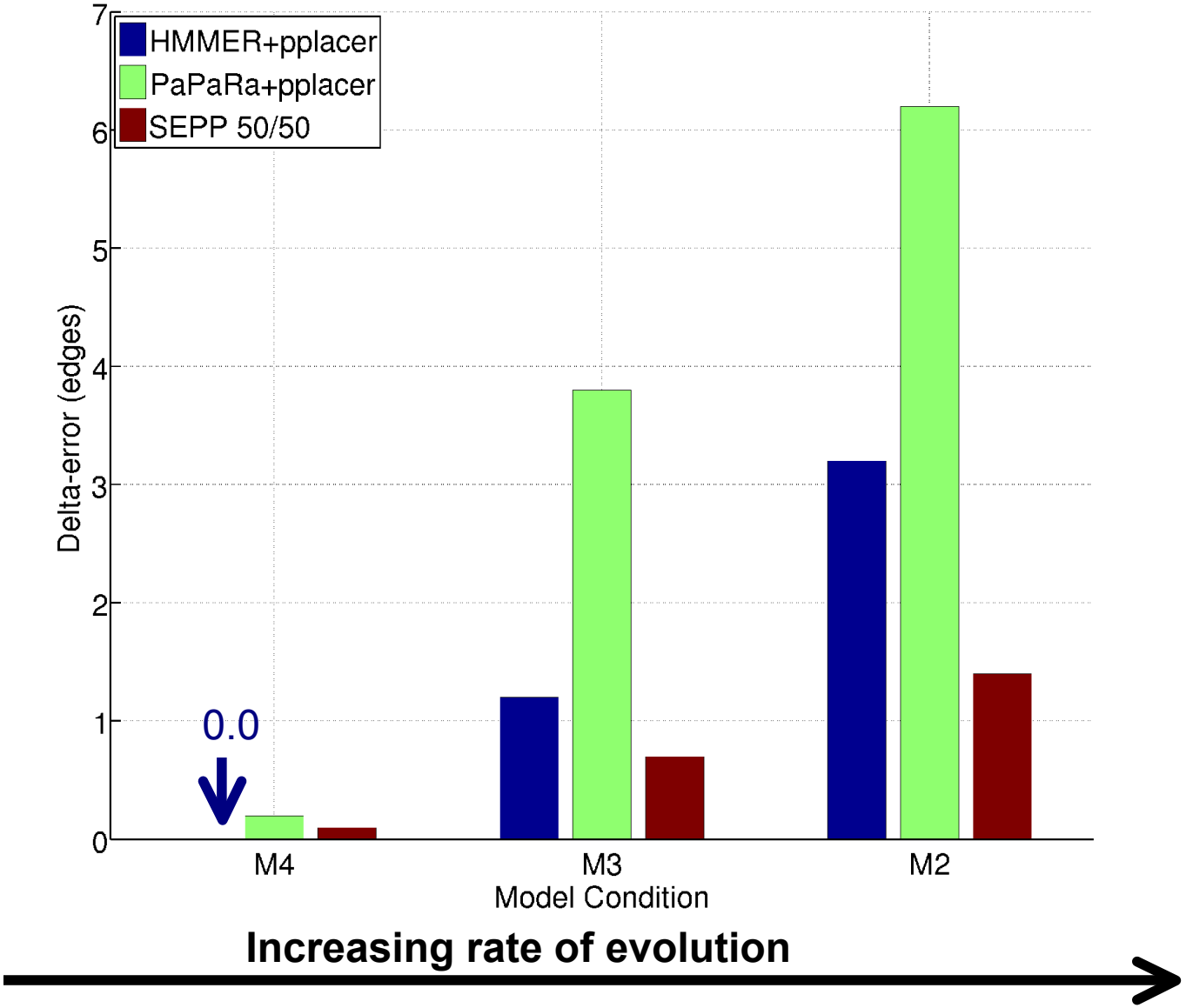
Insights from SATé



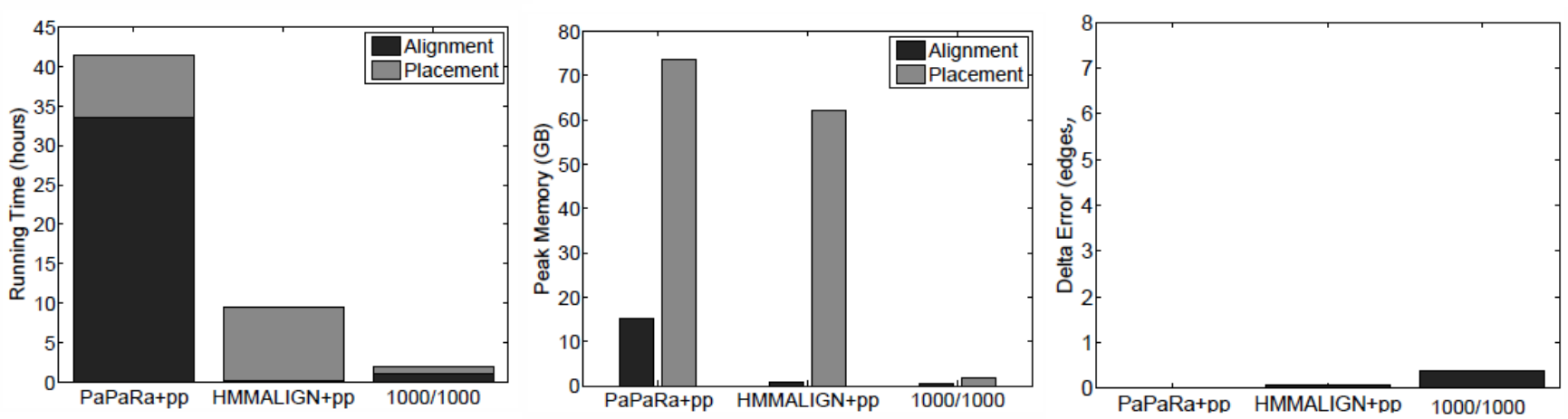
SEPP Parameter Exploration

- 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

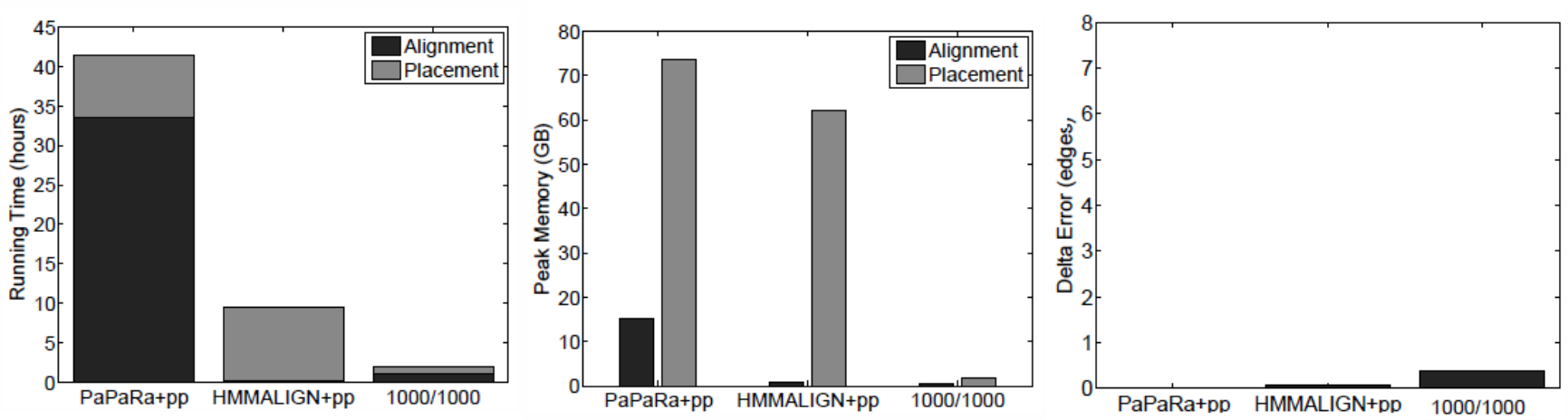


SEPP (10%) on Biological Data



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:

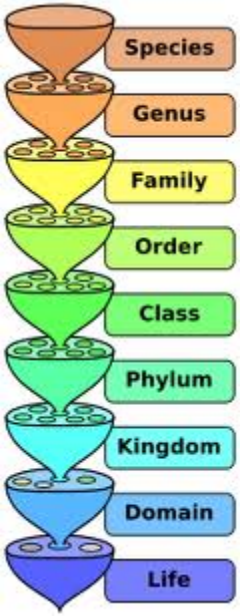
PaPaRa+pplacer: ~133 days

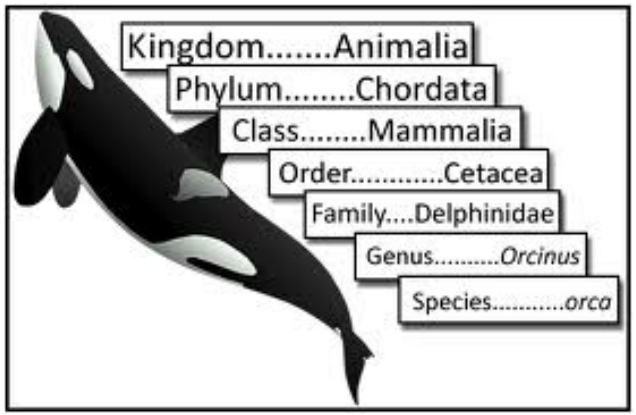
HMMALIGN+pplacer: ~30 days

SEPP 1000/1000: ~6 days

Part IV:
Taxon Identification

Objective: classify short reads in a metagenomic sample





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

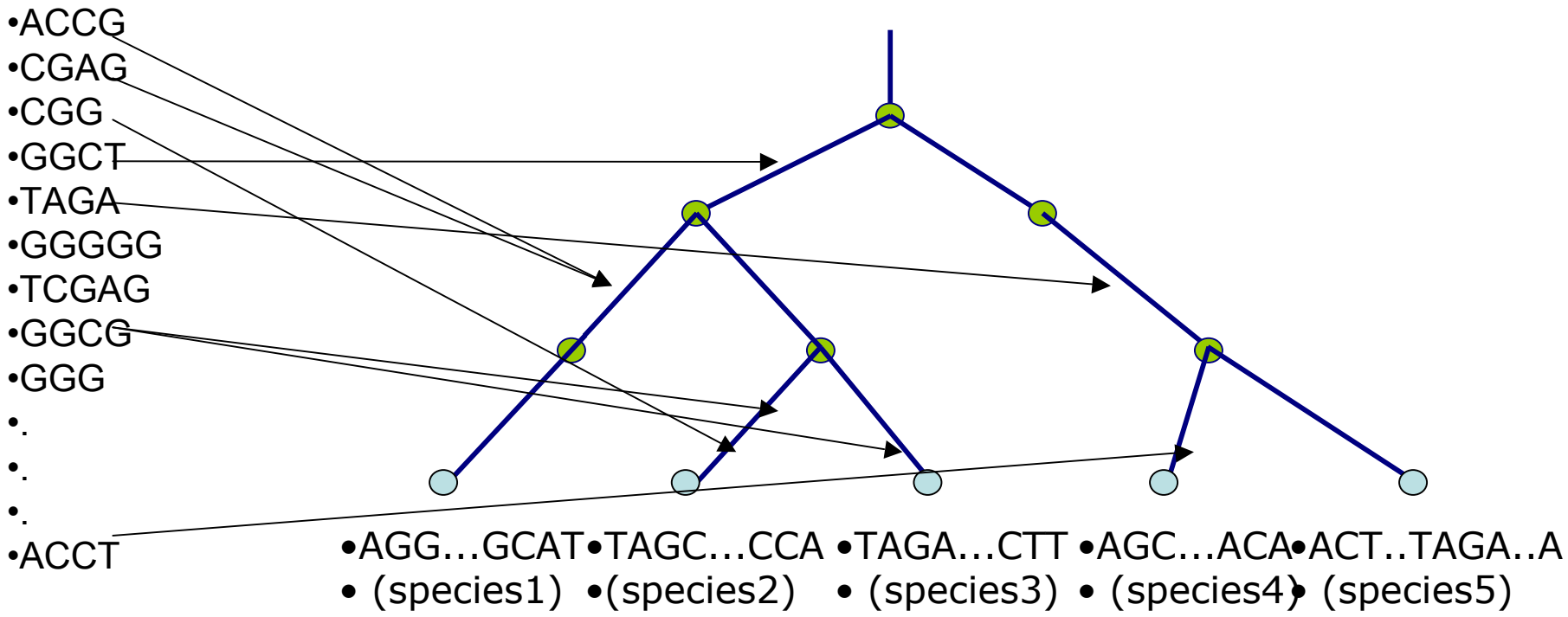
TIPP: Taxon Identification by Phylogenetic Placement

•Fragmentary Unknown Reads:

•(60-200 bp long)

•Known Full length Sequences,
•and an alignment and a tree

•(500-10,000 bp long)



TIPP: Taxon Identification using Phylogenetic Placement - Version 1

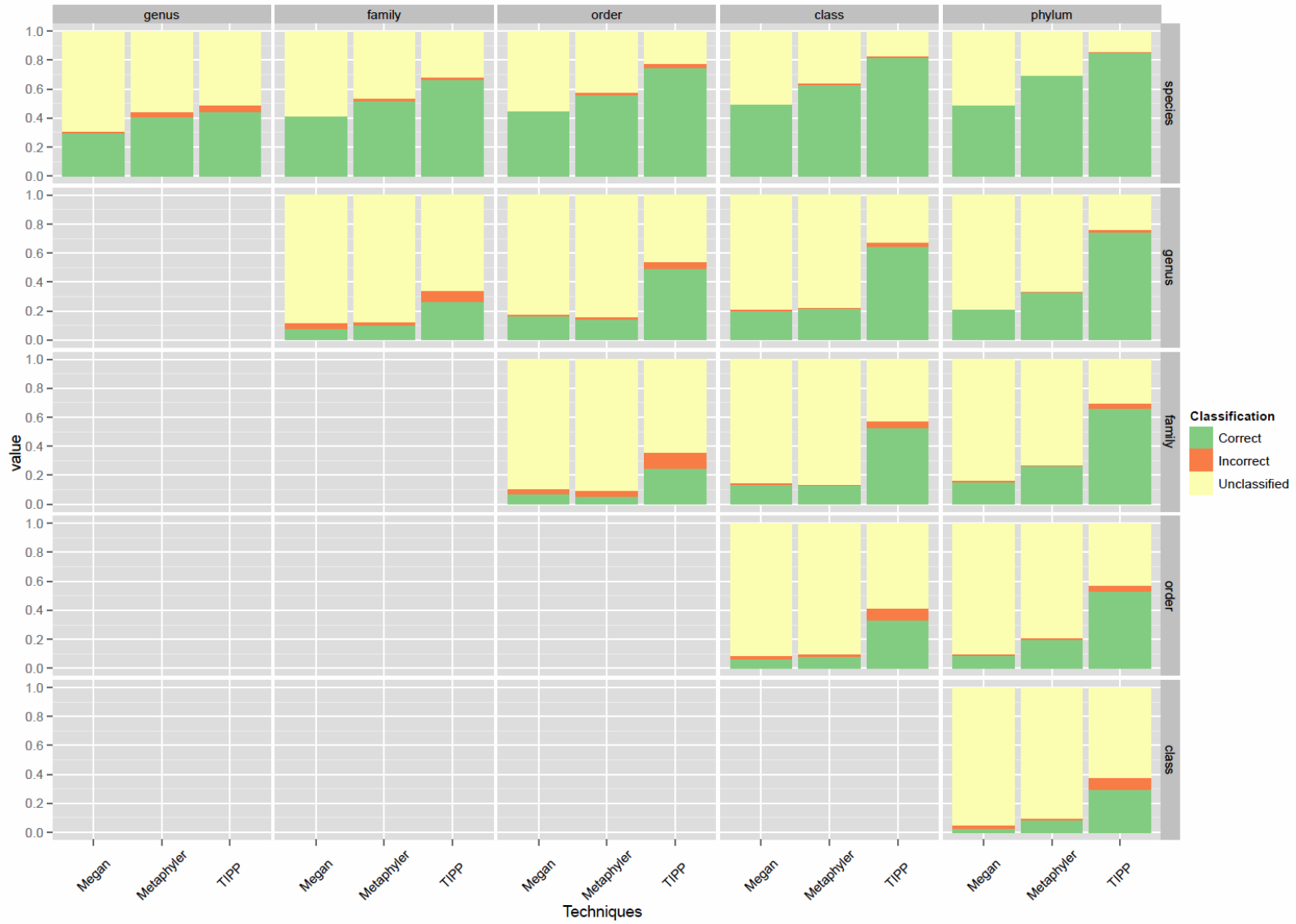
Given a set Q of query sequences for some gene, a taxonomy T, and a set of full-length sequences for the gene,

- Compute reference alignment and tree on the full-length sequences, using SATé
- Use SEPP to place each query sequence into the taxonomy (alignment subsets computed on the reference alignment/tree, then inserted into taxonomy T)

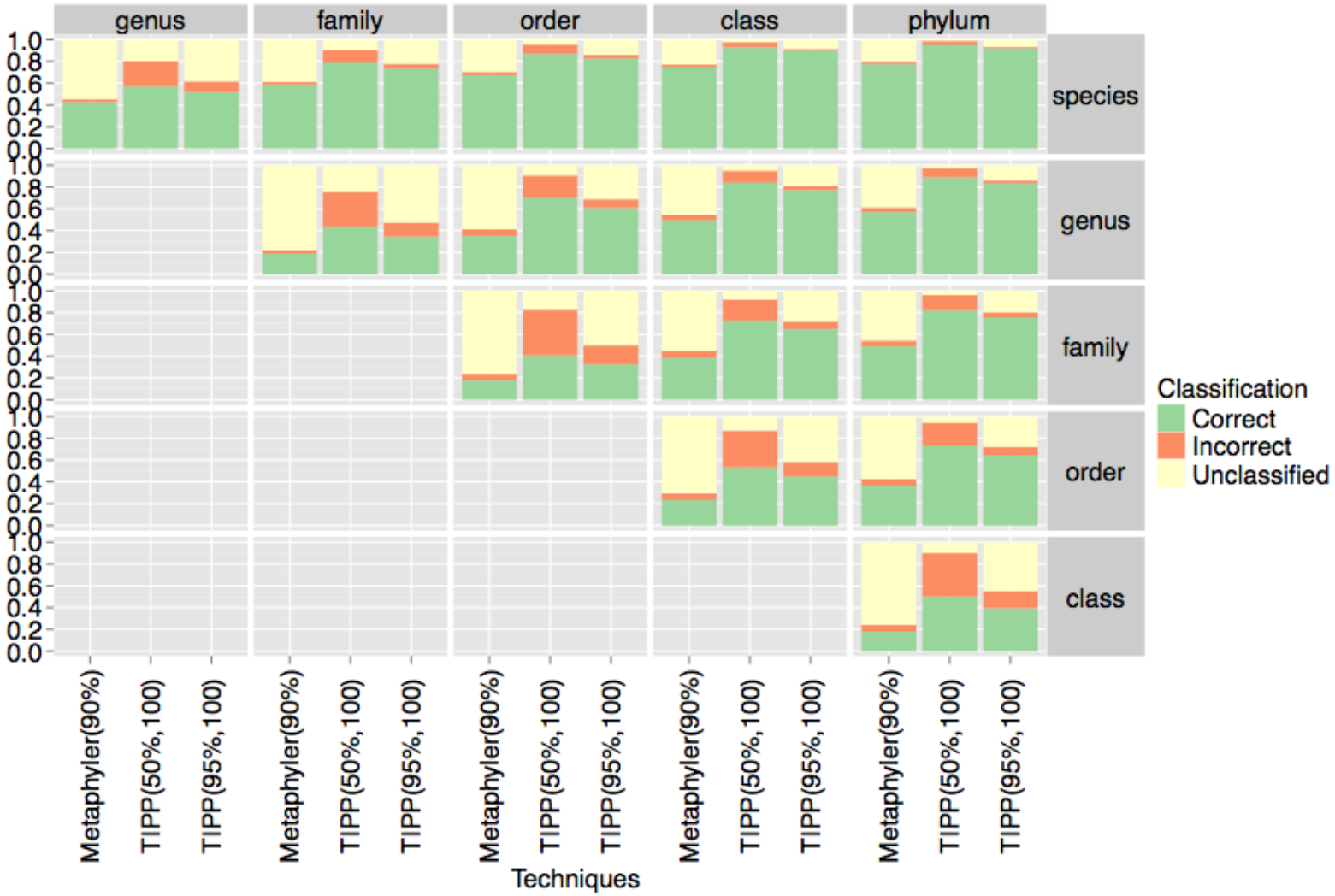
TIPP version 2- considering uncertainty

- TIPP version 1 too aggressive (over-classification)
- TIPP version 2 dramatically reduces false positive rate with small reduction in true positive rate, by considering uncertainty, using statistical techniques.

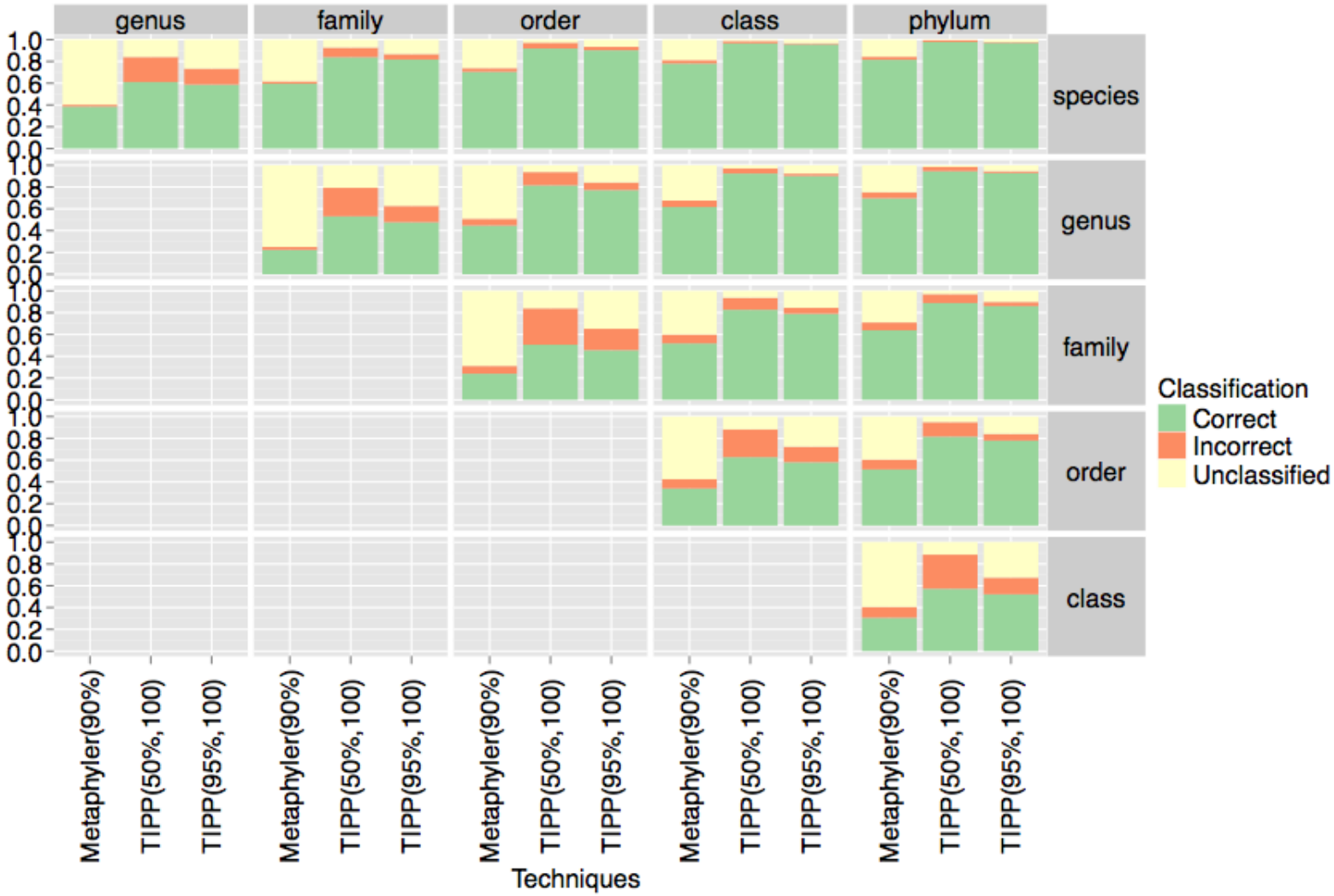
60bp error-free reads on rpsB marker gene



Results on 30 marker genes, leave-one-out experiment with Illumina errors



Results on 30 marker genes, leave-one-out experiment with 454 errors



Five “Boosters”

- DCM: distance-based tree estimation
- SATé: co-estimation of alignments and trees
- DACTAL: large trees without full alignments
- SEPP: phylogenetic placement of short reads
- TIPP: taxon identification of fragmentary data

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

General Observations - Part I

- Relative performance of methods can change dramatically with dataset size
- Statistical inference methods often do not scale well

Observations - Part II

- Meta-methods can improve accuracy and even speed
- Hidden Markov Models (HMMs) can be improved by making a set of HMMs instead of a single HMM
- Algorithmic parameters let you explore sensitivity/specificity
- Parallelism is easily exploited

Overall message

- When data are difficult to analyze, develop better methods - don’t throw out the data.
- BIGDATA problems in biology are an opportunity for computer scientists to have a big impact!

Discussion points

- Applicability to other machine learning problems? Classification and clustering problems, in particular?
- Space issues can arise if multiple solutions are maintained.
- Enabling plug-ins?
- How to enable parameter exploration? Statistically sound parameter selection?

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

- Collaborators:
  - DCM-NJ: Bernard Moret and Katherine St. John
  - SATé: Kevin Liu, Serita Nelesen, Sindhu Raghavan, and Randy Linder (and also Mark Holder at Kansas for public distribution)
  - DACTAL: Serita Nelesen, Kevin Liu, Li-San Wang, and Randy Linder
  - SEPP: Siavash Mirarab and Nam Nguyen
  - TIPP: Siavash Mirarab, Nam Nguyen, Mihai Pop, and Bo Liu