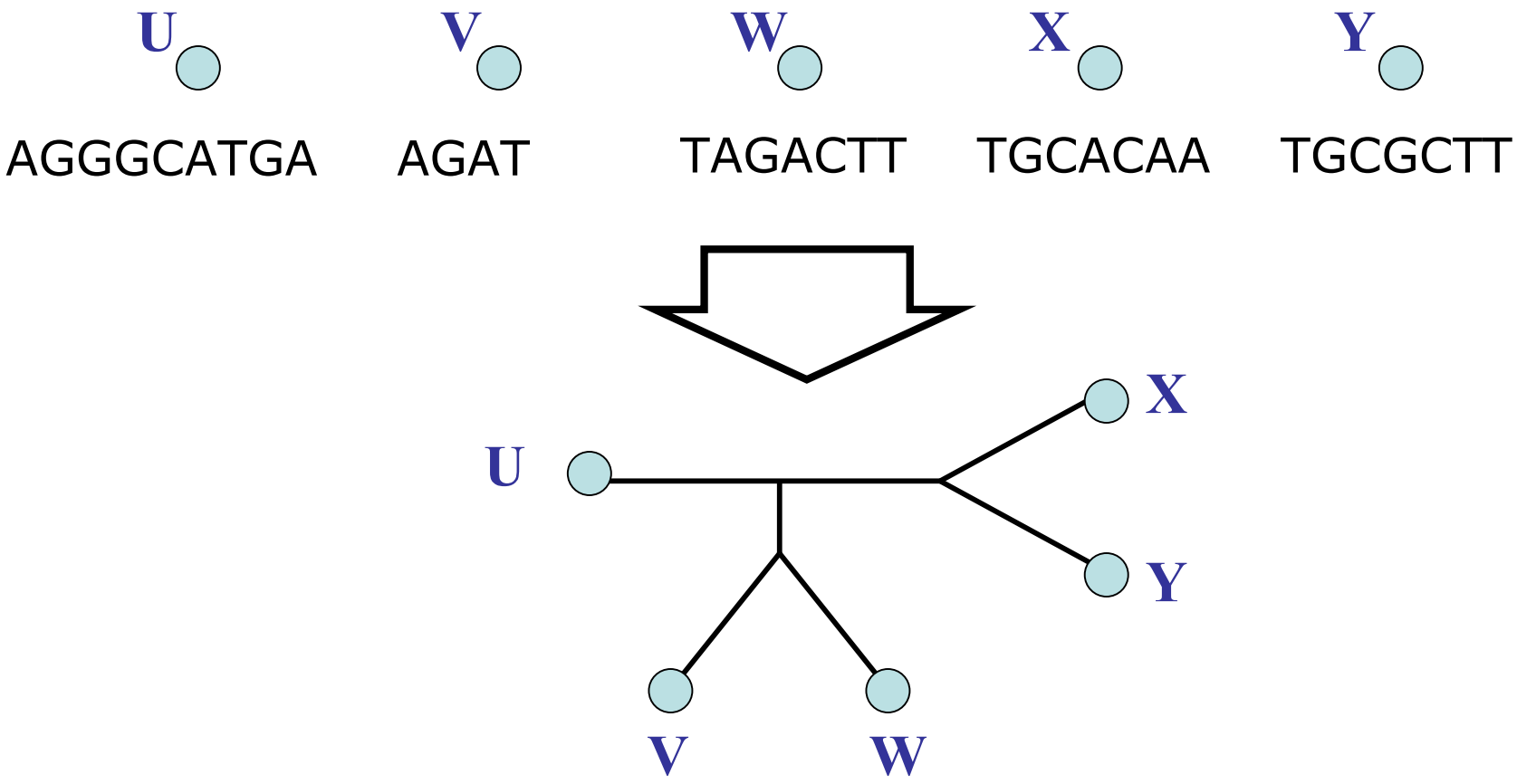
Three approaches to large-scale phylogeny estimation: SATé, DACTAL, and SEPP

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


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

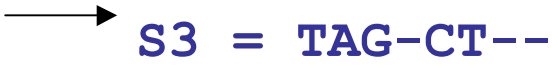
```
S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC
S3 = TAGCTGACCGC
S4 = TCACGACCGACA
```



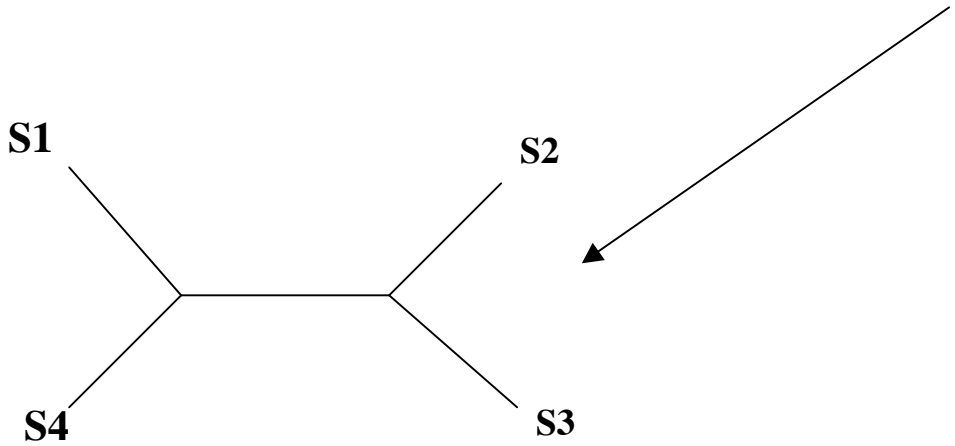
```
S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAC--GACCGC--
S3 = TAG-CT-------GACCGC--
S4 = -------TCAC--GACCGACA
```

Phase 2: Construct Tree

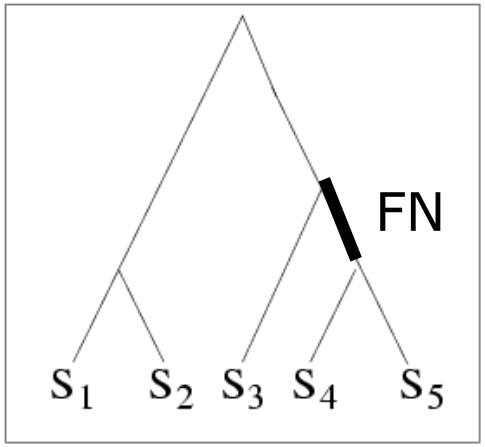
```
S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC
S3 = TAGCTGACCGC
S4 = TCACGACCGACA
```



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



Quantifying Error



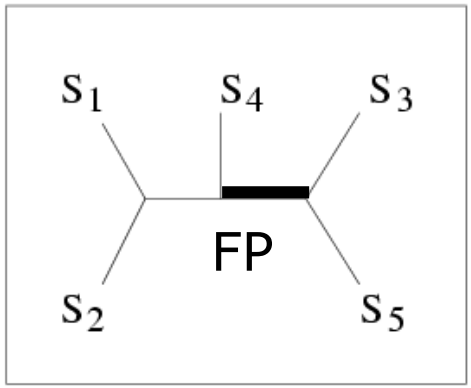
TRUE TREE

S_1	ACAATTAGAAC
S_2	ACCCTTAGAAC
S_3	ACCATTCCAAC
S_4	ACCAGACCAAC
S_5	ACCAGACCGGA

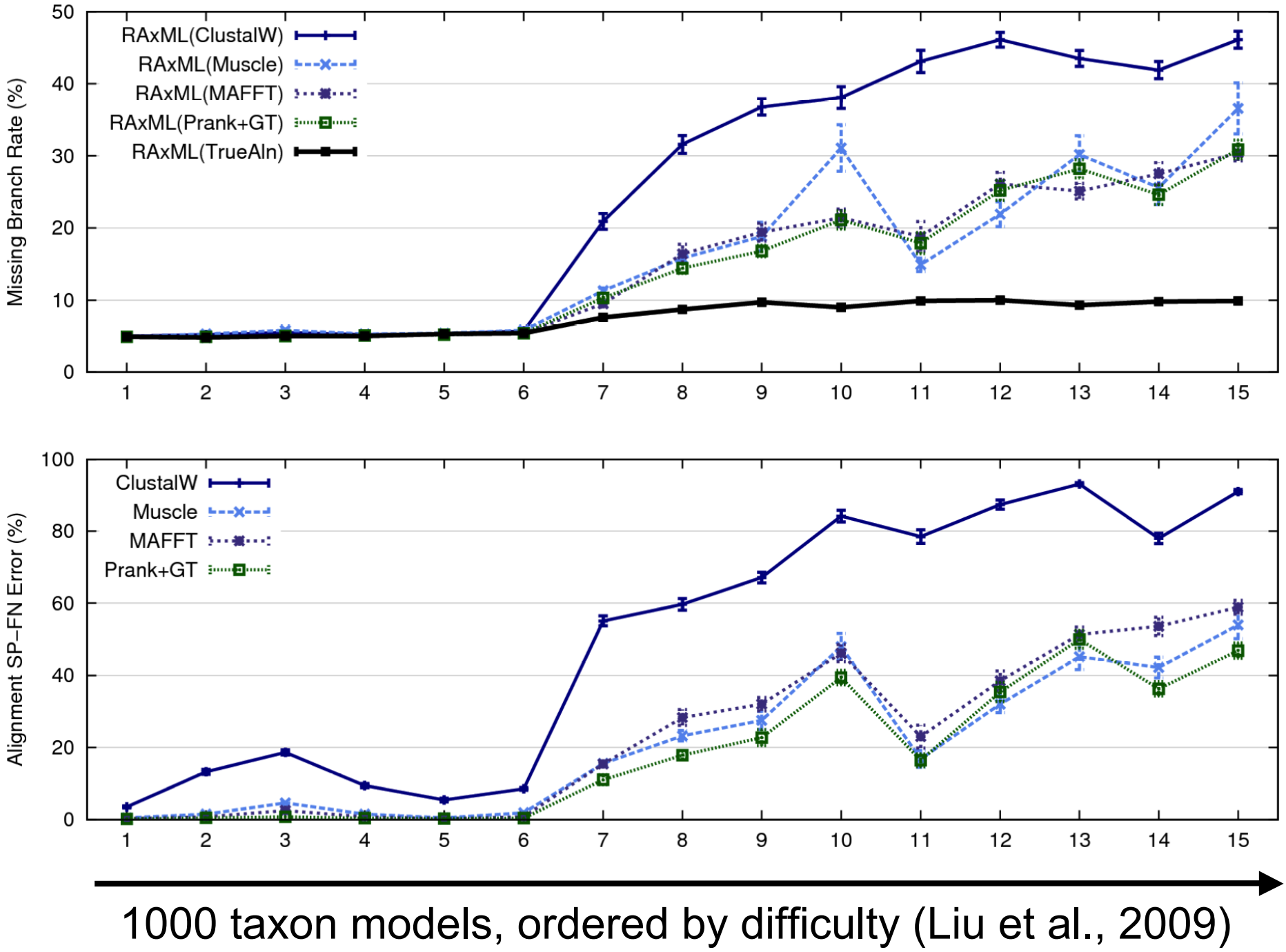
DNA SEQUENCES

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

50% error rate



INFERRED TREE


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)

Co-estimation methods

- POY, and other “treelength” methods, are controversial. Liu and Warnow, PLoS One 2012, showed that although gap penalty impacts tree accuracy, even when using a good affine gap penalty, treelength optimization gives poorer accuracy than maximum likelihood on good alignments.
- Likelihood-based methods based upon statistical models of evolution that include indels as well as substitutions (BAliPhy, Alifritz, StatAlign, and others) provide potential improvements in accuracy. These target small datasets (at most a few hundred sequences). BAli-Phy is the fastest of these methods.

This talk

SATé: Simultaneous Alignment and Tree Estimation

DACTAL: Divide-and-conquer trees (almost) without alignments

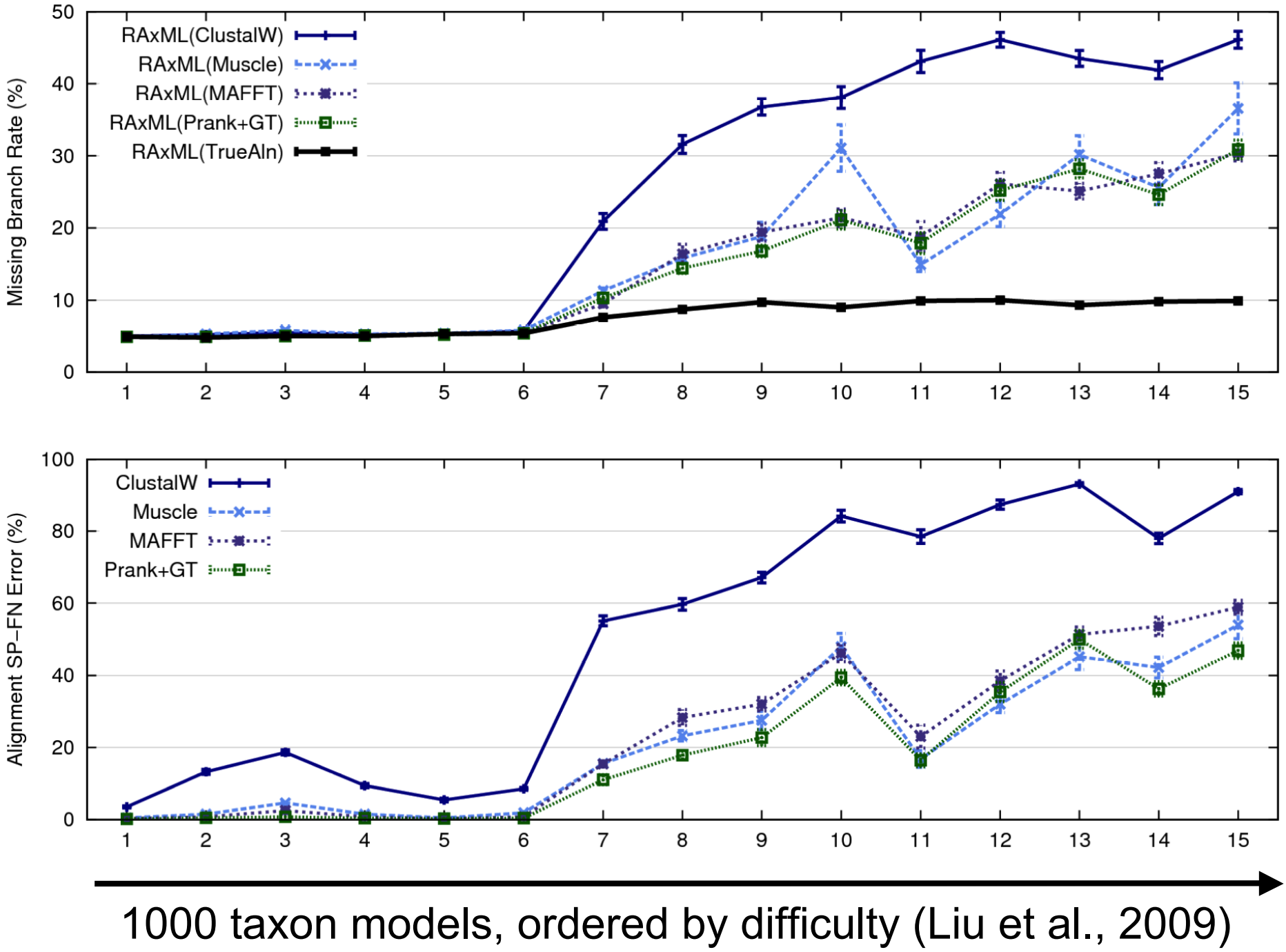
SEPP: SATé-enabled phylogenetic placement (analyses of large numbers of fragmentary sequences)

Part I: SATé

Simultaneous Alignment and Tree Estimation
(for nucleotide or amino-acid analysis)

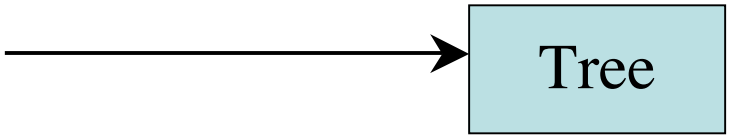
Liu, Nelesen, Raghavan, Linder, and Warnow, *Science*,
Liu et al., *Systematic Biology*, 2012

Public software distribution (open source) through the University of Kansas (Mark Holder)


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)

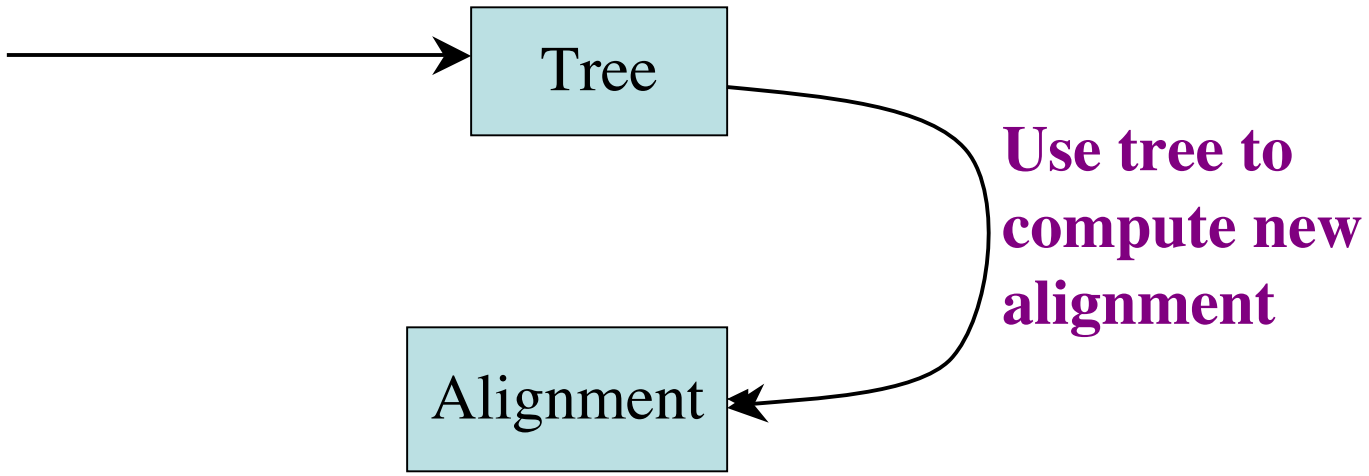
SATé Algorithm

Obtain initial alignment
and estimated ML tree



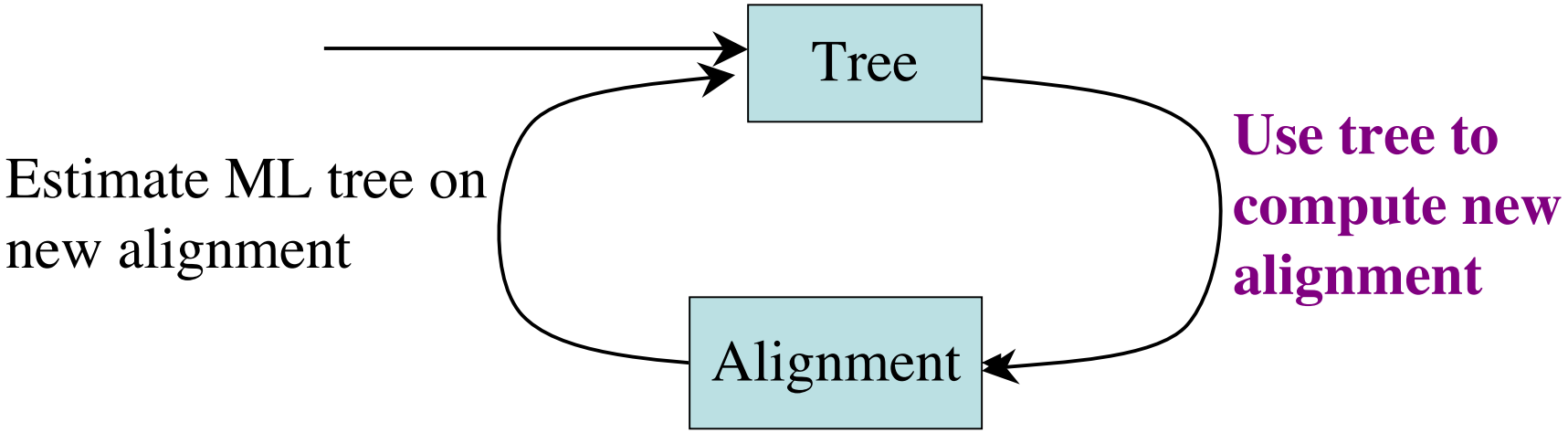
SATé Algorithm

Obtain initial alignment
and estimated ML tree

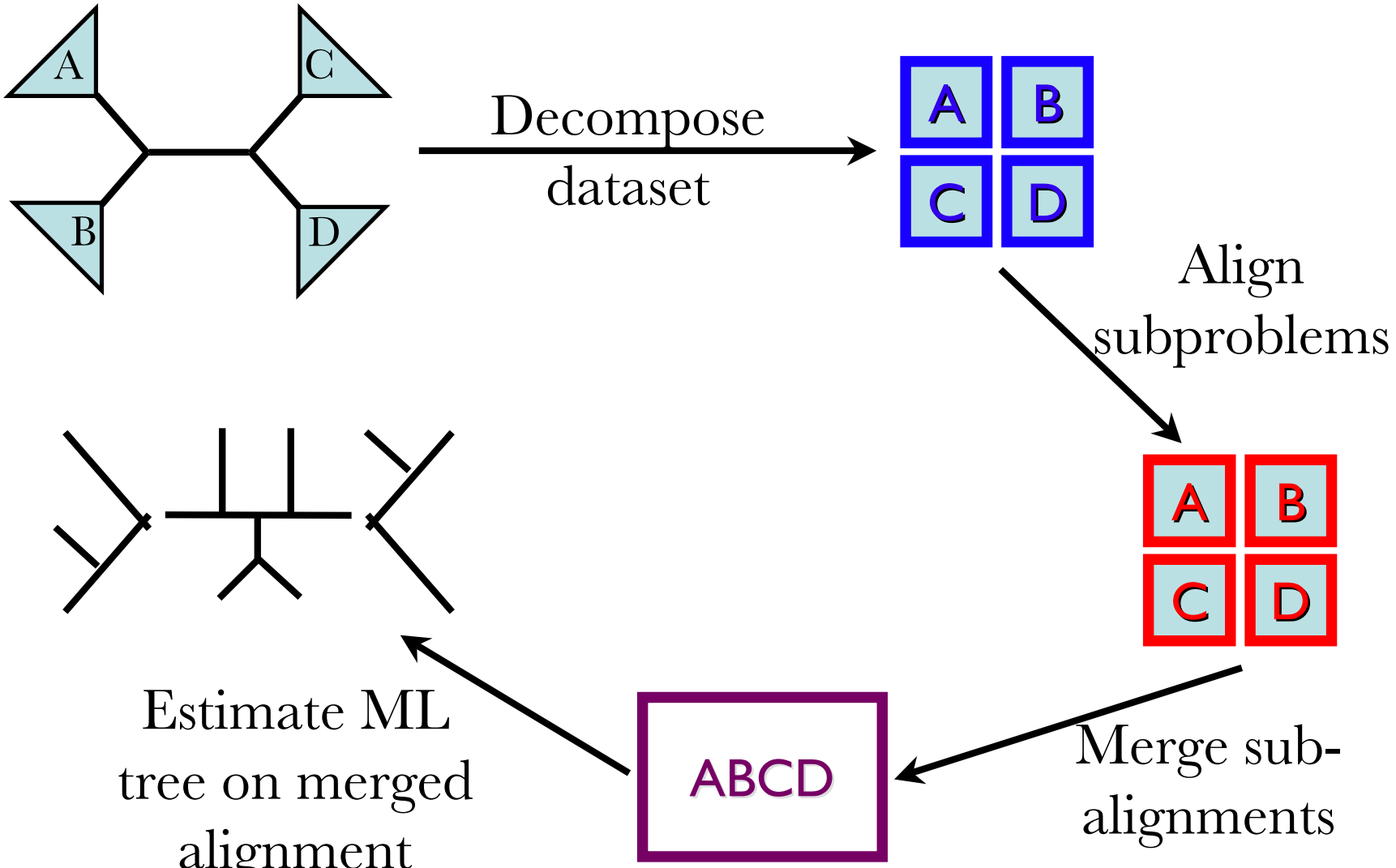


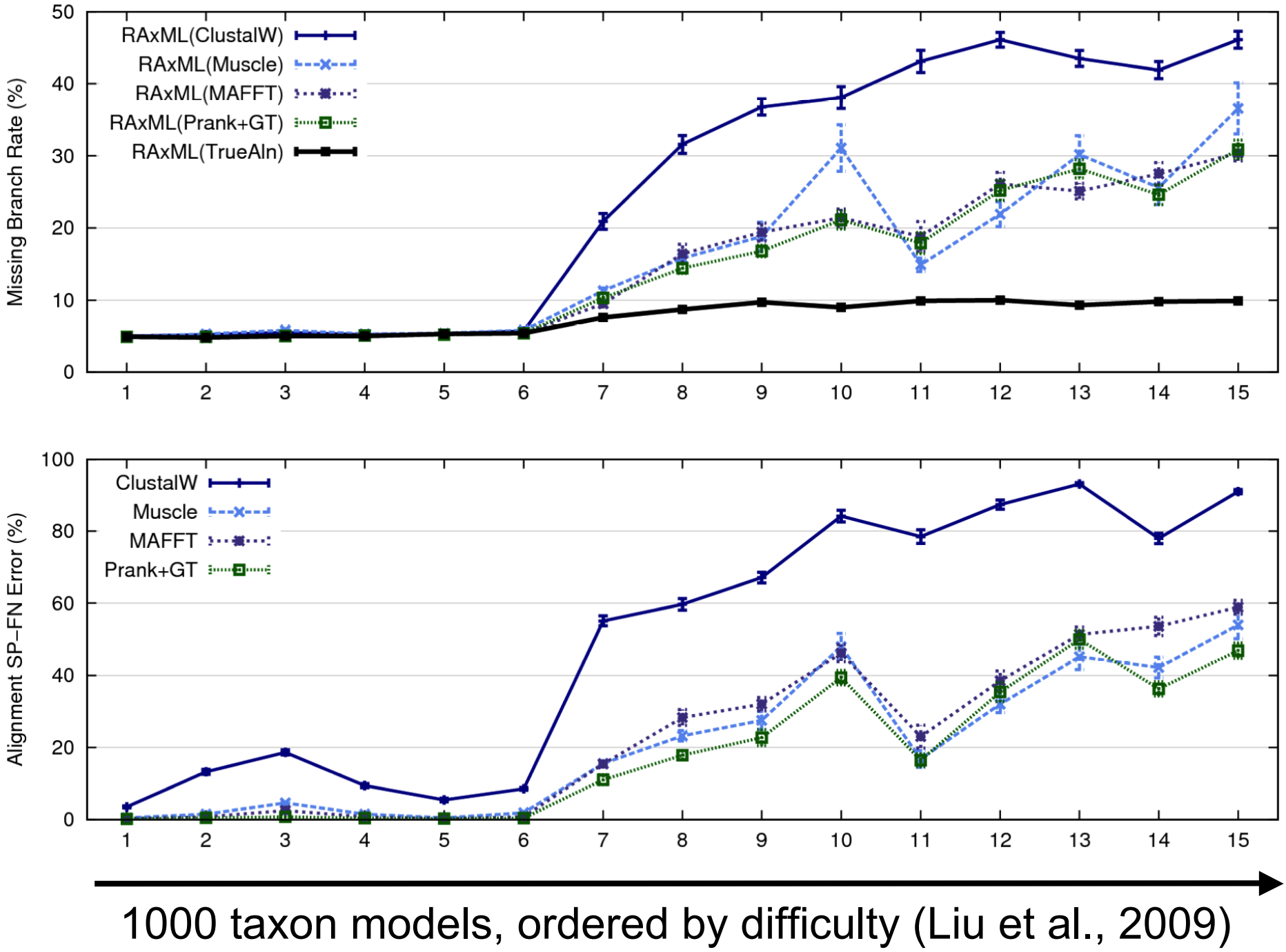
SATé Algorithm

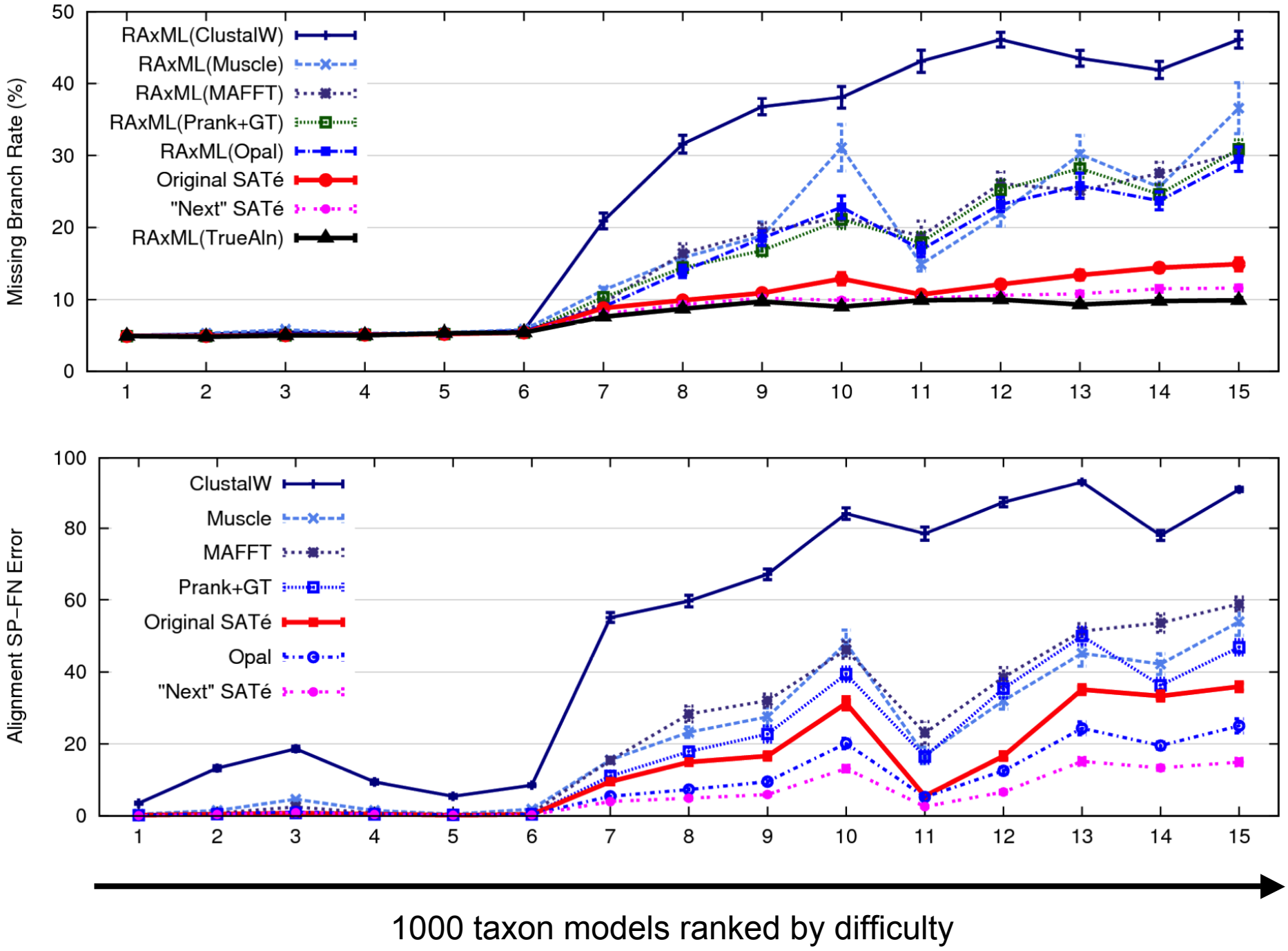
Obtain initial alignment
and estimated ML tree



Re-aligning on a Tree




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)


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

SATé Summary

Improved tree and alignment accuracy compared to two-phase methods, on both simulated and biological data.

Public software distribution (open source) through the University of Kansas (Mark Holder)

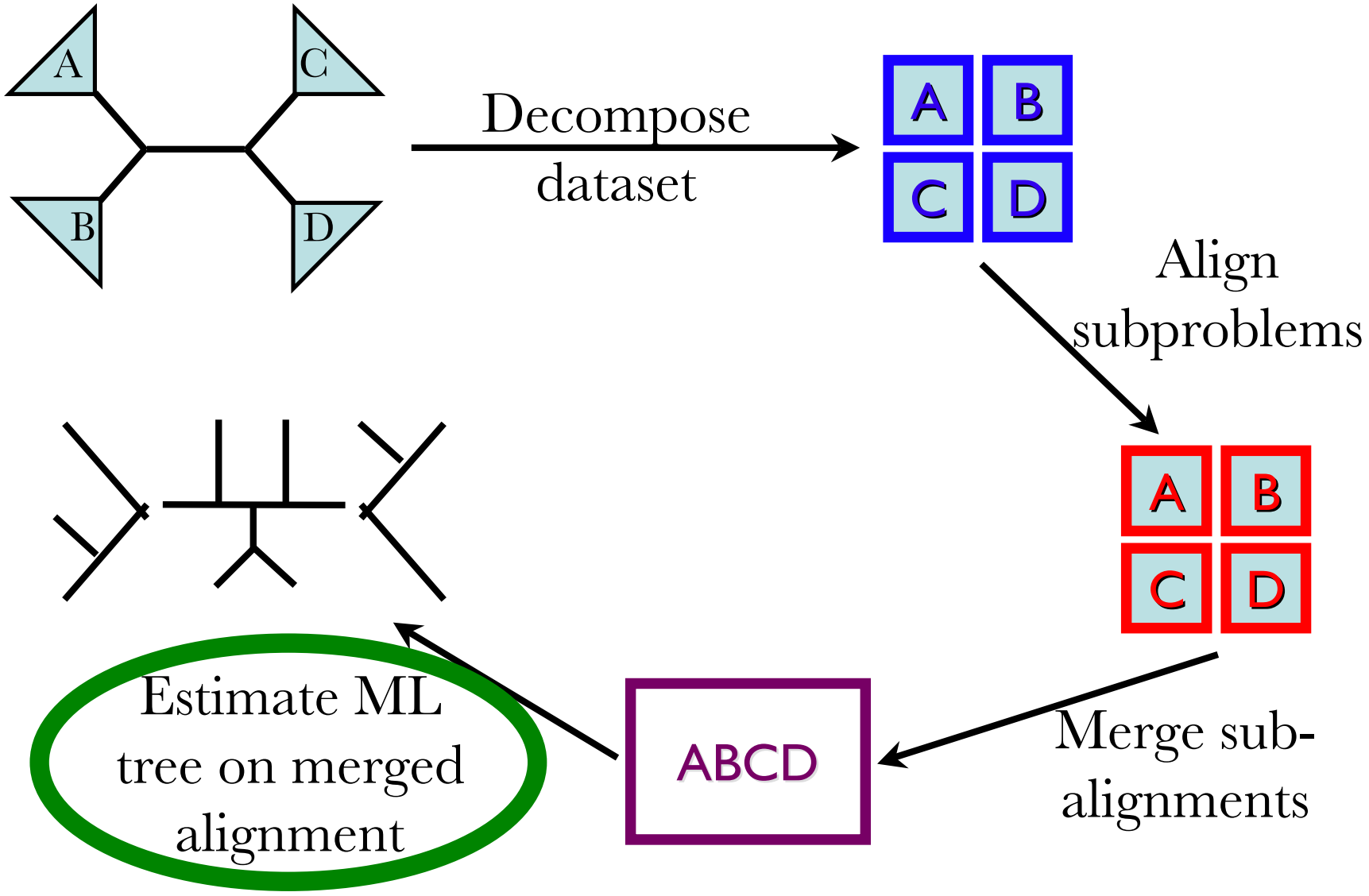
Workshops Monday and Tuesday

References:

Liu, Nelesen, Raghavan, Linder, and Warnow, *Science*,

Liu et al., *Systematic Biology*, 2012

Limitations



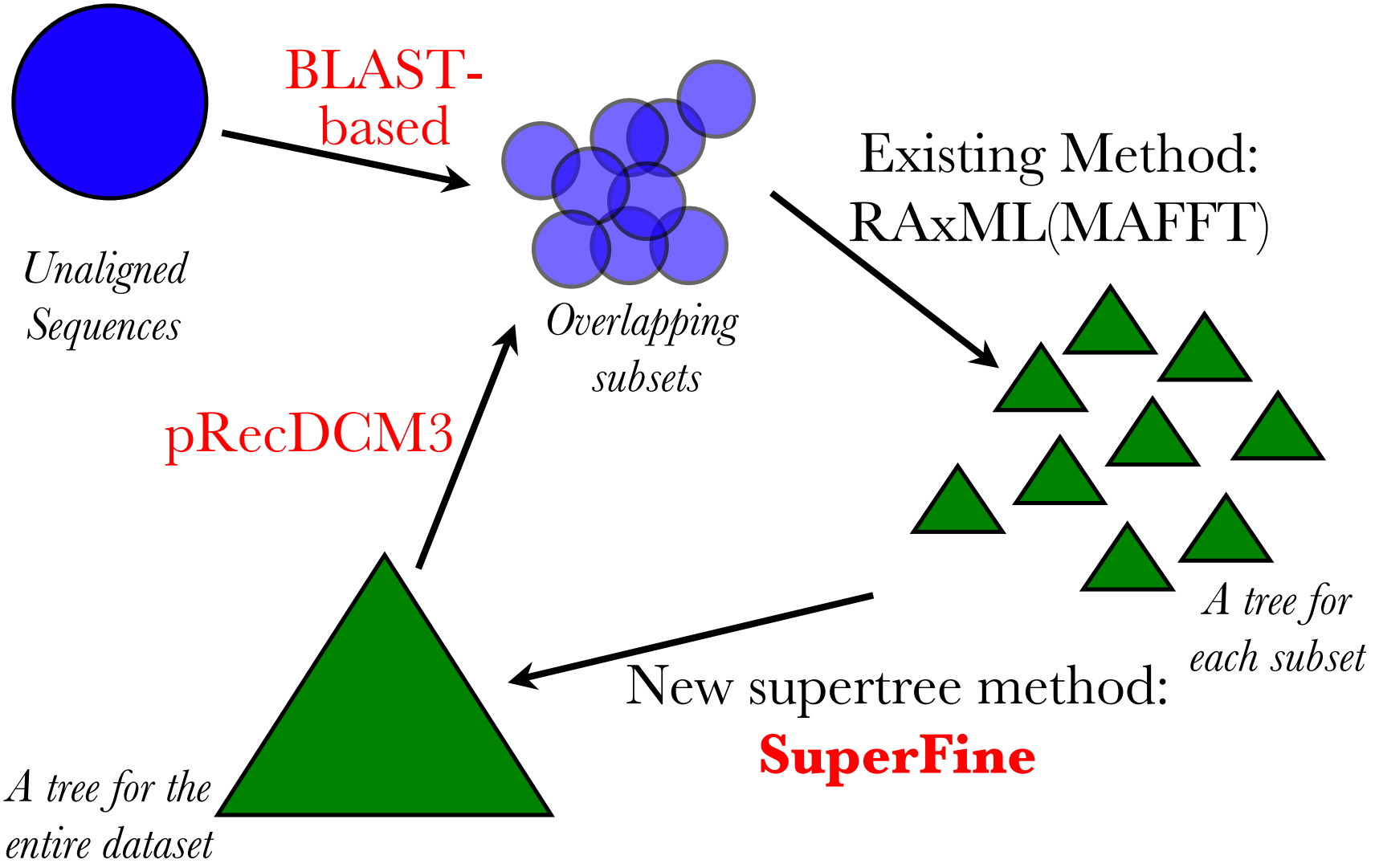
Part II: DACTAL

(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, In Press, ISMB 2012 and Bioinformatics 2012

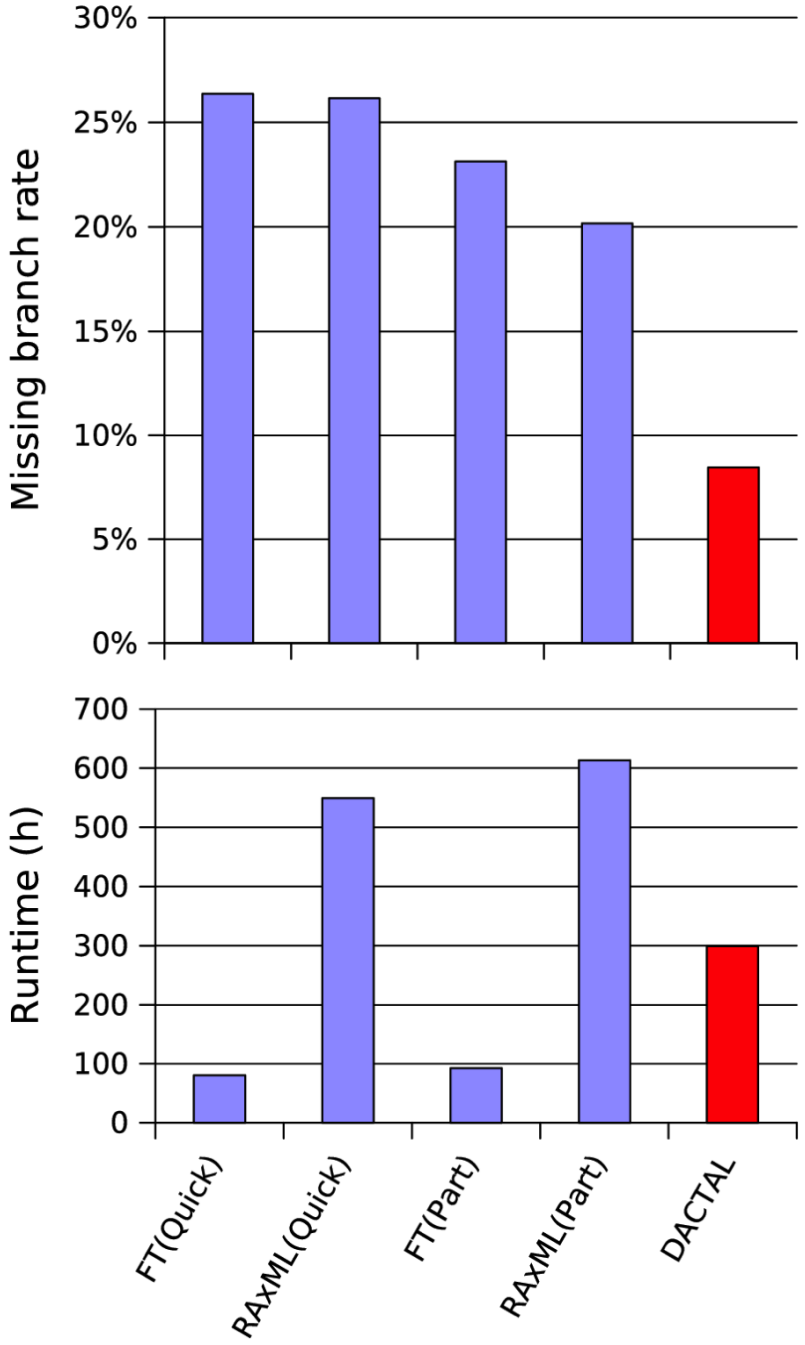
DACTAL


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

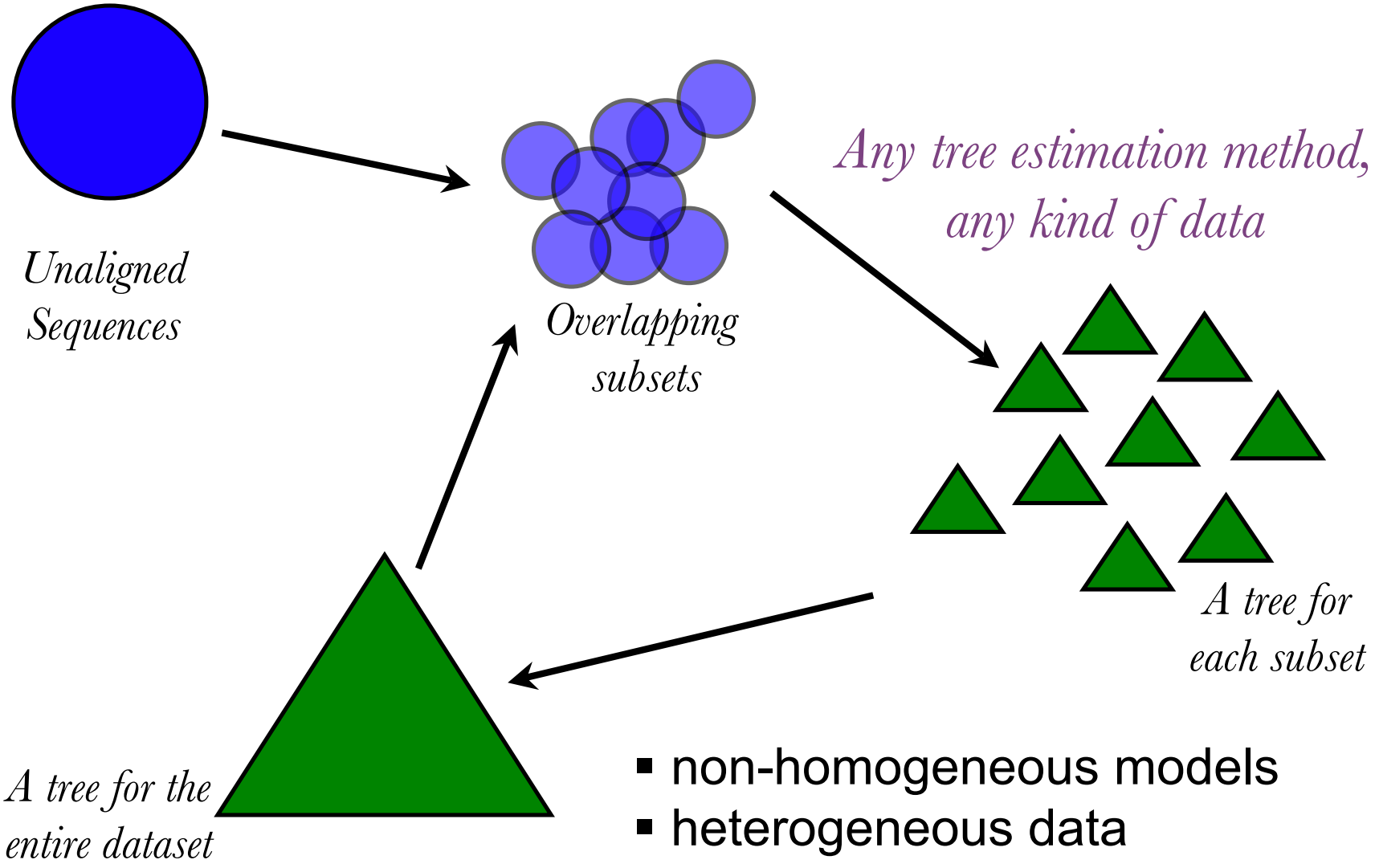
DACTAL: Better results than 2-phase methods

- Three 16S datasets from Gutell's database (CRW) 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



DACTAL is Flexible



- non-homogeneous models
- heterogeneous data

Part III: SEPP

- SEPP: SATé-enabled phylogenetic placement

- Mirarab, Nguyen, and Warnow. Pacific Symposium on Biocomputing, 2012.

NGS and metagenomic data

- Fragmentary data (e.g., short reads):
  - How to align? How to insert into trees?
- Unknown taxa
  - How to identify the species, genus, family, etc?

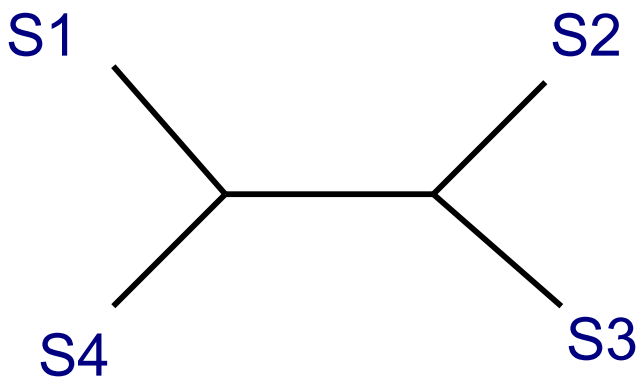
Phylogenetic Placement

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

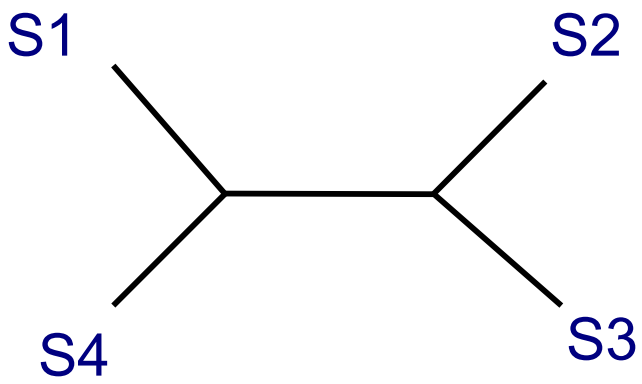
Align Sequence

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



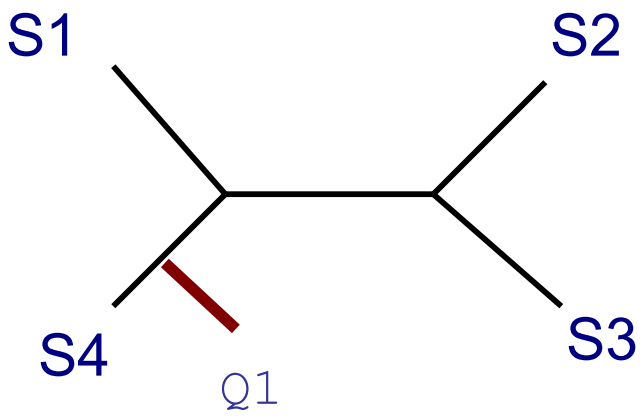
Align Sequence

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

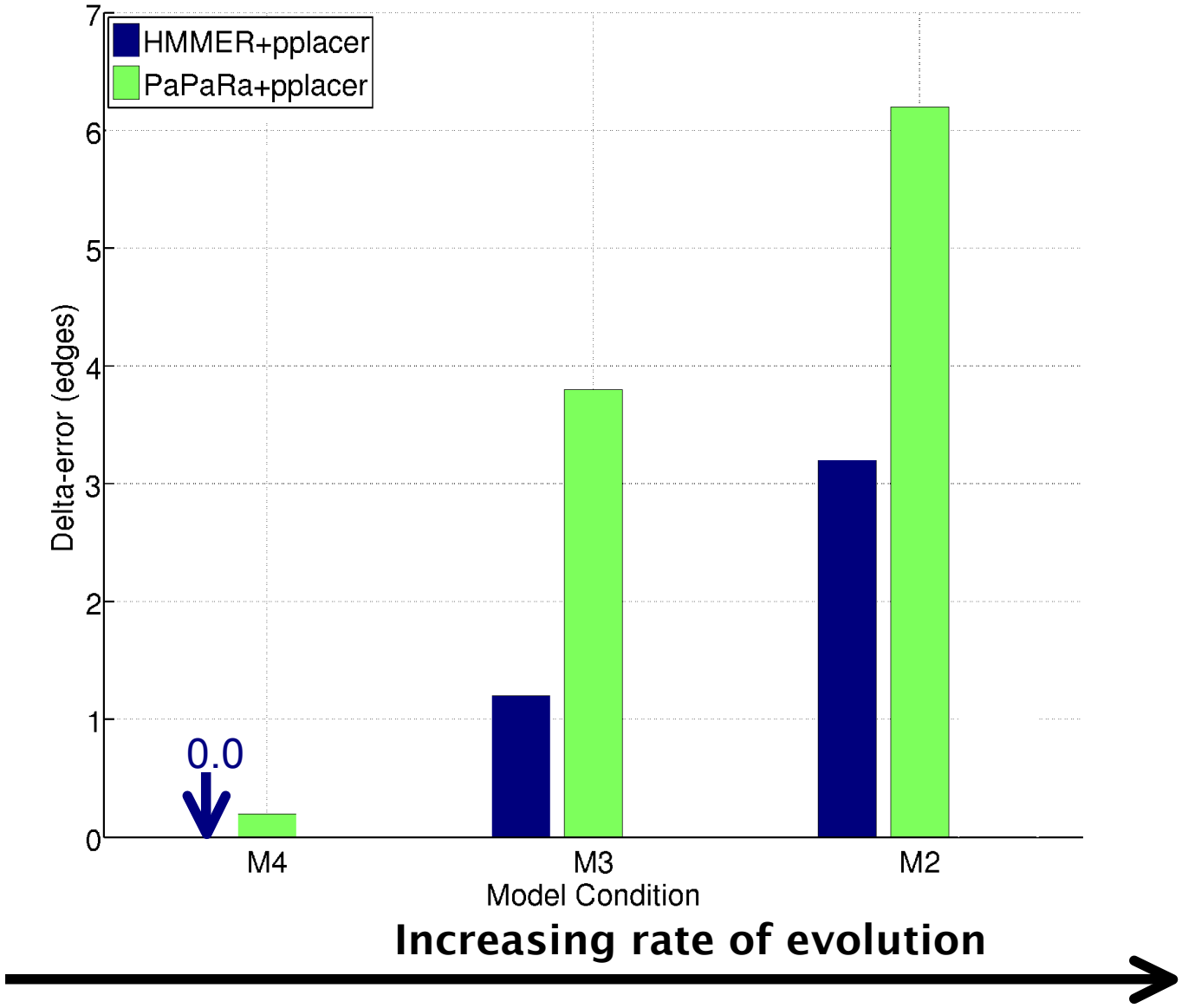


Phylogenetic Placement

- 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

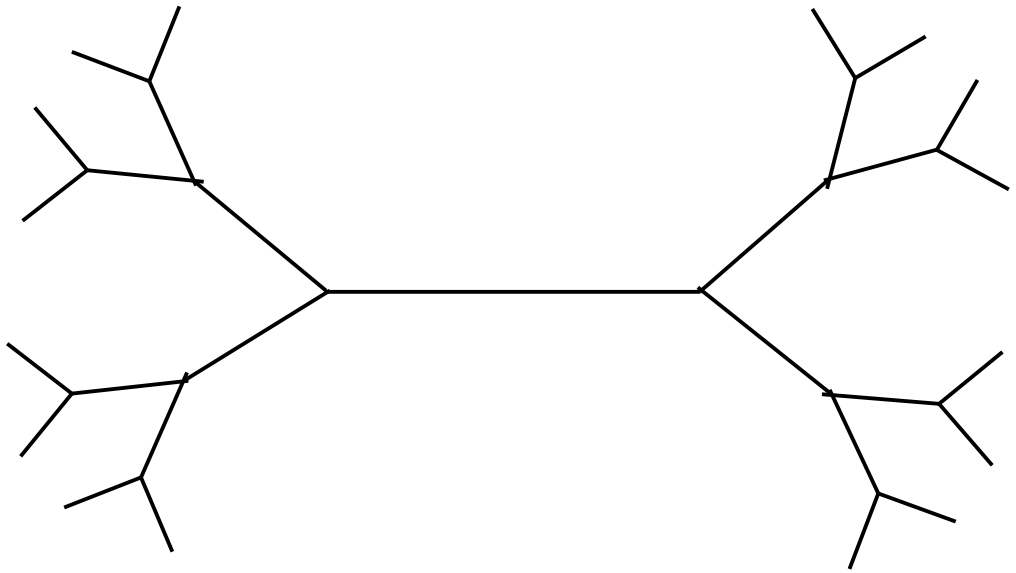
HMMER vs. PaPaRa



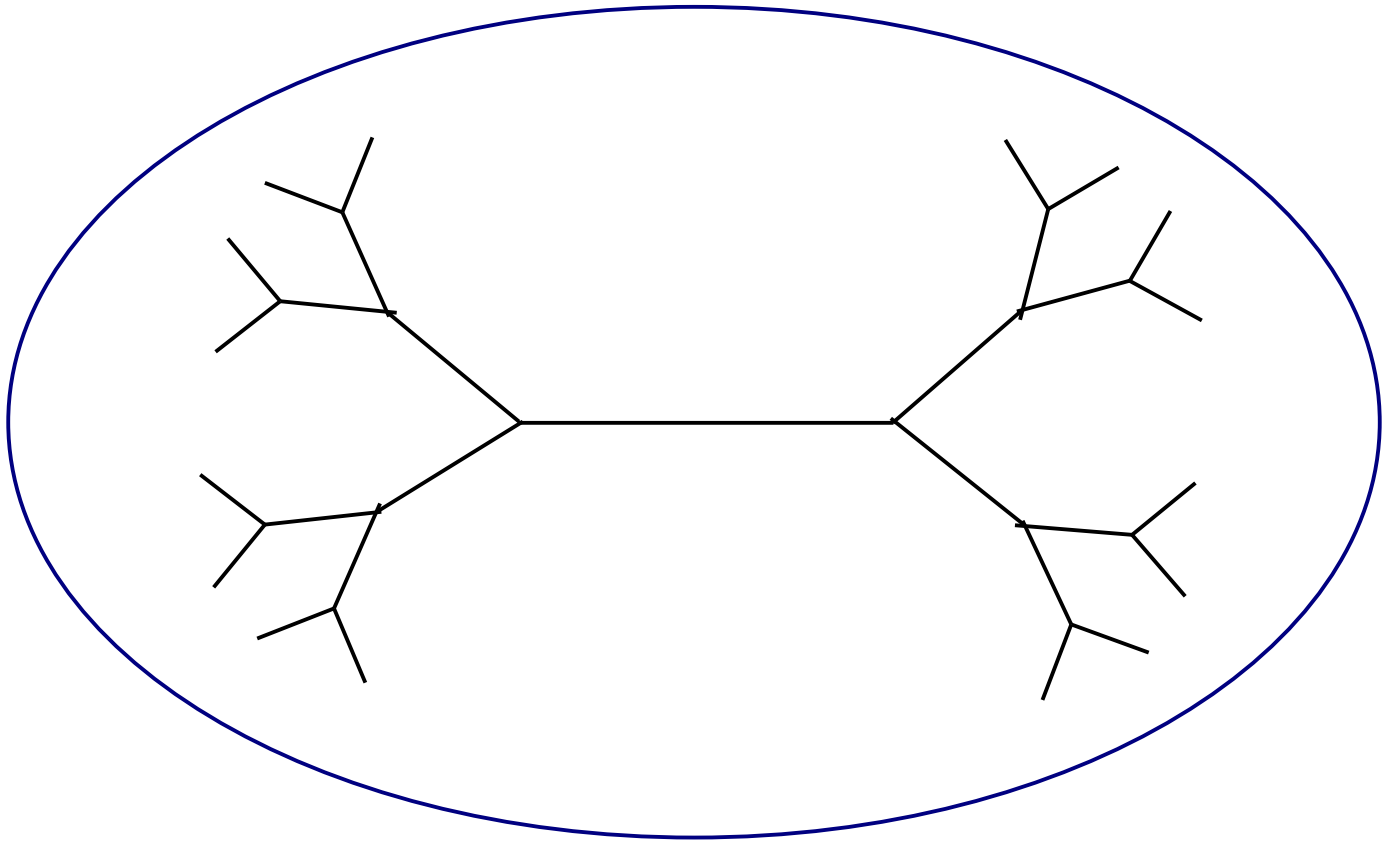
SEPP

- Key insight: HMMs are not very good at modelling MSAs on large, divergent datasets.
- Approach: insert fragments into taxonomy using estimated alignment of full-length sequences, and multiple HMMs (on different subsets of taxa).

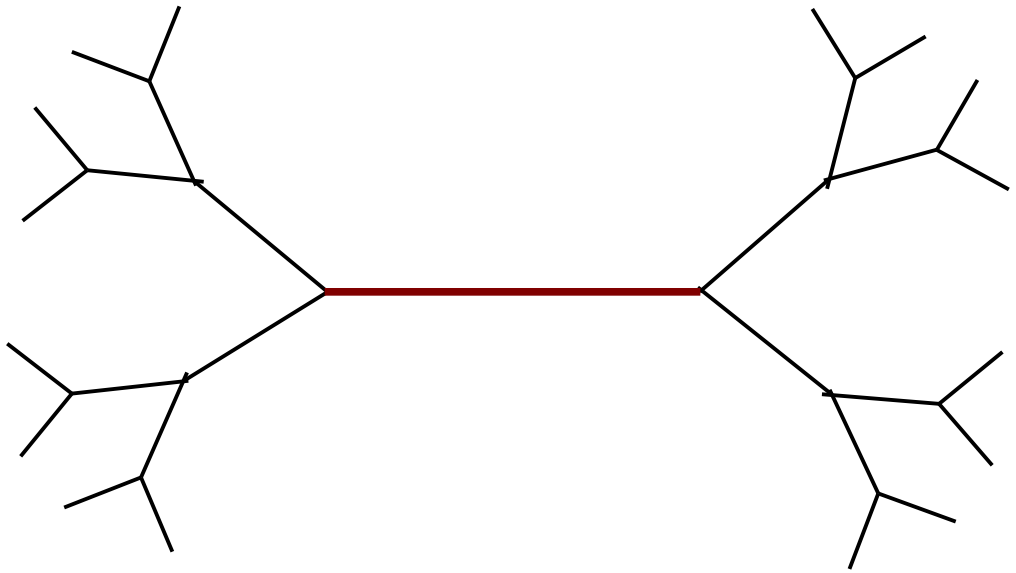
SEPP: SATé-enabled Phylogenetic Placement



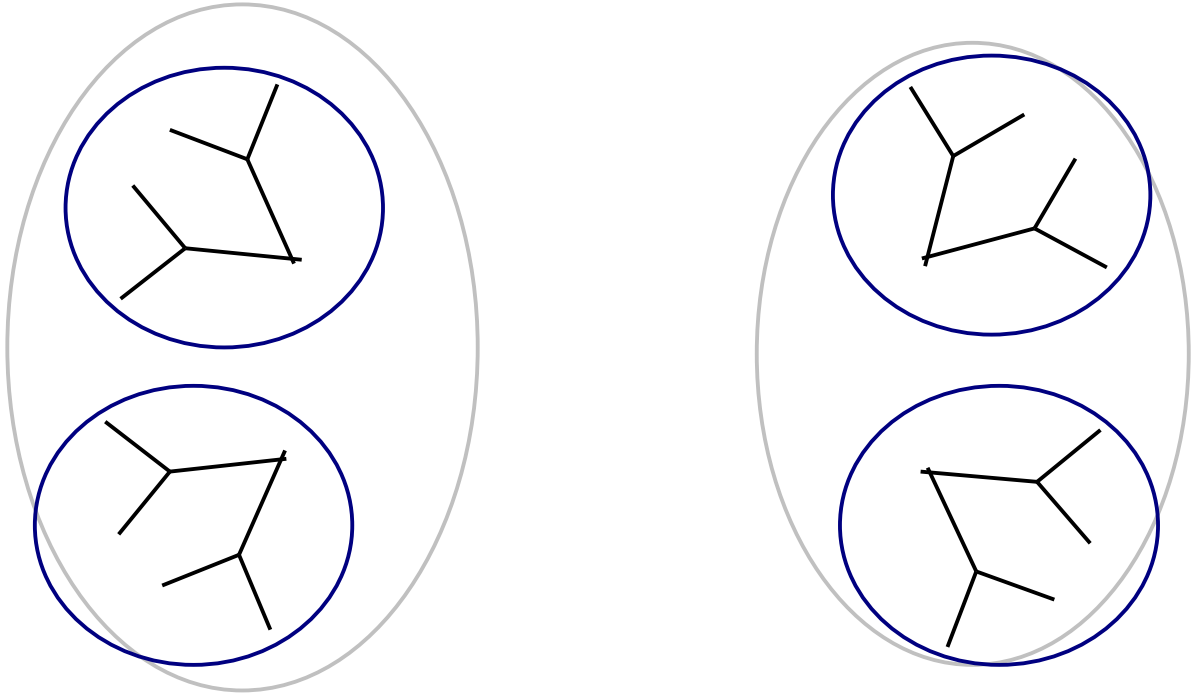
SEPP: SATé-enabled Phylogenetic Placement



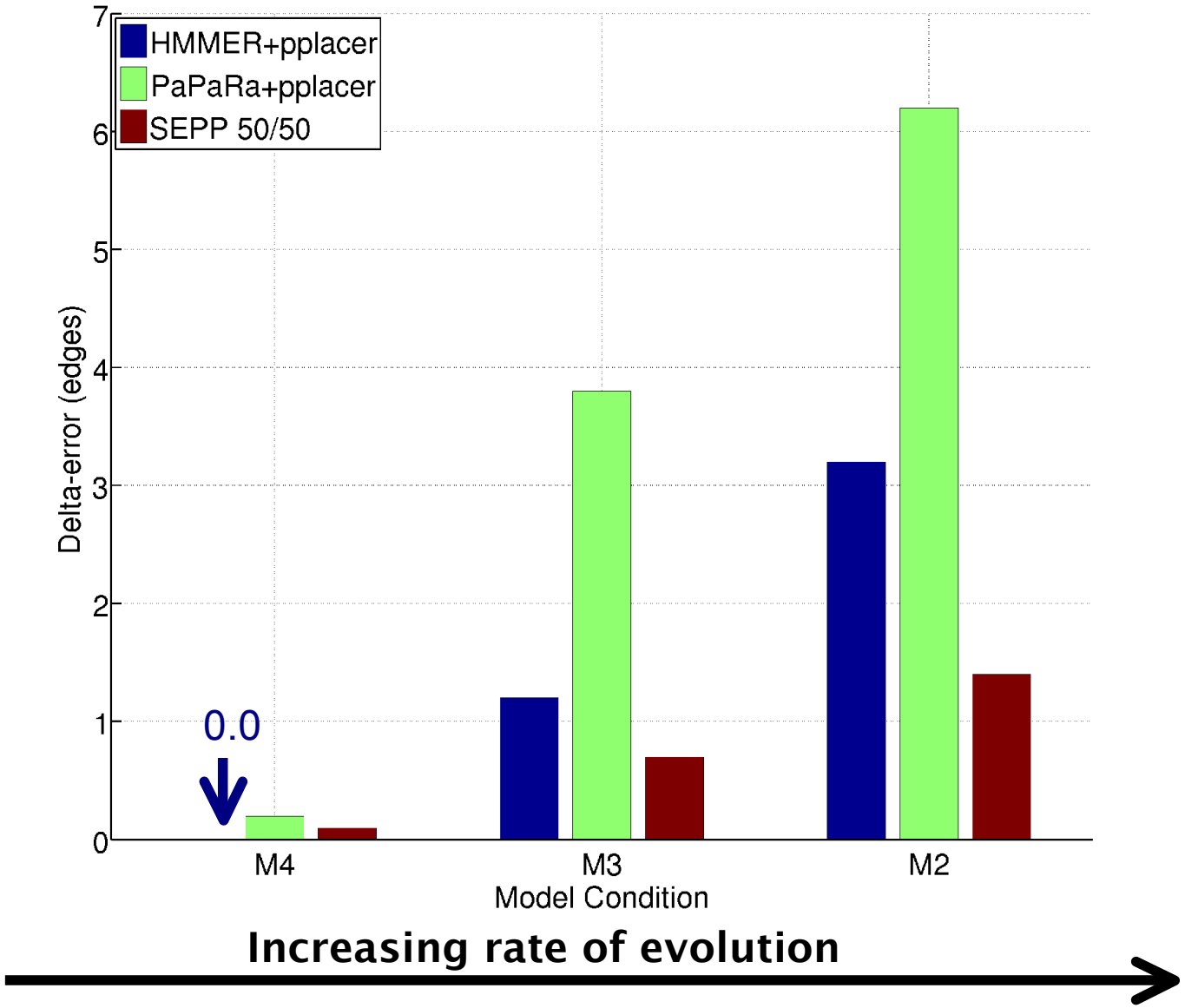
SEPP: SATé-enabled Phylogenetic Placement



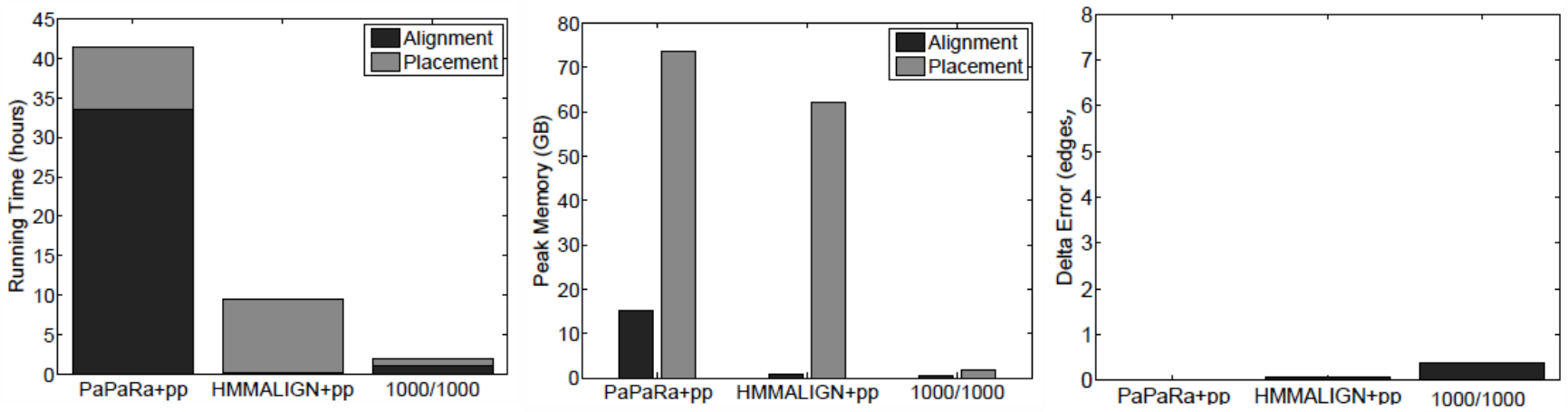
SEPP: SATé-enabled Phylogenetic Placement



SEPP (10%-rule) on Simulated Data

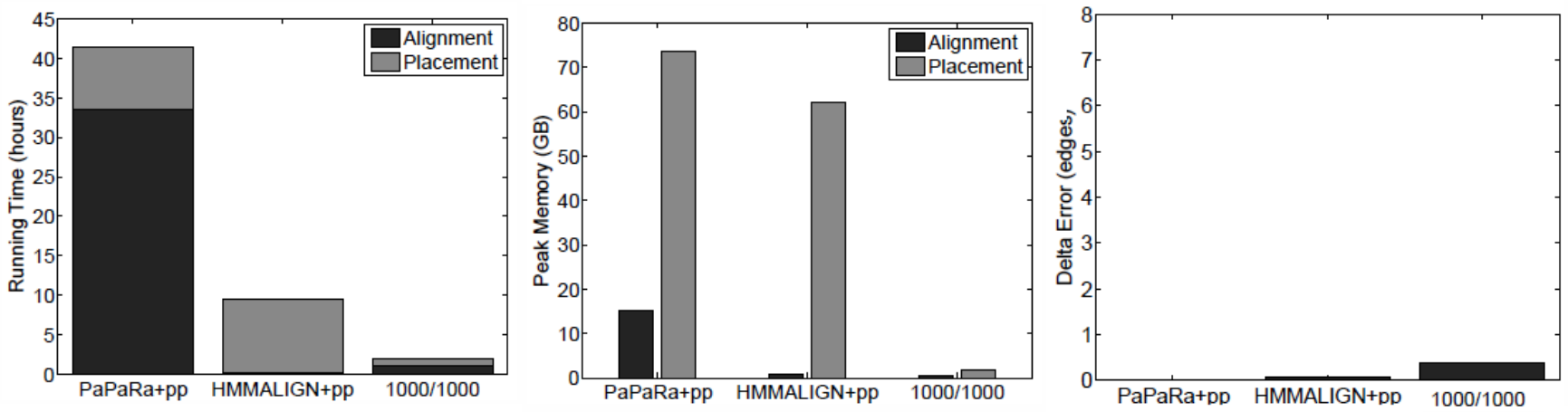


SEPP (10%) on Biological Data



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

Summary

SATé improves accuracy for large-scale alignment and tree estimation

DACTAL enables phylogeny estimation for very large datasets, and may be robust to model violations

SEPP is useful for phylogenetic placement of short reads

Main observation: divide-and-conquer can make base methods more accurate (and maybe even faster)

References

For papers, see http://www.cs.utexas.edu/users/tandy/papers.html and note numbers listed below

SATé: Science 2009 (papers #89, #99)

DACTAL: To appear, Bioinformatics (special issue for ISMB 2012)

SEPP: Pacific Symposium on Biocomputing (#104)

For software, see

http://www.cs.utexas.edu/~phylo/software/

Research Projects

Theory: Phylogenetic estimation under statistical models

Method development:

- "Absolute fast converging" methods
- Very large-scale multiple sequence alignment and phylogeny estimation
- Estimating species trees and networks from gene trees
- Supertree methods
- Comparative genomics (genome rearrangement phylogenetics)
- Metagenomic taxon identification
- Alignment and Phylogenetic Placement of NGS data

Dataset analyses

- Avian Phylogeny: 50 species and 8000+ genes
- Thousand Transcriptome (1KP) Project: 1000 species and 1000 genes
- Chloroplast genomics

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:
  - SATé: Mark Holder, Randy Linder, Kevin Liu, Siavash Mirarab, Serita Nelesen, Sindhu Raghavan, Li-San Wang, and Jiaye Yu
  - DACTAL: Serita Nelesen, Kevin Liu, Li-San Wang, and Randy Linder
  - SEPP/TIPP: Siavash Mirarab and Nam Nguyen

- Software: see http://www.cs.utexas.edu/users/phylo/software/