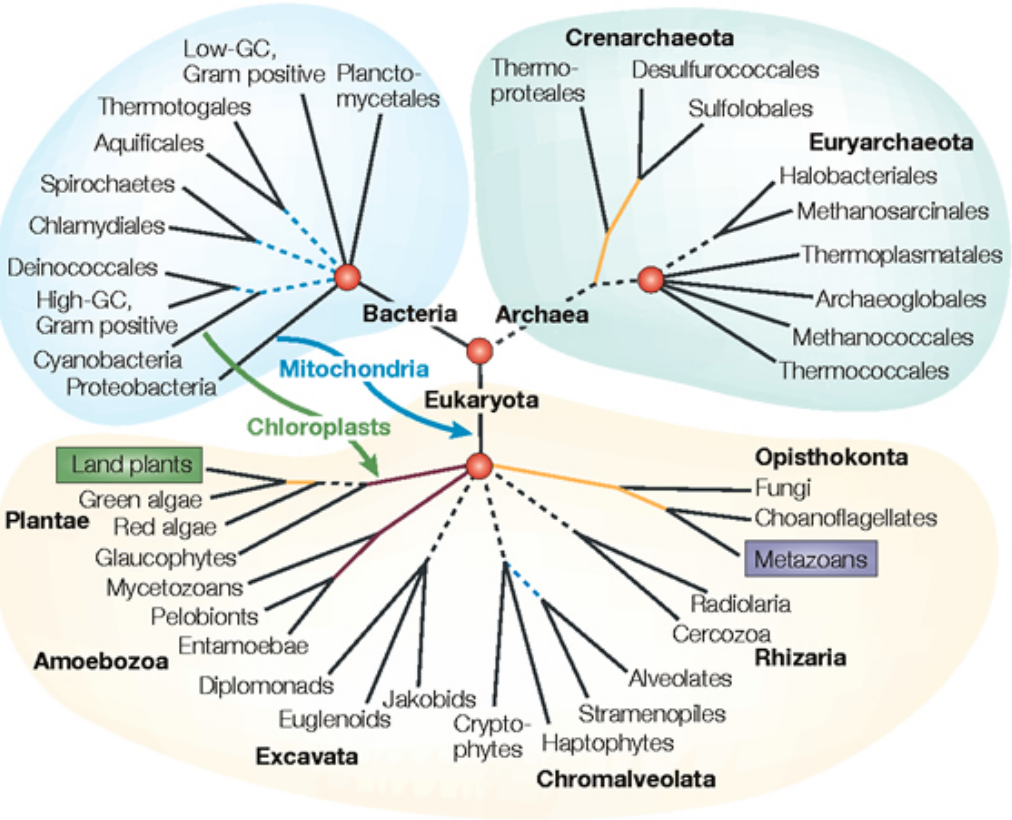
New Multiple Sequence Alignment Methods

Tandy Warnow

The Department of Computer Science

The University of Texas at Austin

The “Tree of Life”



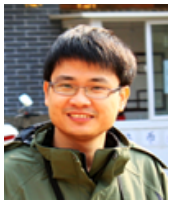
Nature Reviews | Genetics

Avian Phylogenomics Project

Erich Jarvis, HHMI

MTP Gilbert, Copenhagen

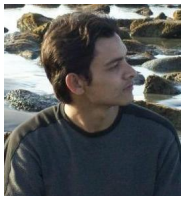
G Zhang, BGI



T. Warnow UT-Austin

S. Mirarab UT-Austin

Md. S. Bayzid UT-Austin



Plus many many other people…

- Approx. 50 species, whole genomes
- 8000+ genes, UCEs
- Gene sequence alignments and trees computed using SATé

Challenges:

Maximum likelihood tree estimation on multi-million-site sequence alignments

Massive gene tree incongruence

1kp: Thousand Transcriptome Project

G. Ka-Shu Wong
U Alberta

J. Leebens-Mack
U Georgia

N. Wickett
Northwestern

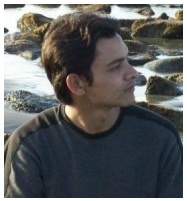
N. Matasci
iPlant

T. Warnow,
UT-Austin

S. Mirarab,
UT-Austin

N. Nguyen,
UT-Austin

Md. S.Bayzid
UT-Austin

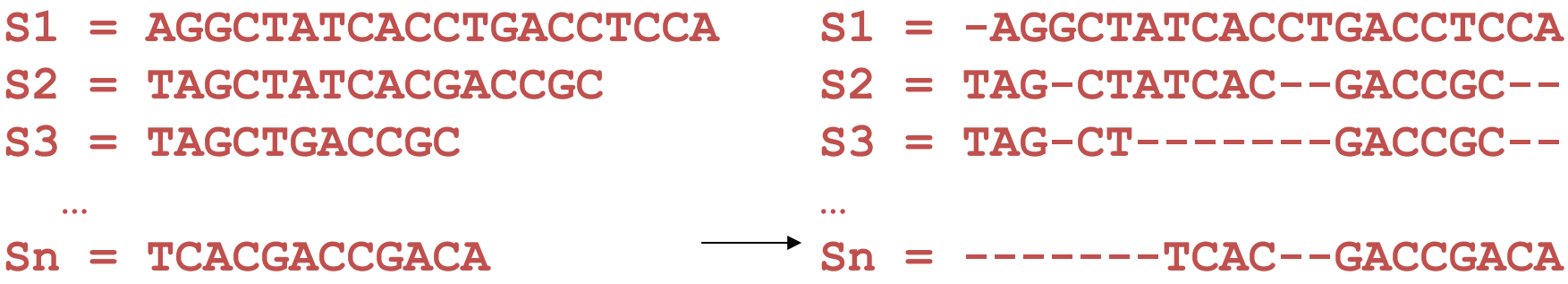


Plus many many other people…

- Plant Tree of Life based on transcriptomes of ~1200 species
- More than 13,000 gene families (most not single copy)

Challenge:
Alignment of datasets with > 100,000 sequences
Gene tree incongruence

Multiple Sequence Alignment (MSA): *another grand challenge*[1]



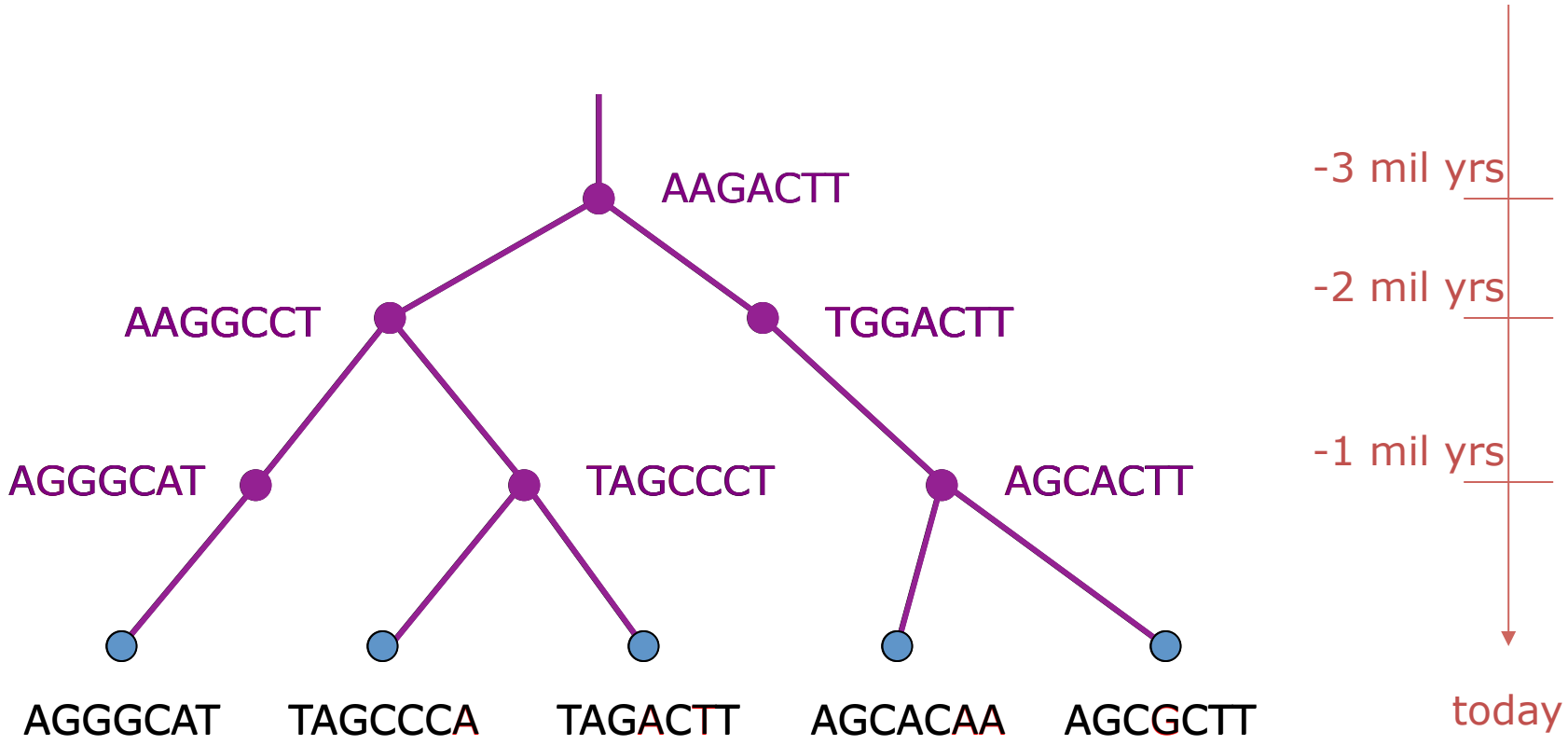
Novel techniques needed for scalability and accuracy

- NP-hard problems and large datasets
- Current methods do not provide good accuracy
- Few methods can analyze even moderately large datasets

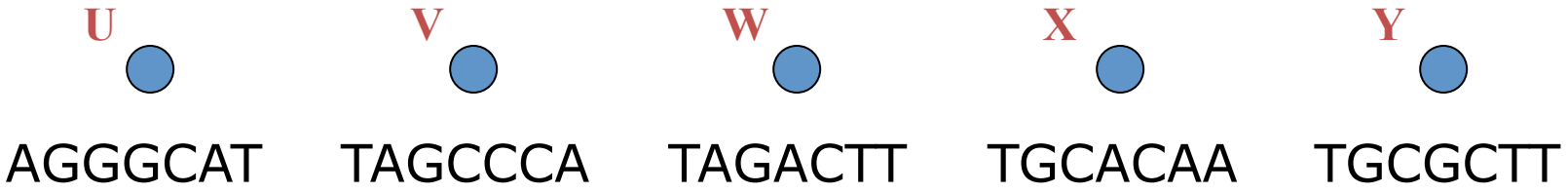
Many important applications besides phylogenetic estimation

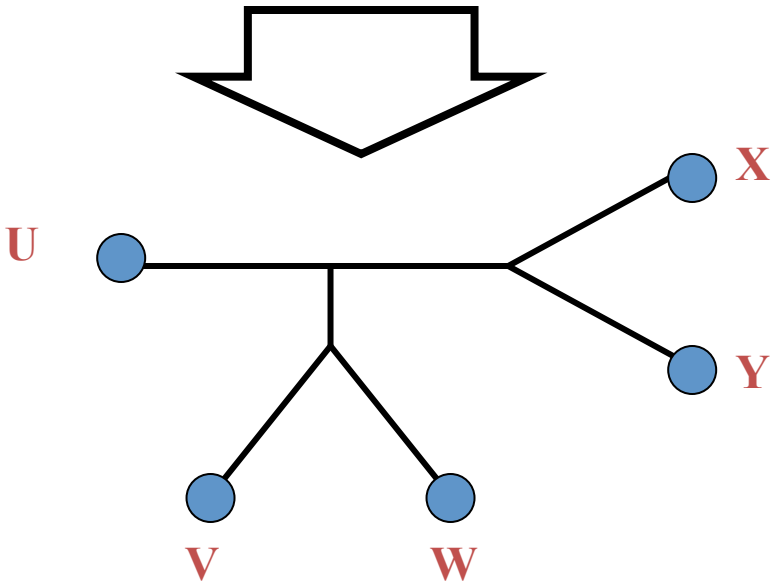
[1] Frontiers in Massive Data Analysis, National Academies Press, 2013

DNA Sequence Evolution

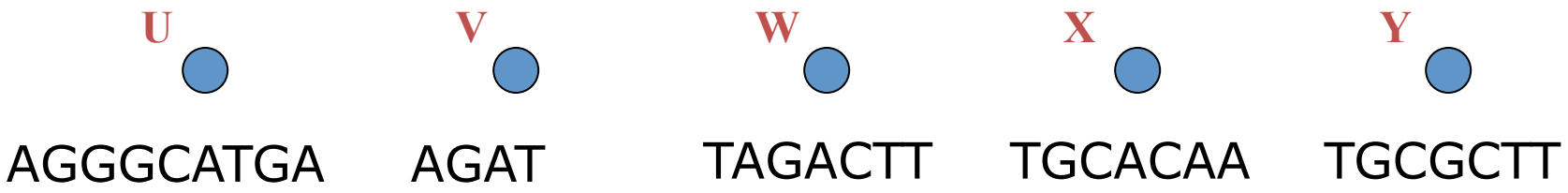


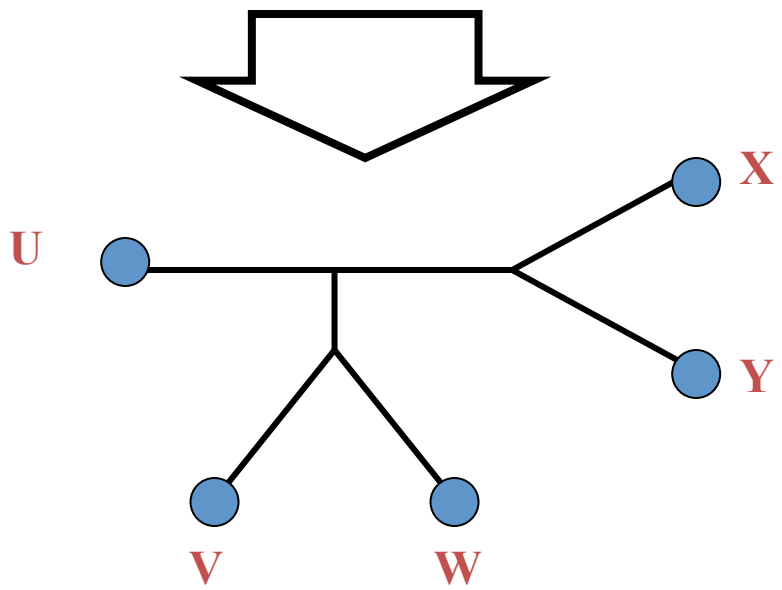
Phylogeny Problem



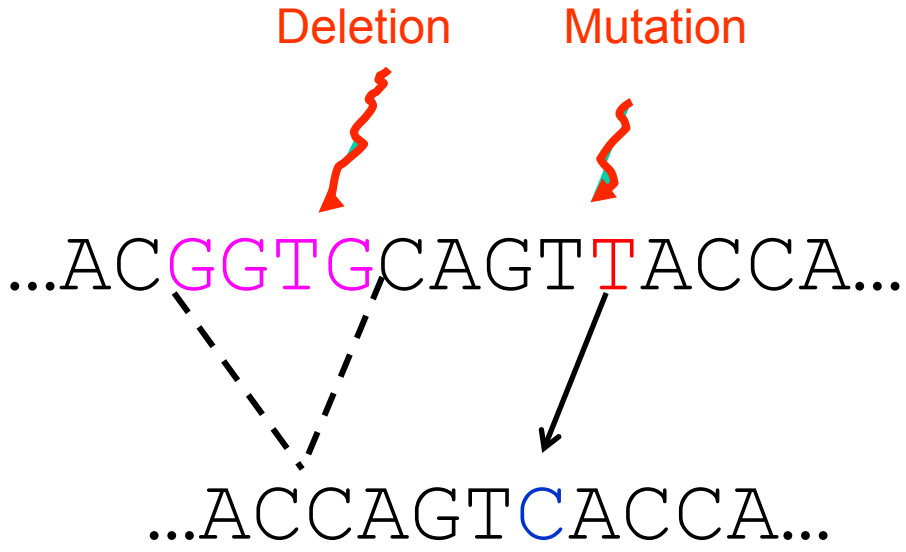


The “real” problem





Indels (insertions and deletions)



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**

Input: unaligned sequences

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

Phase 1: Alignment

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



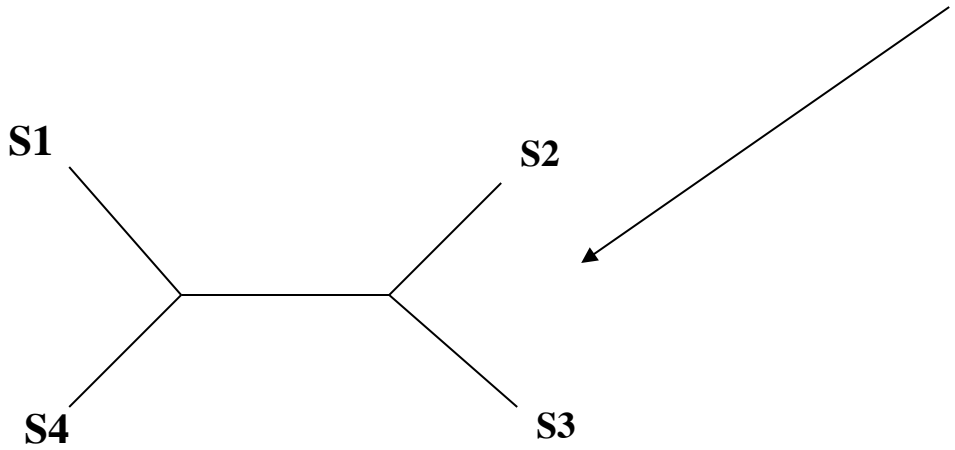
```
S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAC--GACCGC--
S3 = TAG-CT-------GACCGC--
S4 = -------TCAC--GACCGACA
```

Phase 2: Construct tree

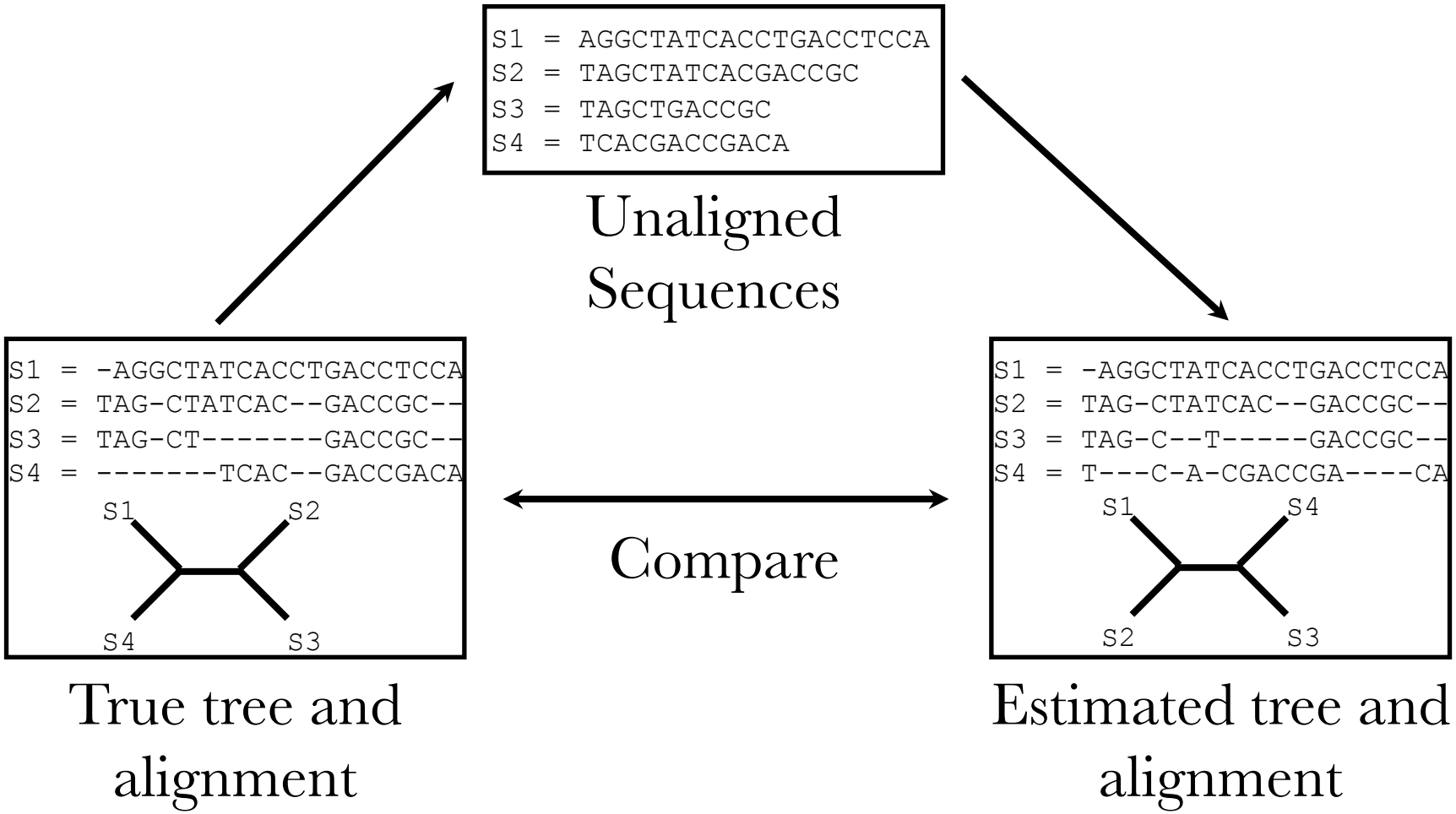
```
S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC
S3 = TAGCTGACCGC
S4 = TCACGACCGACA
```

→

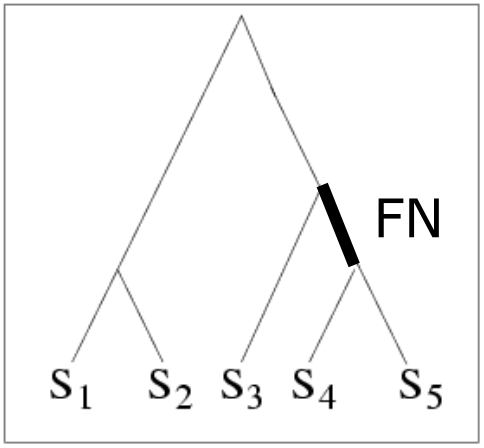
```
S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAC--GACCGC--
S3 = TAG-CT-------GACCGC--
S4 = -------TCAC--GACCGACA
```



Simulation Studies



Quantifying Error



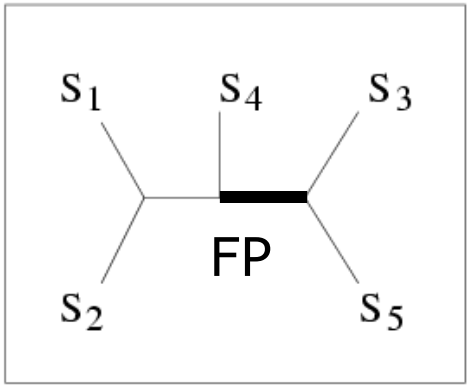
TRUE TREE

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

Two-phase estimation

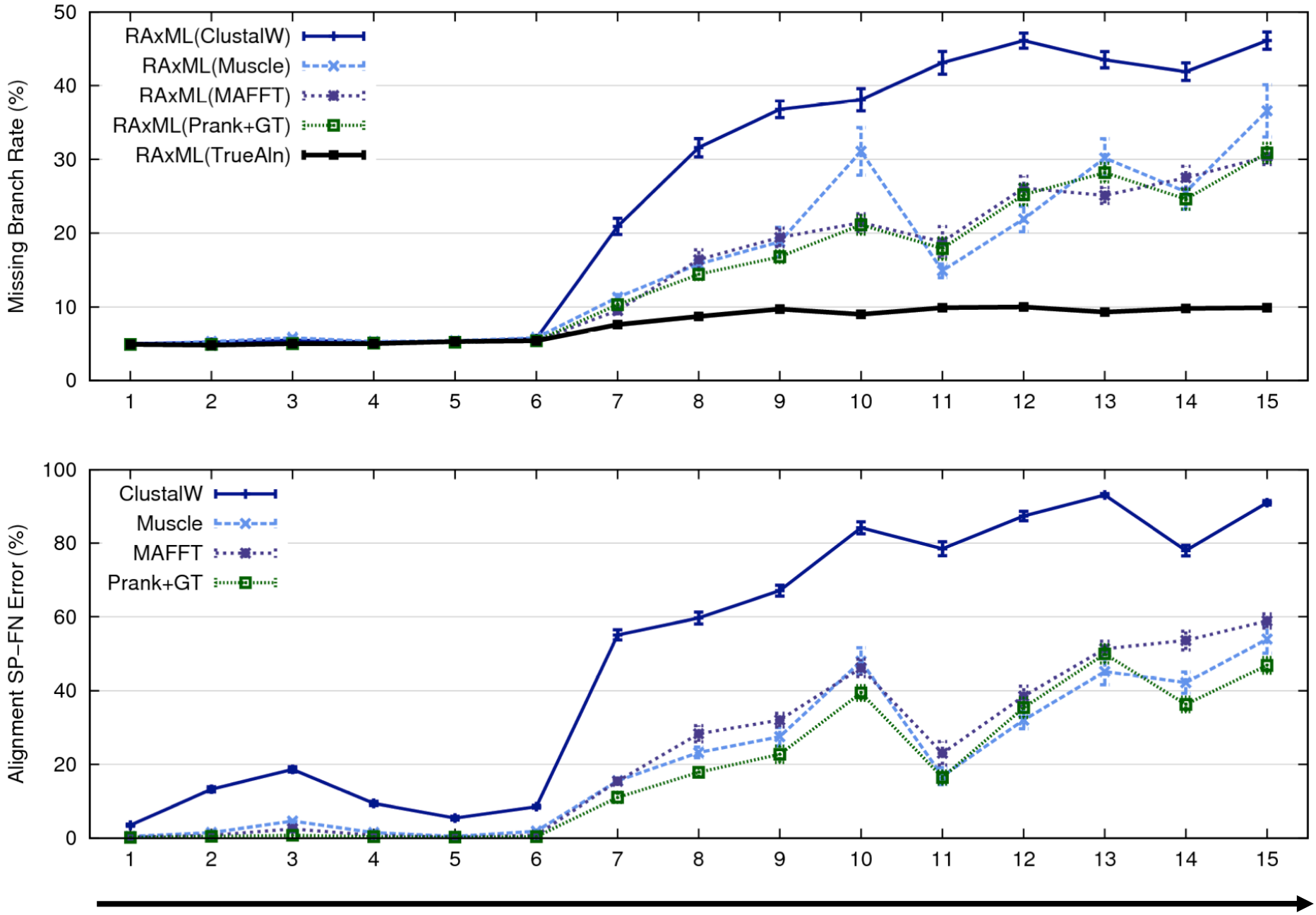
Alignment methods

- 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 with the two-phase approach

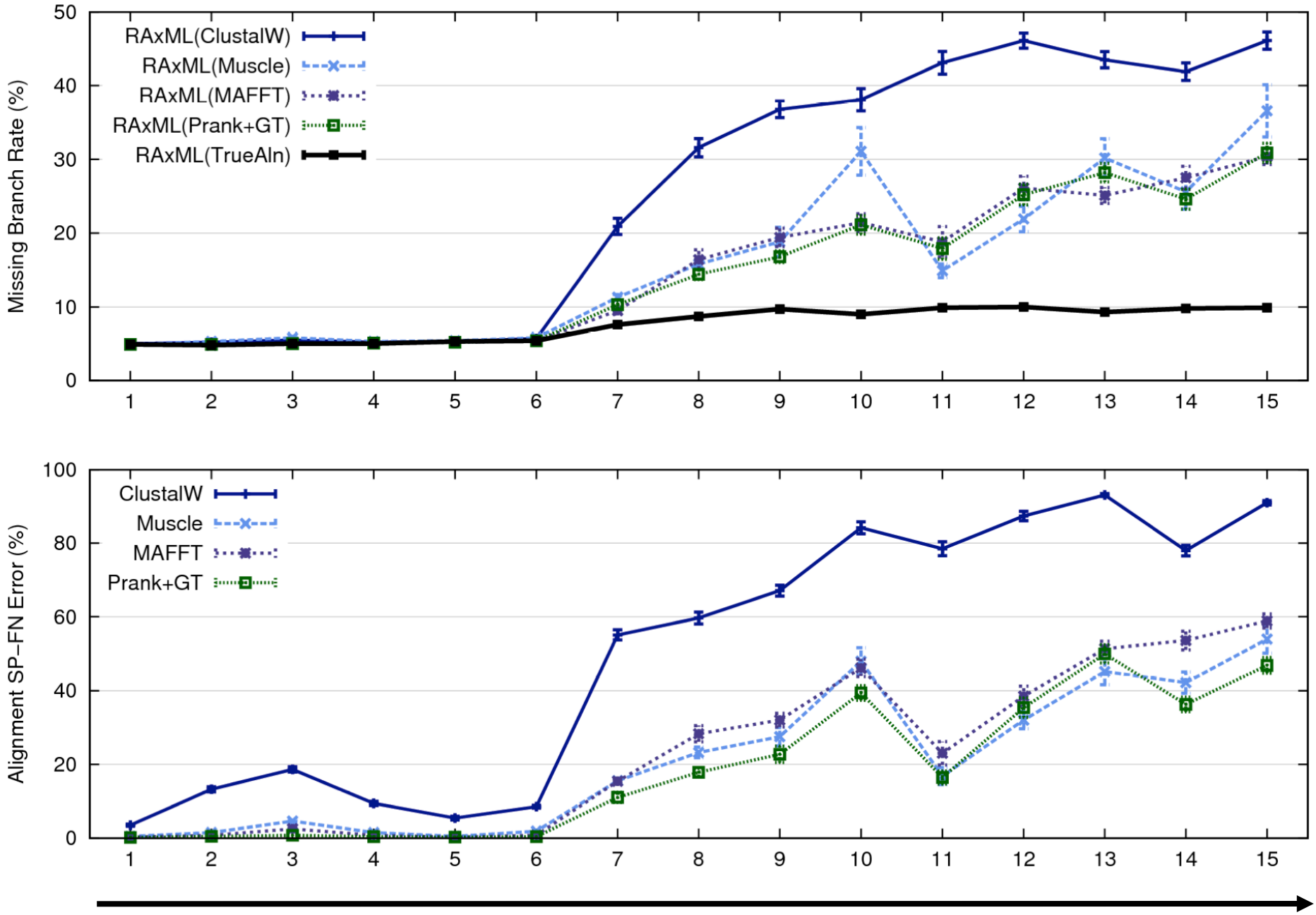
- Current alignment methods fail to return reasonable alignments on large datasets with high rates of indels and substitutions.
- Manual alignment is time consuming and subjective.
- *Systematists discard potentially useful markers* if they are difficult to align.

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

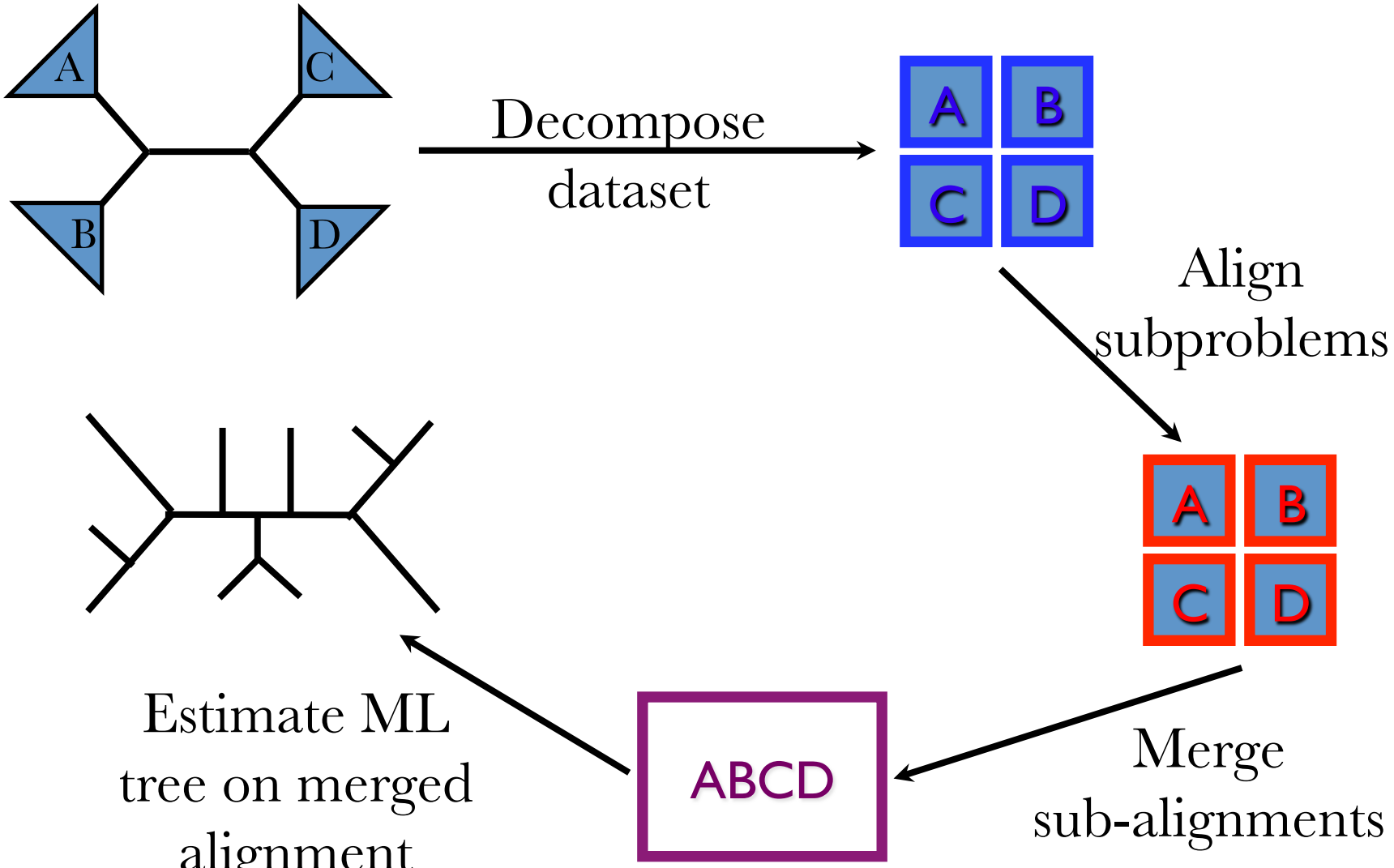
SATé

SATé (Simultaneous Alignment and Tree Estimation)

- Liu et al., Science 2009
- Liu et al., Systematic Biology 2012
- Public distribution (open source software) and user-friendly GUI

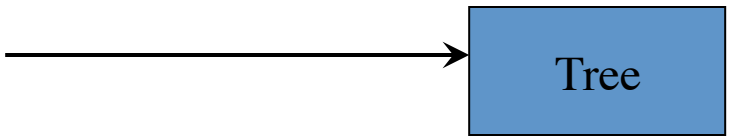

Missing Branch Rate (%)
RAxML(ClustalW)
RAxML(Muscle)
RAxML(MAFFT)
RAxML(Prank+GT)
RAxML(TrueAln)
Alignment SP–FN Error (%)
ClustalW
Muscle
MAFFT
Prank+GT
1000-taxon models, ordered by difficulty (Liu et al., 2009)

Re-aligning on a tree



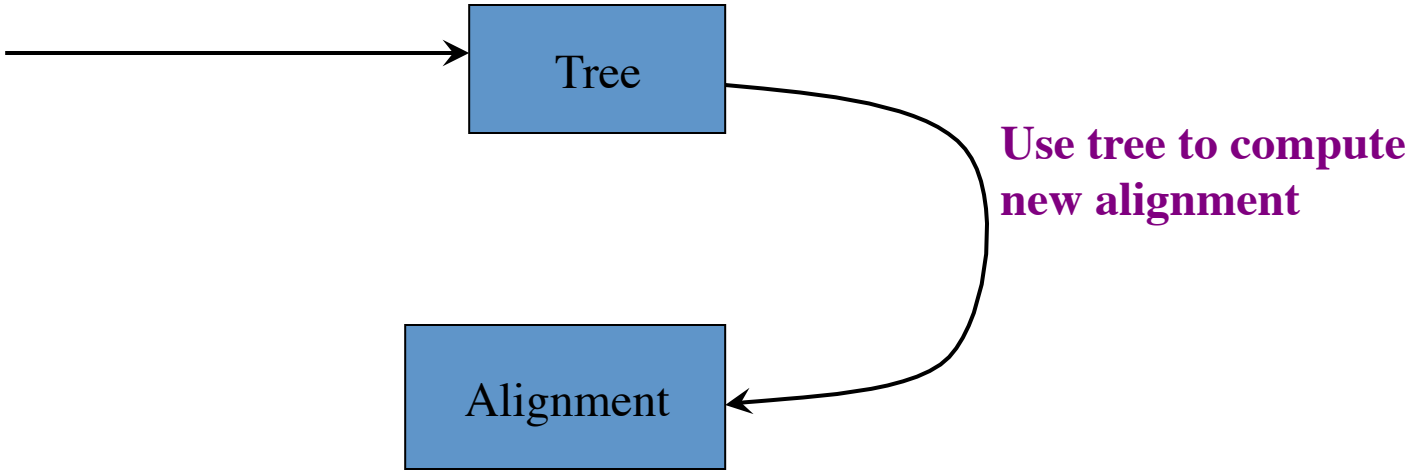
SATé Algorithm

Obtain initial alignment and
estimated ML tree



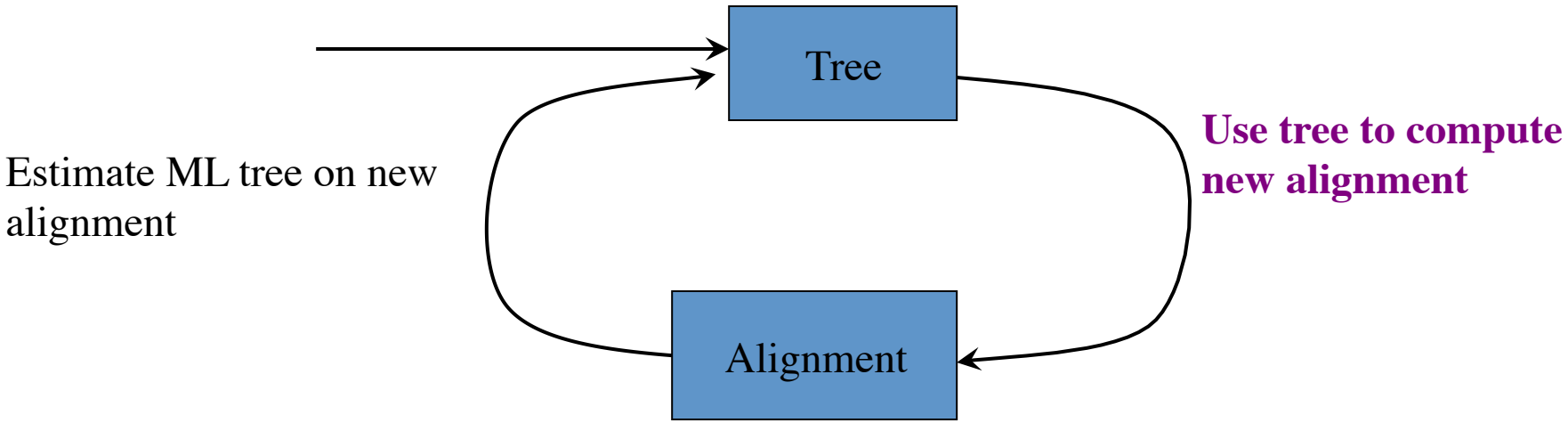
SATé Algorithm

Obtain initial alignment and
estimated ML tree



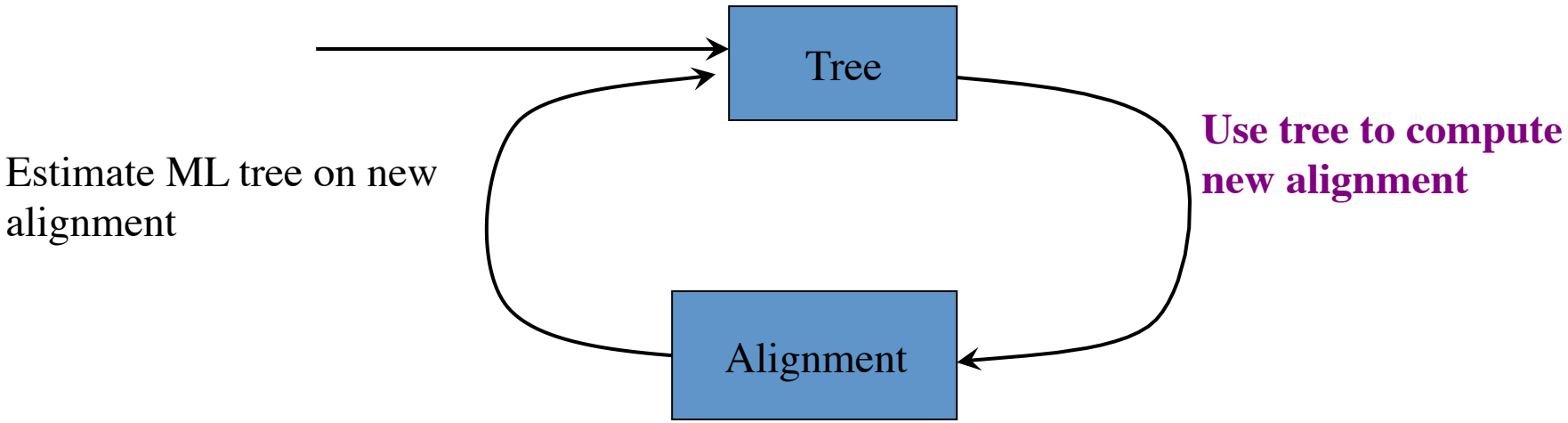
SATé Algorithm

Obtain initial alignment and estimated ML tree



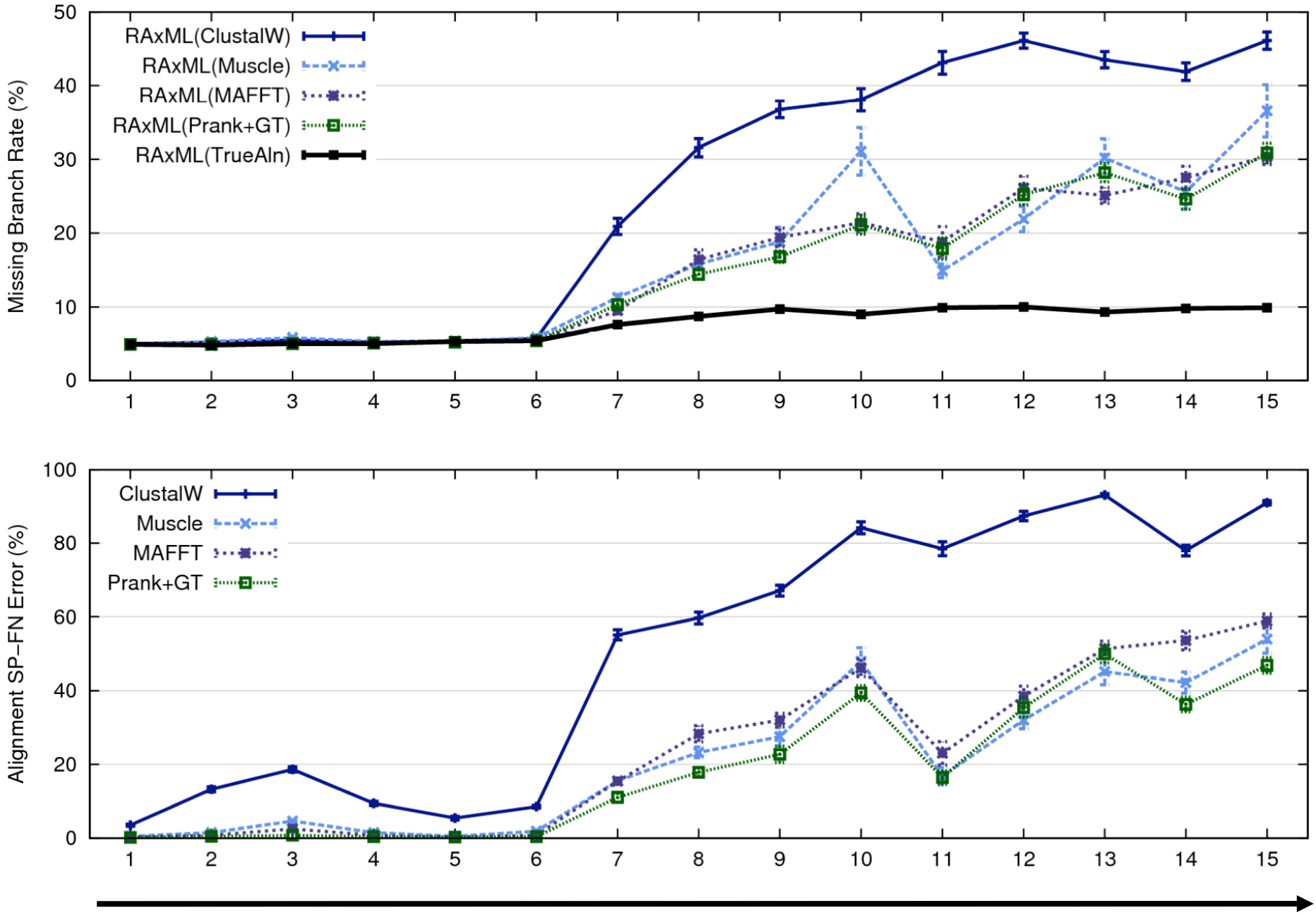
SATé Algorithm

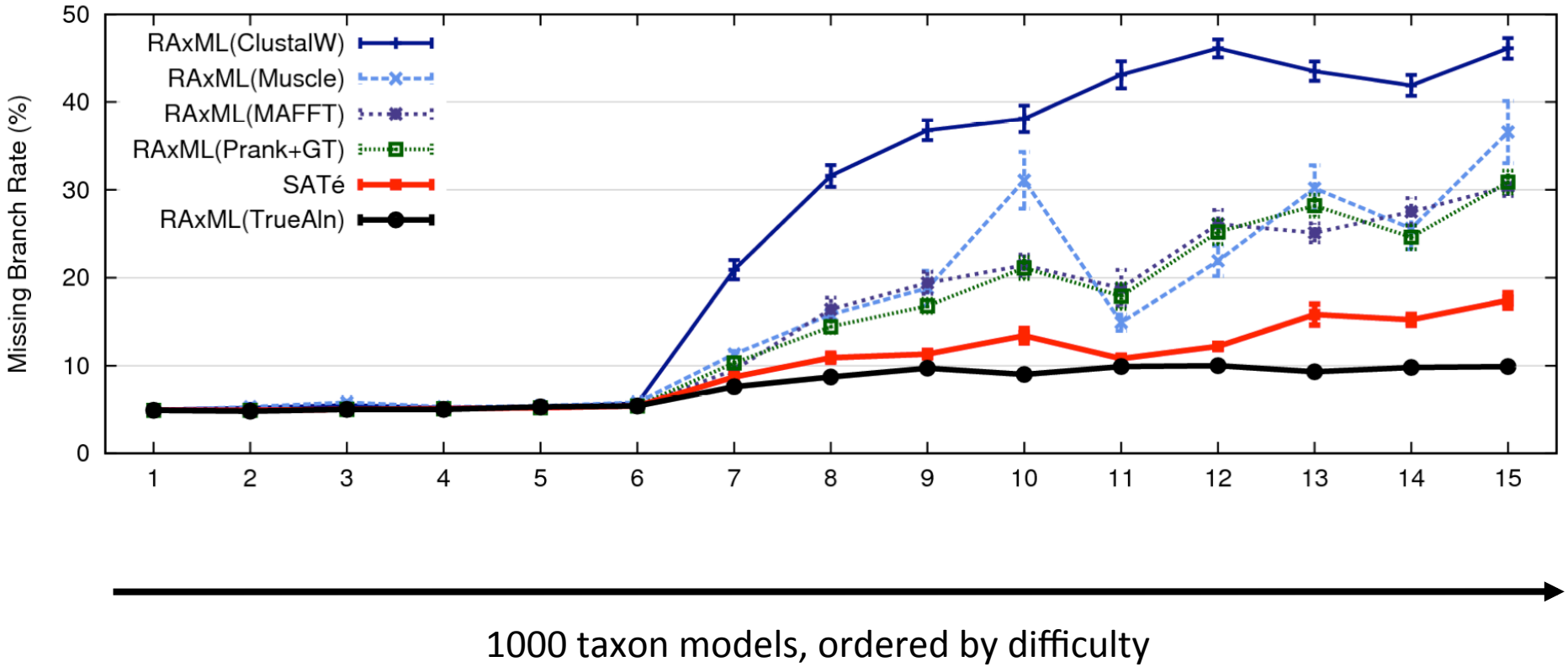
Obtain initial alignment and estimated ML tree



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

Repeat until termination condition (typically, 24 hours)

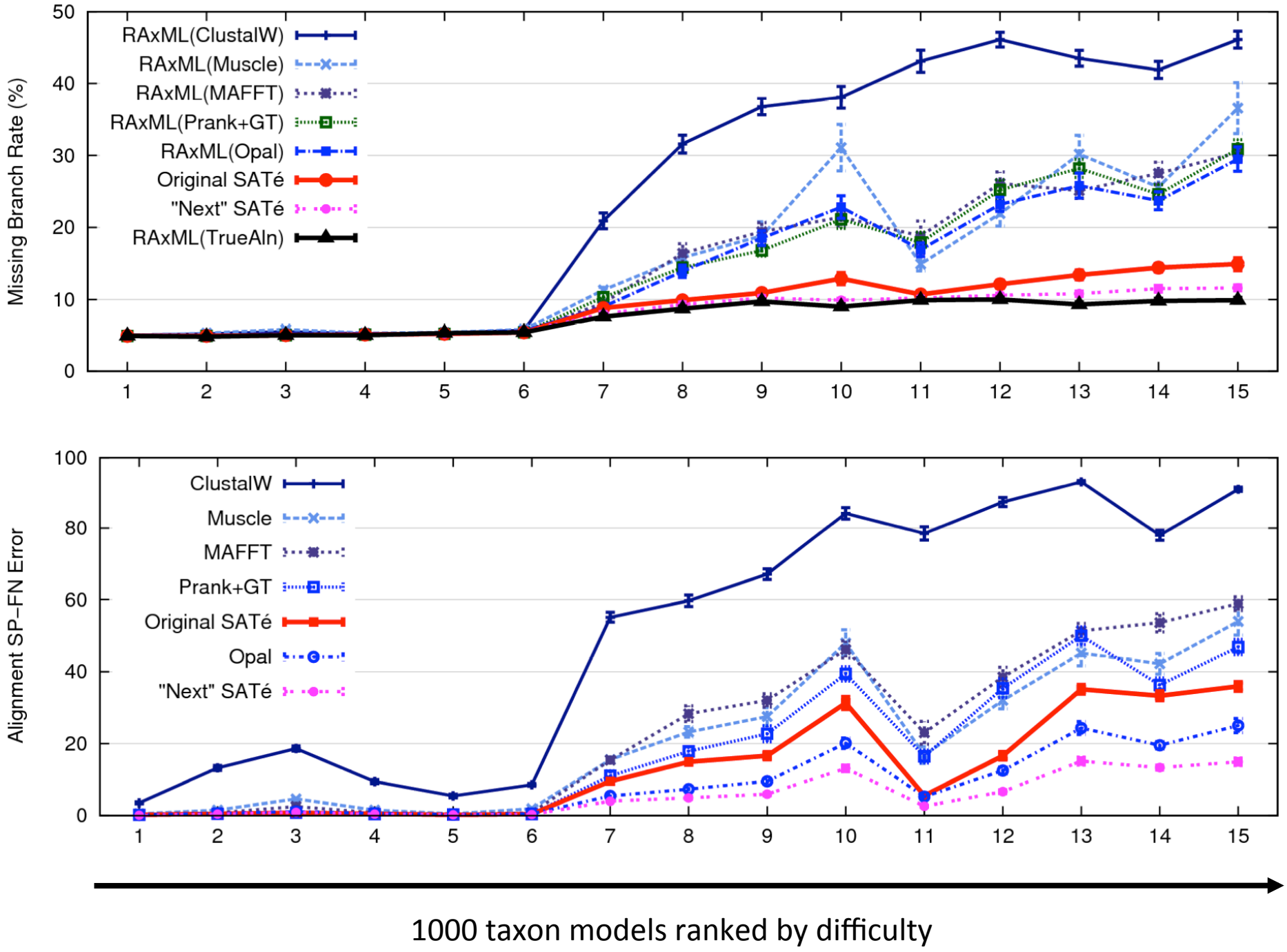

Missing Branch Rate (%)
RAxML(ClustalW)
RAxML(Muscle)
RAxML(MAFFT)
RAxML(Prank+GT)
RAxML(TrueAln)
Alignment SP–FN Error (%)
ClustalW
Muscle
MAFFT
Prank+GT
1000-taxon models, ordered by difficulty (Liu et al., 2009)



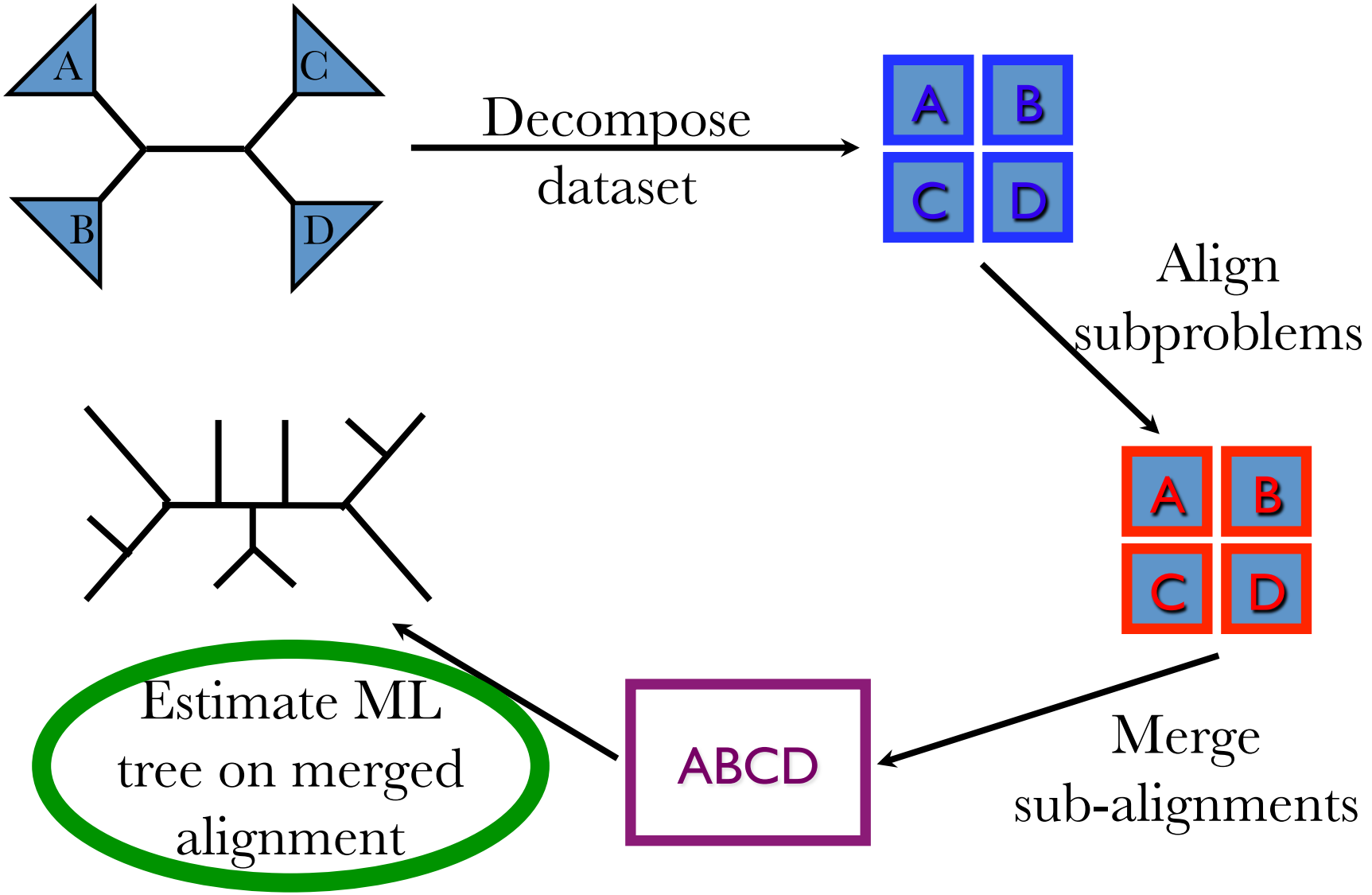
1000 taxon models, ordered by difficulty

24 hour SATé analysis, on desktop machines

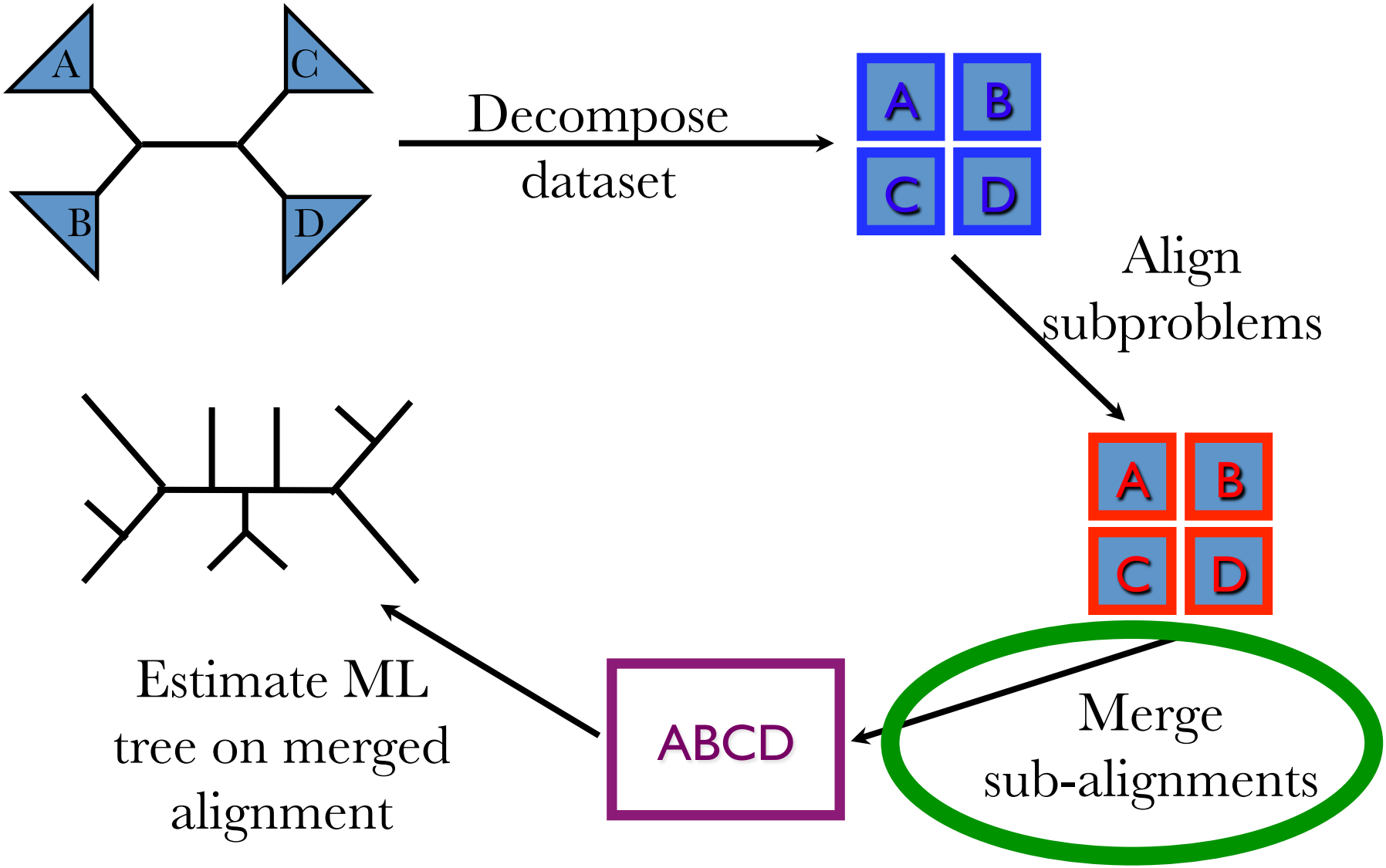
(Similar improvements for biological datasets)


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

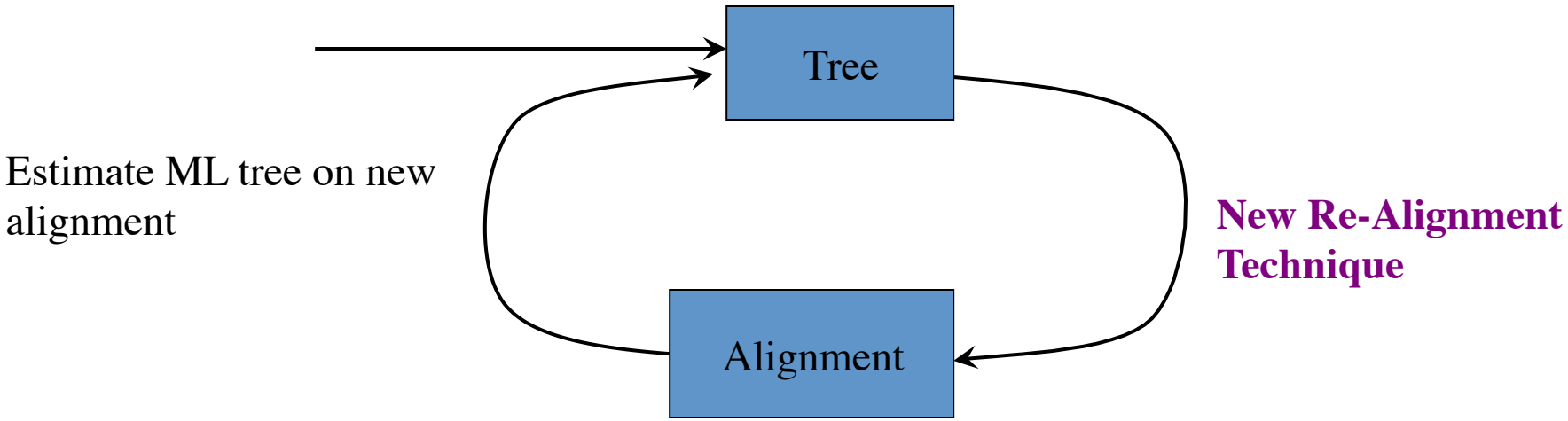


Limitations



PASTA Algorithm (Improving SATé)

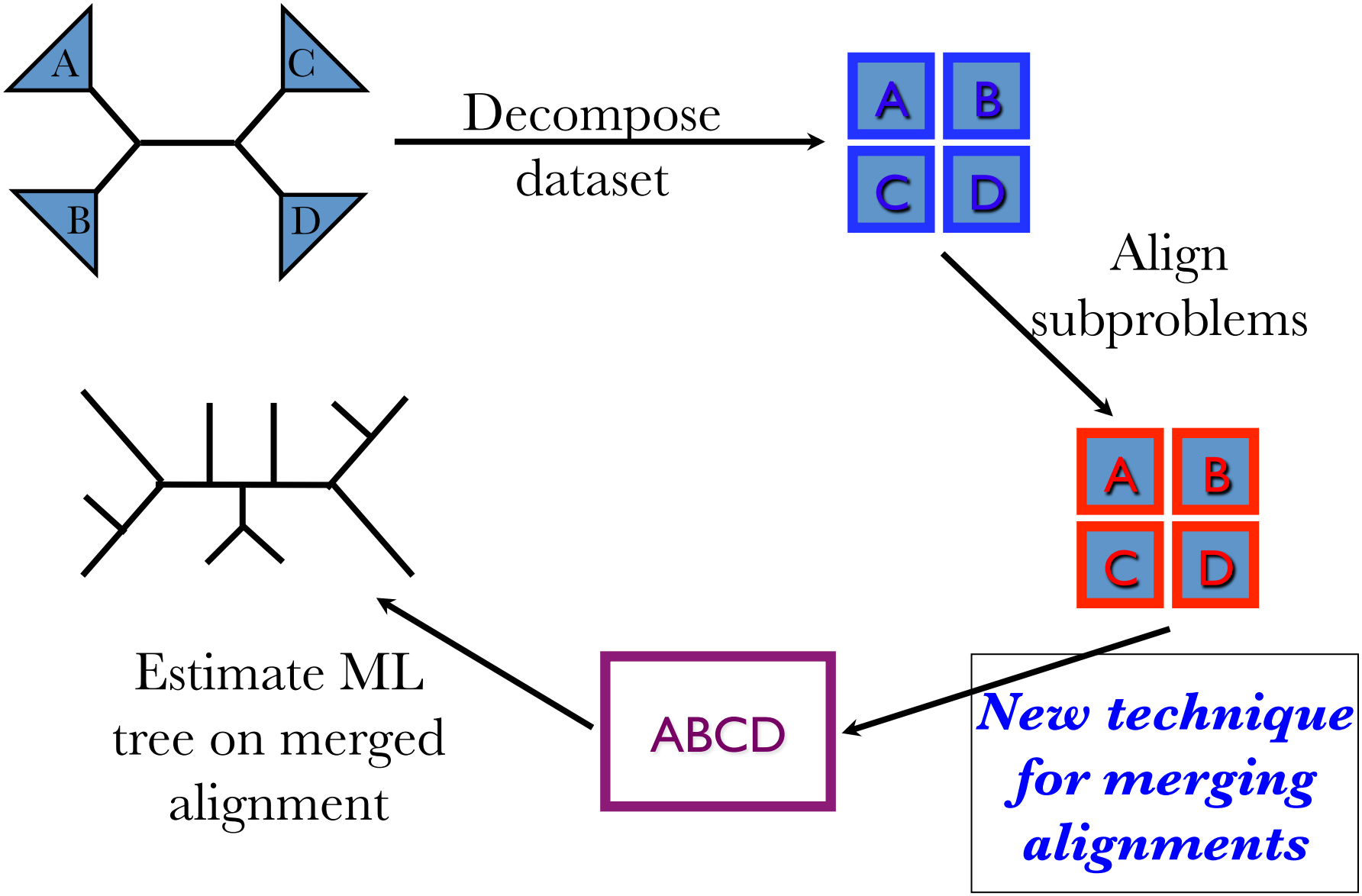
Obtain initial alignment and estimated ML tree



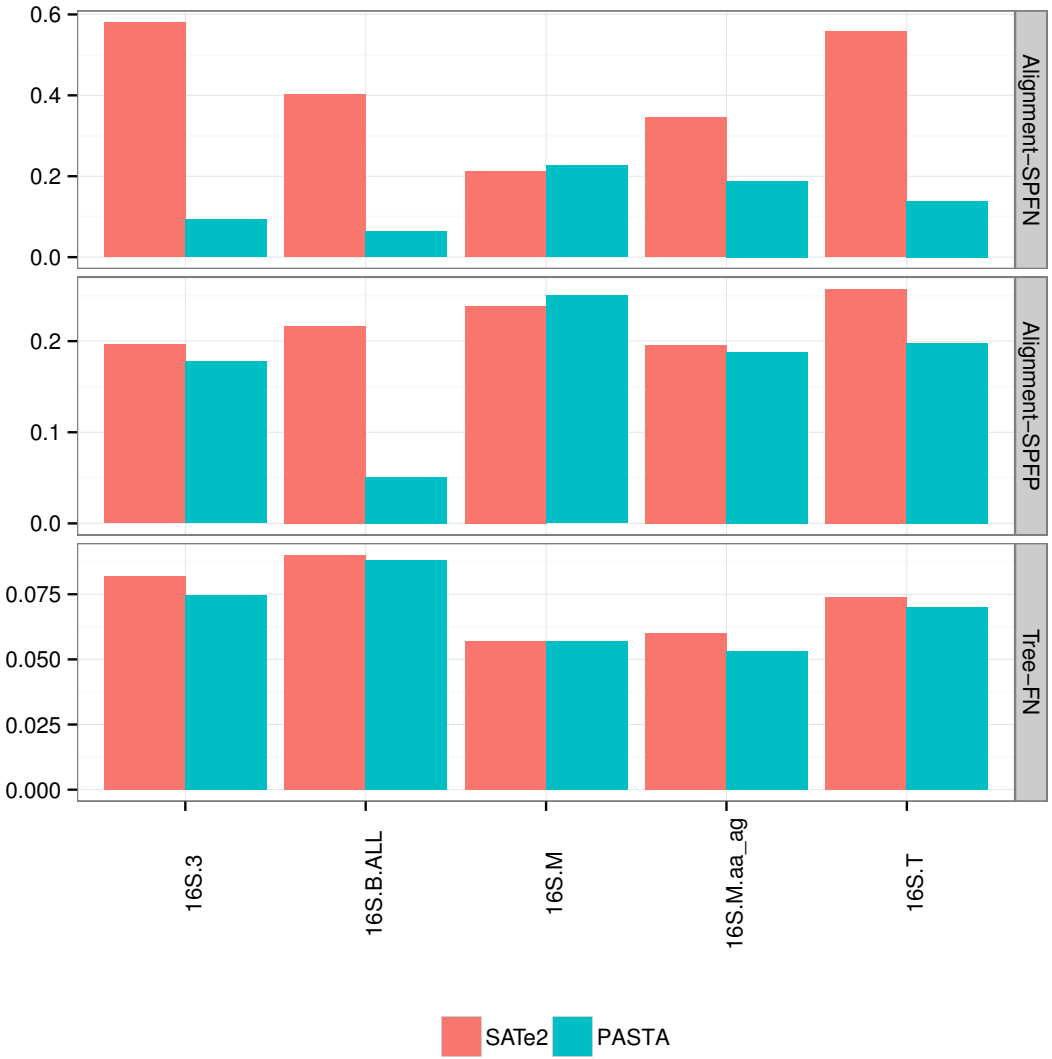
If new alignment/tree pair has worse ML score, realign using a different decomposition

Repeat until termination condition (typically, 24 hours)

Each PASTA iteration



PASTA vs. SATé: better alignments and trees, better scalability

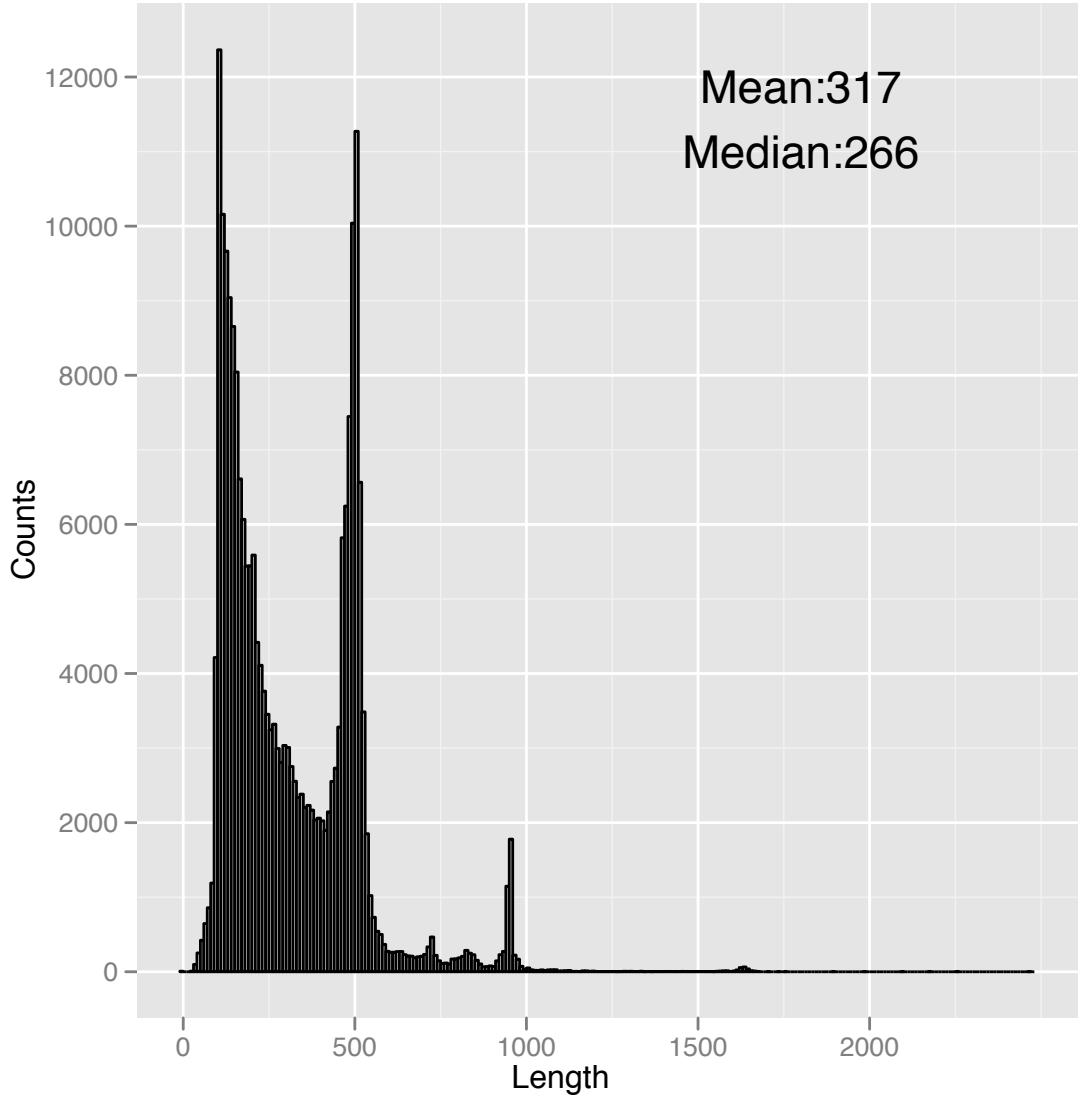


Analyses of Gutell's 16S datasets with curated structural alignments and reference trees using maximum likelihood with bootstrapping.

PASTA can analyze datasets with up to 200,000 sequences efficiently.

SATé maxes out around 50,000 sequences.

1KP dataset: Large *and* Highly Fragmentary!



Cytochrome dataset sequence length distribution

>100,000 AA sequences
Transcriptome data

UPP: basic idea

Input: set S of unaligned sequences

Output: alignment on S

- Select random subset X of S
- Estimate “backbone” alignment A and tree T on X
- Independently align each sequence in S-X to A
- Use transitivity to produce multiple sequence alignment A* for entire set S

Input: Unaligned Sequences

```
S1  = AGGCTATCACCTGACCTCCAAT
S2  = TAGCTATCACGACCGCGCT
S3  = TAGCTGACCGCGCT
S4  = TACTCACGACCGACAGCT
S5  = TAGGTACAACCTAGATC
S6  = AGATACGTCGACATATC
```

Step 1: Pick random subset (backbone)

```
S1  = AGGCTATCACCTGACCTCCAAT
S2  = TAGCTATCACGACCGCGCT
S3  = TAGCTGACCGCGCT
S4  = TACTCACGACCGACAGCT
S5  = TAGGTACAACCTAGATC
S6  = AGATACGTCGACATATC
```

Step 2: Compute backbone alignment

```
S1  = -AGGCTATCACCTGACCTCCA-AT
S2  = TAG-CTATCAC--GACCGC--GCT
S3  = TAG-CT-------GACCGC--GCT
S4  = TAC----TCAC--GACCGACAGCT
S5  = TAGGTAAAACCTAGATC
S6  = AGATAAAACTACATATC
```

Step 3: Align each remaining sequence to backbone

First we add S5 to the backbone alignment

```
S1  = -AGGCTATCACCTGACCTCCA-AT-
S2  = TAG-CTATCAC--GACCGC--GCT-
S3  = TAG-CT-------GACCGC--GCT-
S4  = TAC----TCAC--GACCGACAGCT-
S5  = TAGG---T-A-CAA-CCTA--GATC
```

Step 3: Align each remaining sequence to backbone

Then we add S6 to the backbone alignment

```
S1  = -AGGCTATCACCTGACCTCCA-AT-
S2  = TAG-CTATCAC--GACCGC--GCT-
S3  = TAG-CT-------GACCGC--GCT-
S4  = TAC----TCAC--GACCGACAGCT-
S6  = -AG---AT-A-CGTC--GACATATC
```

Step 4: Use transitivity to obtain MSA on entire set

```
S1  = -AGGCTATCACCTGACCTCCA-AT--
S2  = TAG-CTATCAC--GACCGC--GCT--
S3  = TAG-CT-------GACCGC--GCT--
S4  = TAC----TCAC--GACCGACAGCT--
S5  = TAGG---T-A—CAA-CCTA--GATC-
S6  = -AG---AT-A-CGTC--GACATAT-C
```

UPP: details

Input: set S of unaligned sequences

Output: alignment on S

- Select random subset X of S
- Estimate “backbone” alignment A and tree T on X
- Independently align each sequence in S-X to A
- Use transitivity to produce multiple sequence alignment A* for entire set S

UPP: details

Input: set S of unaligned sequences

Output: alignment on S

- Select random subset X of S
- Estimate “backbone” alignment A and tree T on X
- Independently align each sequence in S-X to A
- Use transitivity to produce multiple sequence alignment A* for entire set S

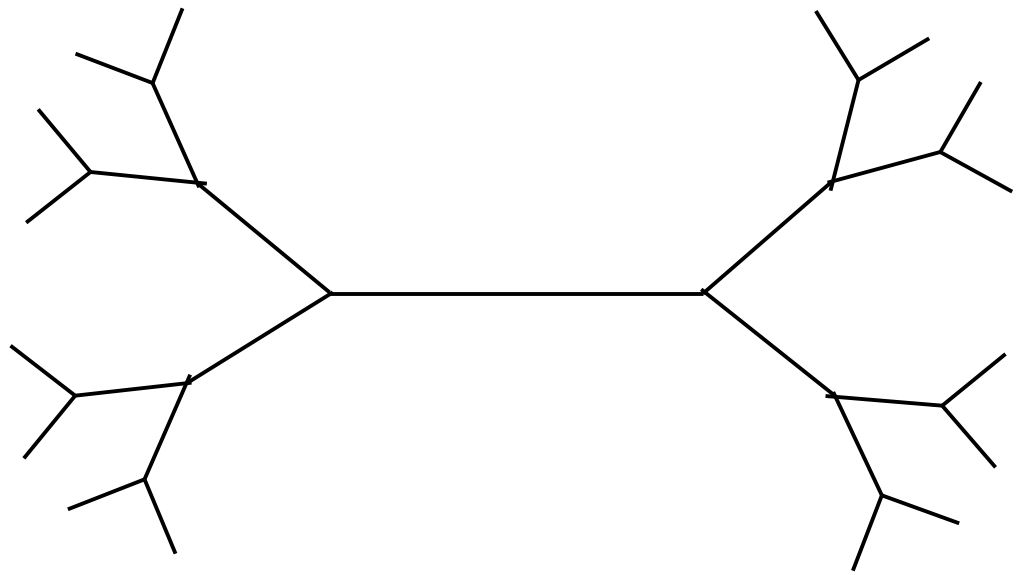
How to align sequences to a backbone alignment?

Standard machine learning technique:

- Build HMM (Hidden Markov Model) for backbone alignment, and use it to align remaining sequences

We use HMMER (Eddy, HHMI) for this purpose

1: Build Hidden Markov Model (HMM) for backbone MSA
2: Use HMM to align remaining sequences (one by one)
3: Use transitivity to infer MSA on entire dataset



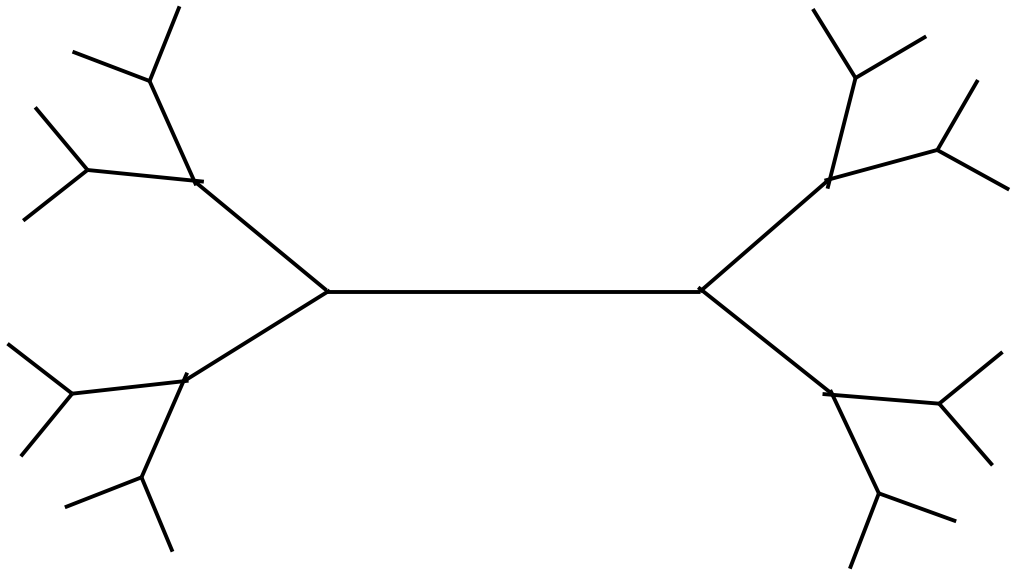
Using HMMER

Using HMMER works well…

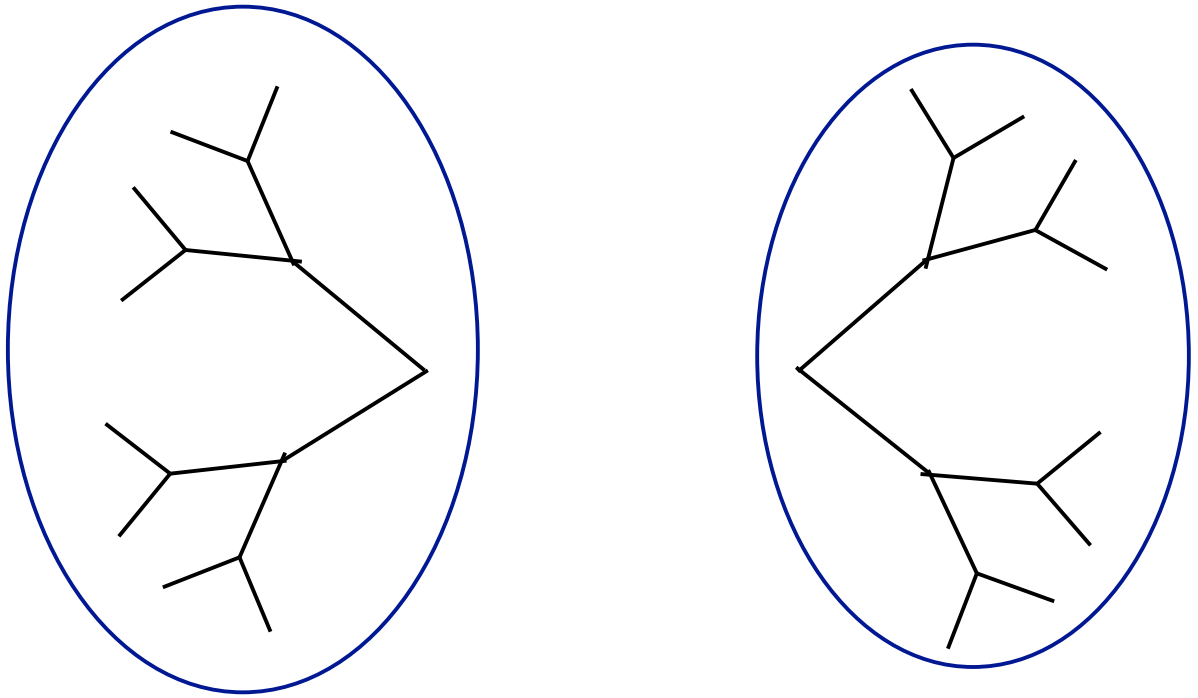
Using HMMER

Using HMMER works well…except when the dataset is big!

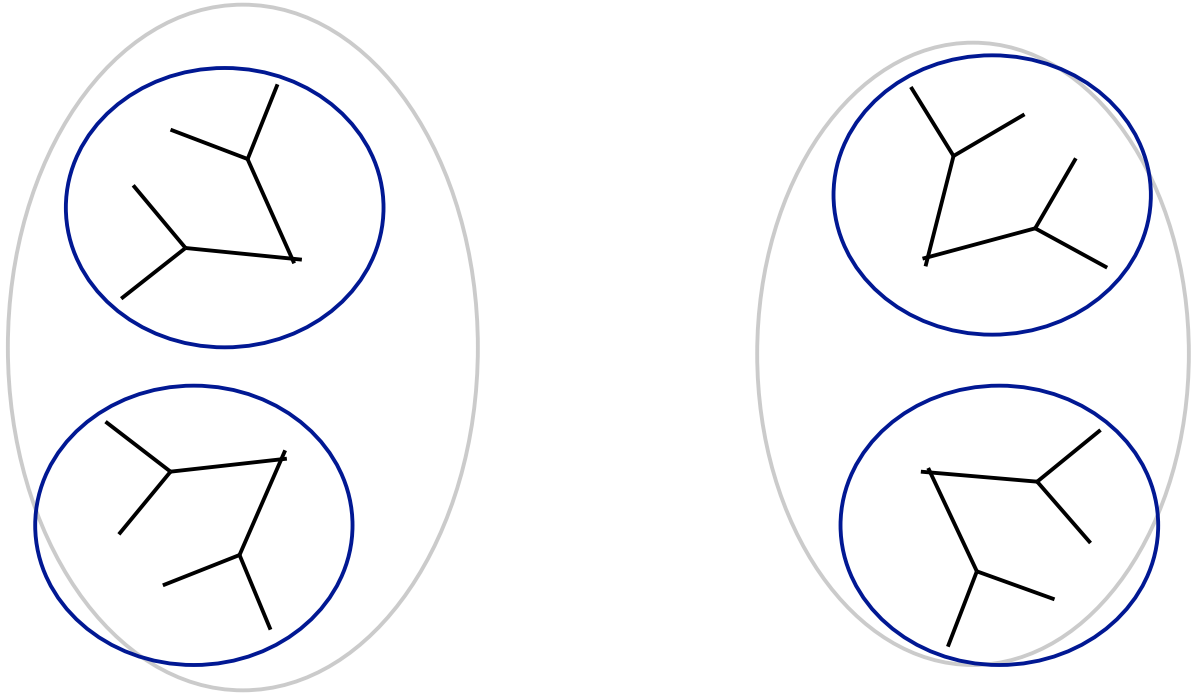
One Hidden Markov Model for the entire alignment?



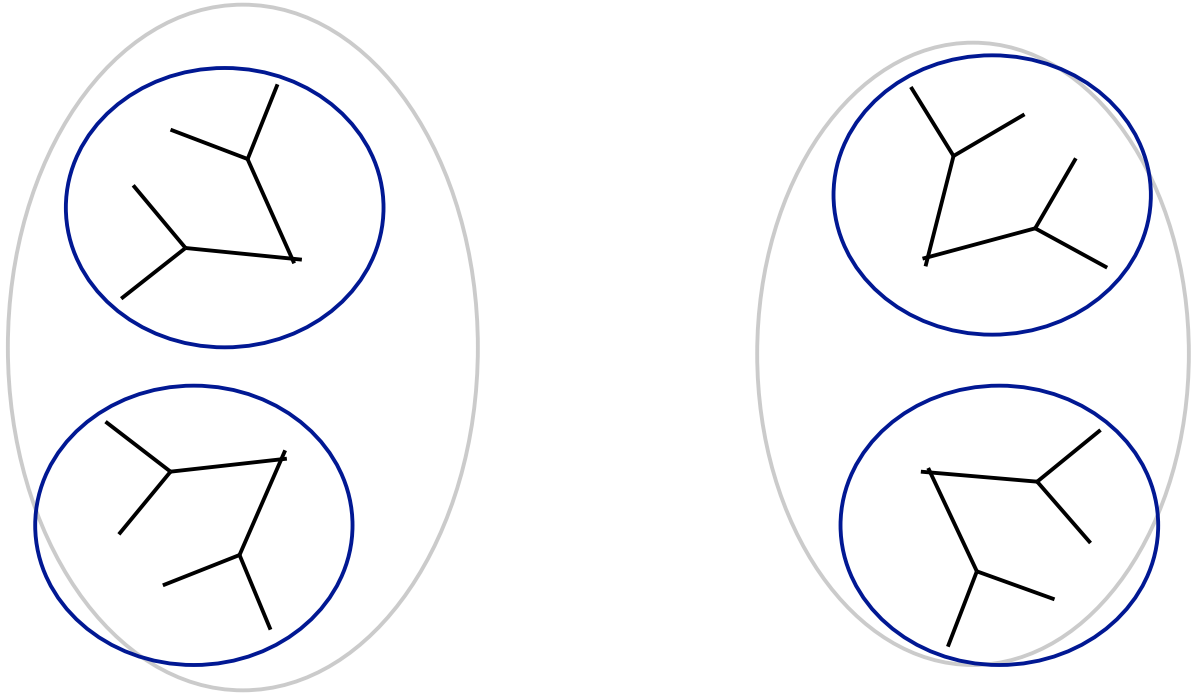
Or 2 HMMs?



Or 4 HMMs?
(or 8 or 16 or…)



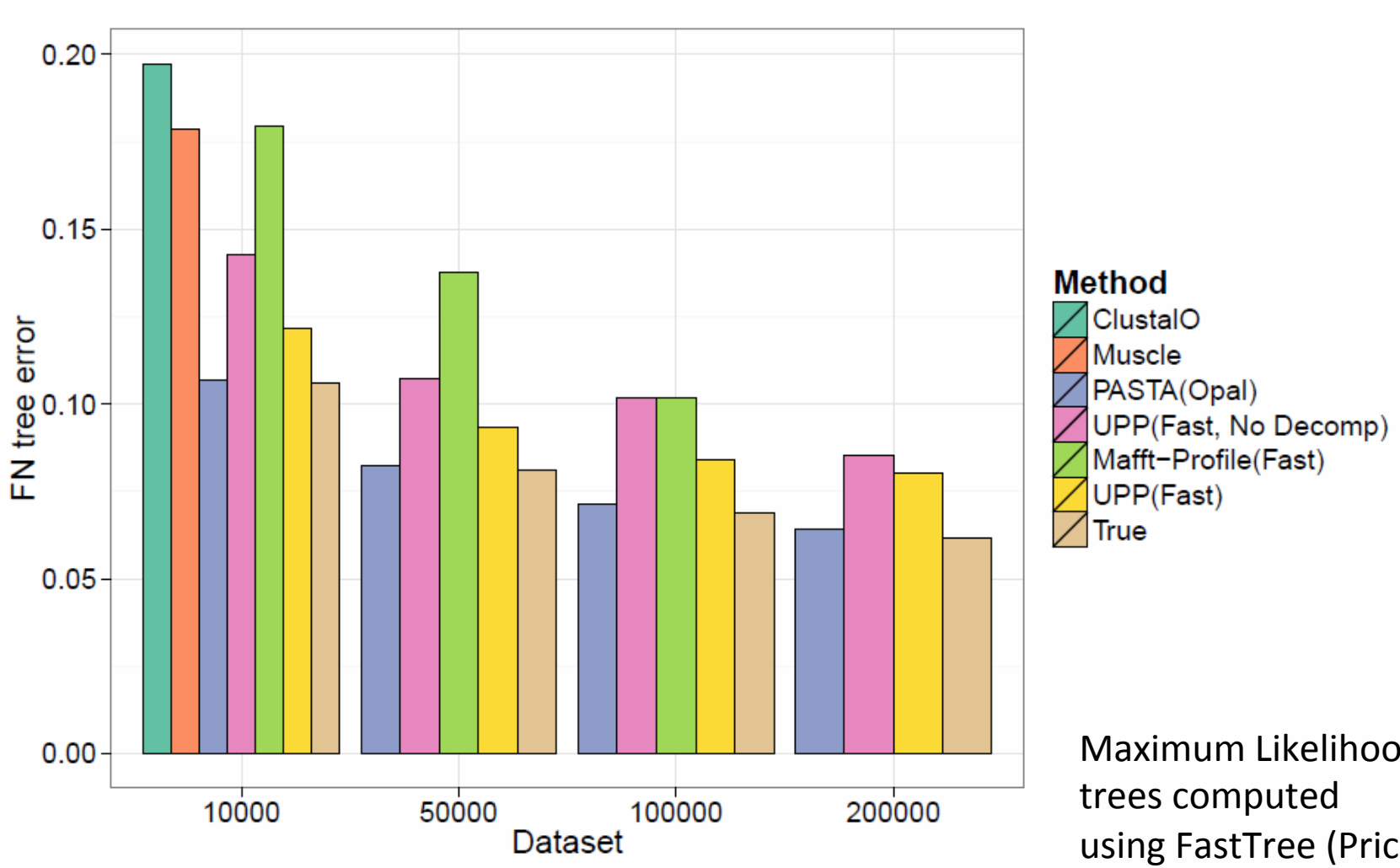
Or what about many different HMMs (HMMs on subsets of size 10, 20, 40, …, n)?



Evaluation

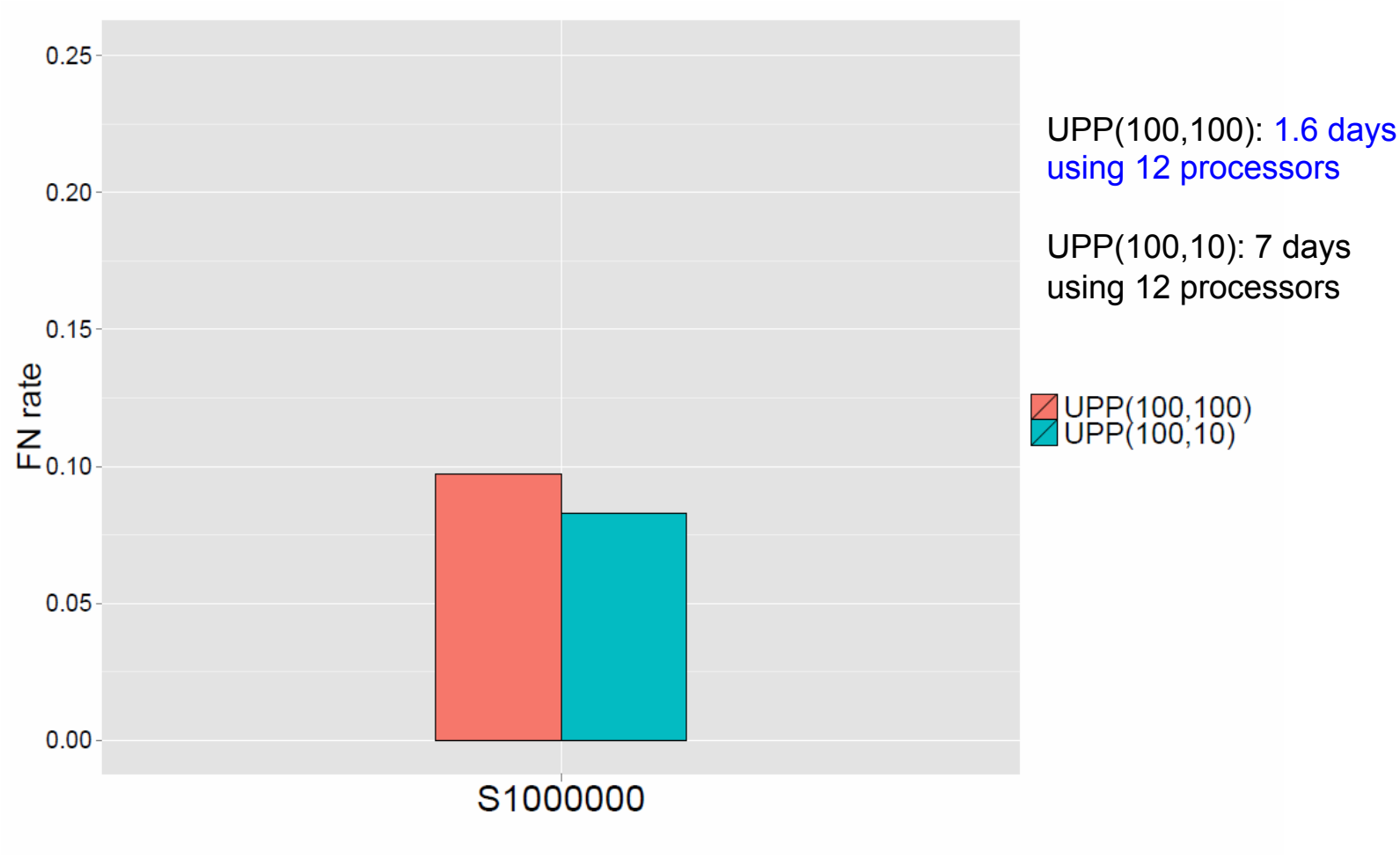
- Simulated datasets (some have fragmentary sequences):
  - 10K to 1,000,000 sequences in RNASim (Sheng Guo, Li-San Wang, and Junhyong Kim, arxiv)
  - 1000-sequence nucleotide datasets from SATe papers
  - 5000-sequence AA datasets (from FastTree paper)
  - 10,000-sequence Indelible nucleotide simulation
- Biological datasets:
  - Proteins: largest BaliBASE and HomFam
  - RNA: 3 CRW (Gutell) datasets up to 28,000 sequences

RNASim: Tree error



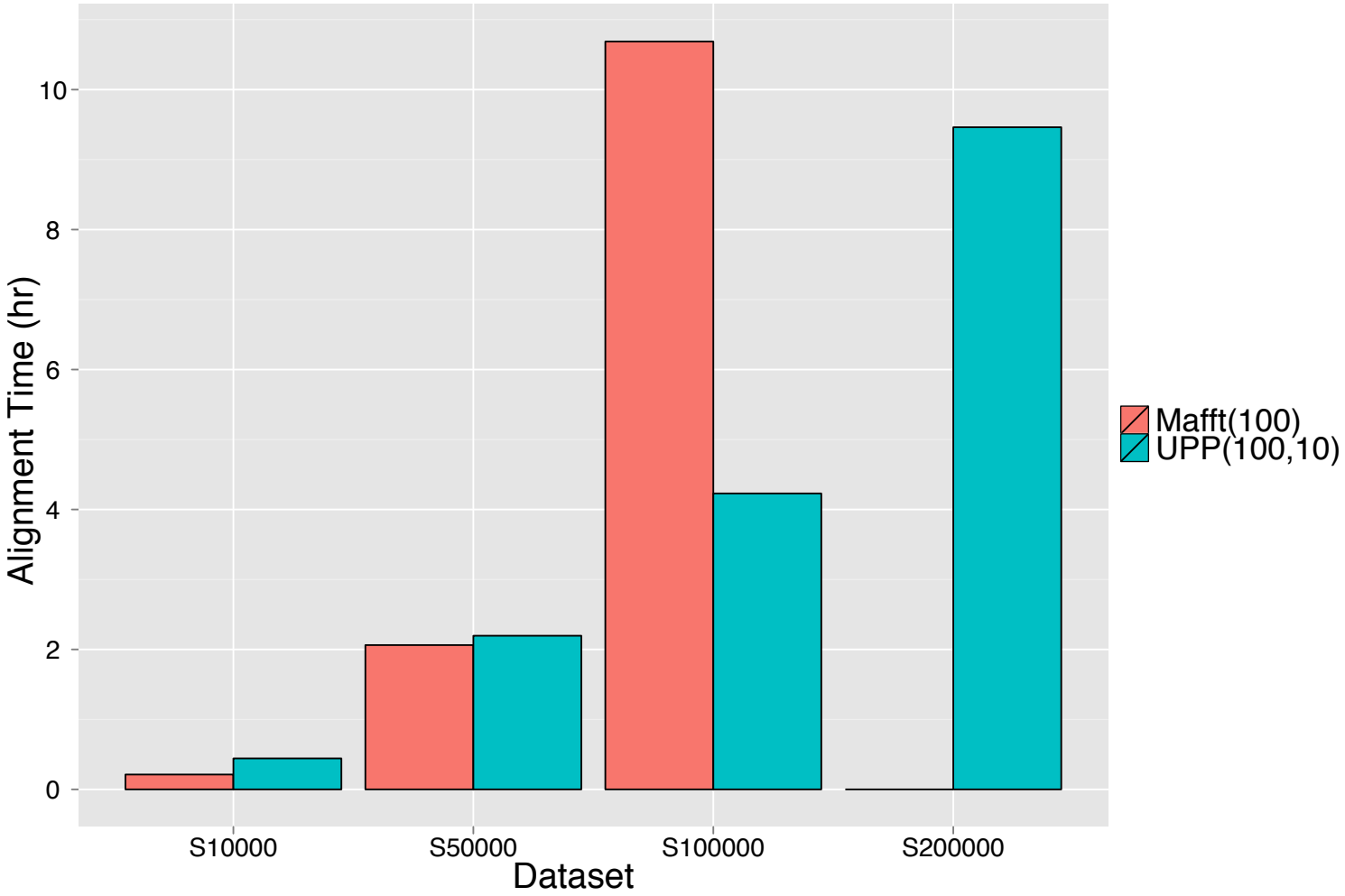
Maximum Likelihood trees computed using FastTree (Price)

One Million Sequences: Tree Error

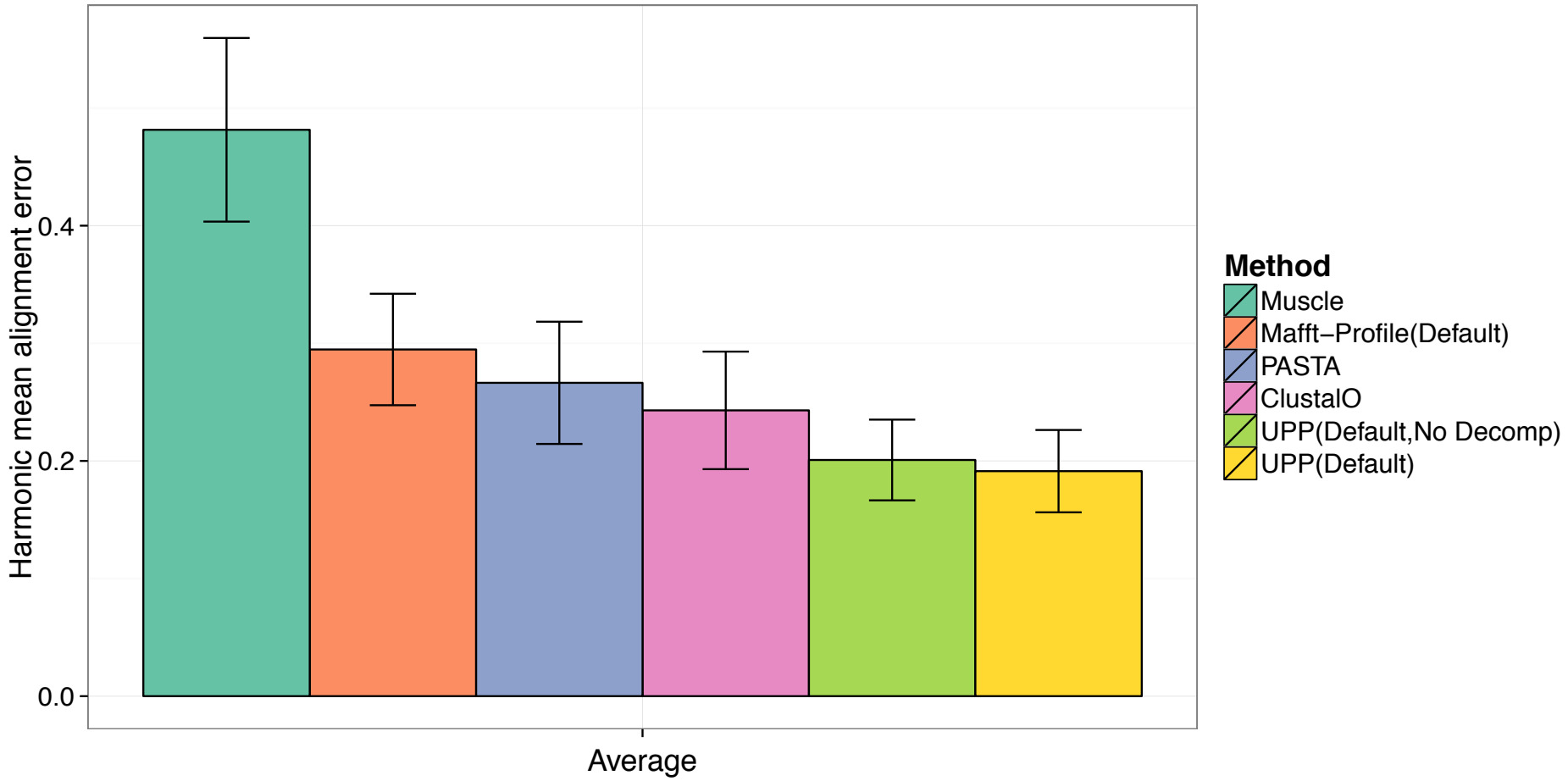


Note: UPP Decomposition improves accuracy

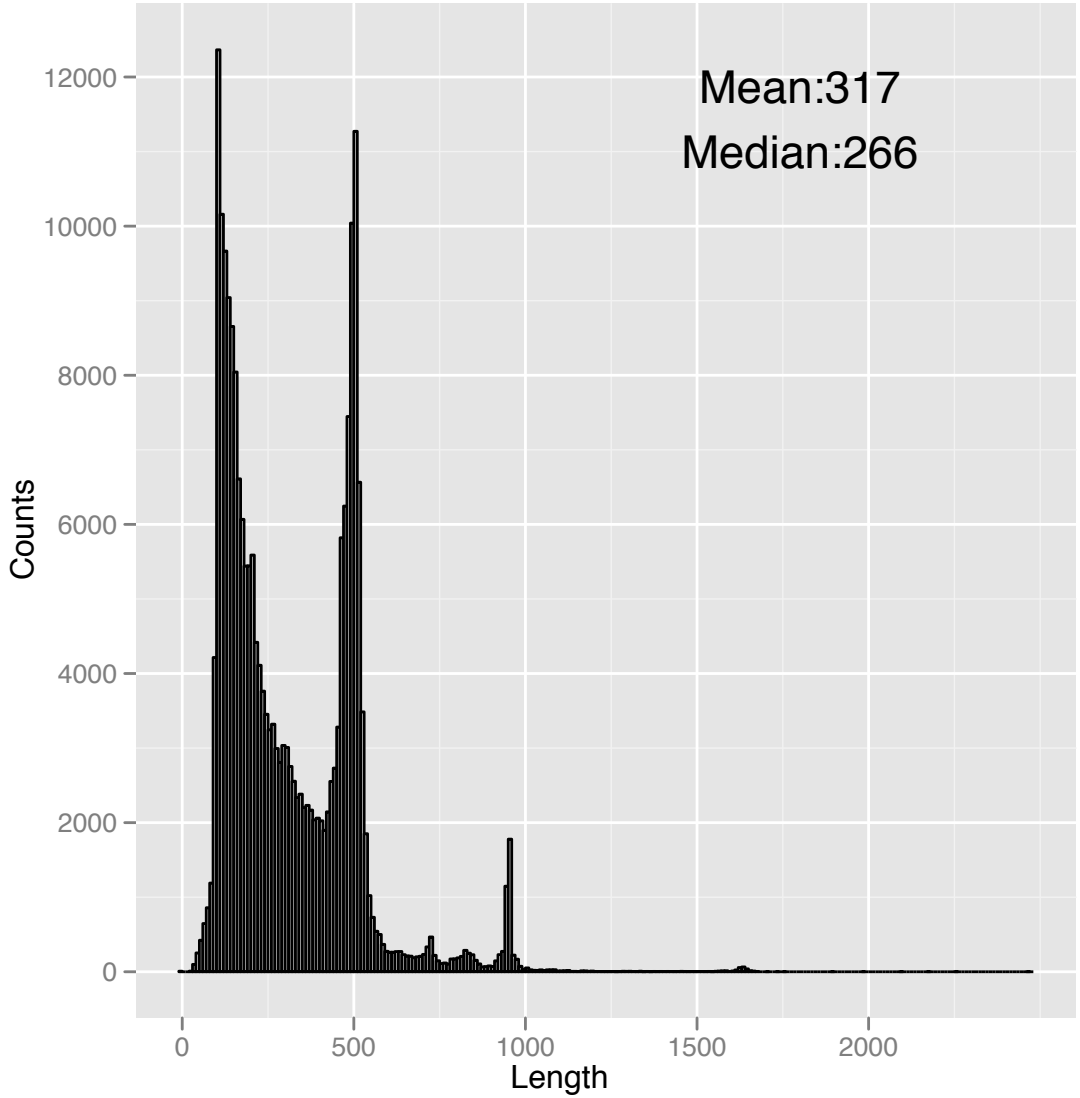
UPP vs. MAFFT-profile Running Time

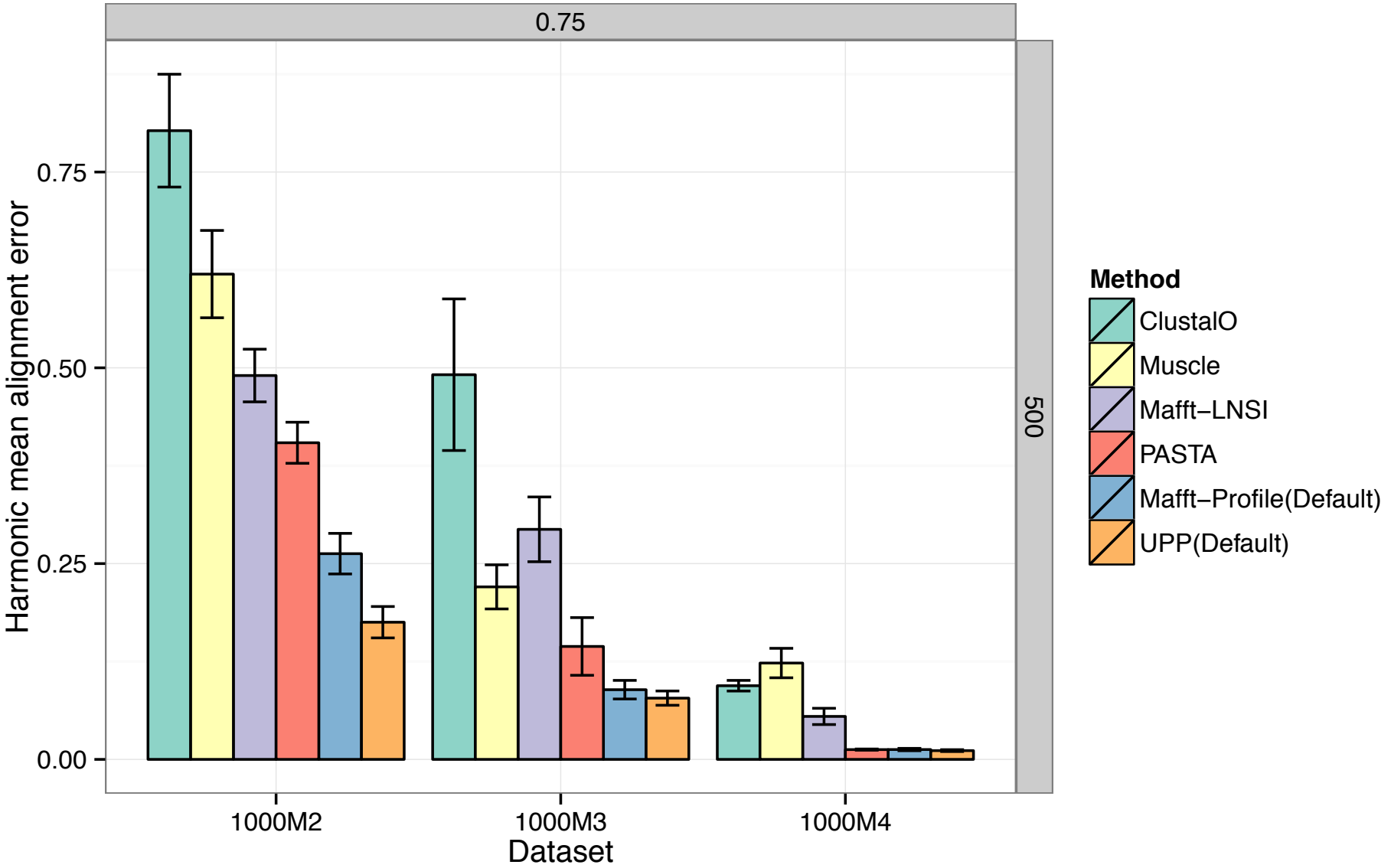


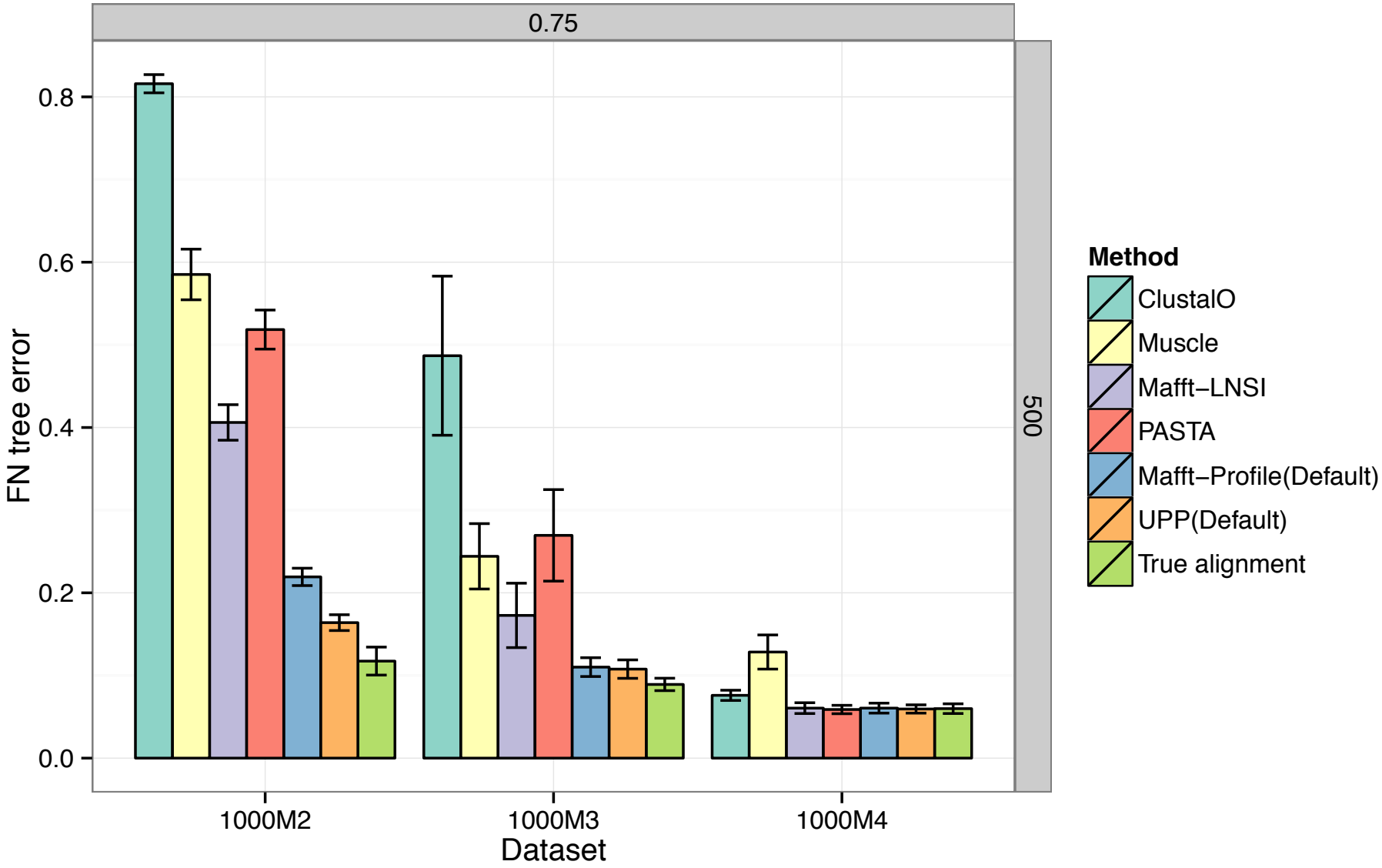
AA Sequence Alignment Error (18 HomFam datasets)



1KP Cytochrome dataset sequence length distribution (>100K seqs)




0.75
500
Harmonic mean alignment error
0.00
0.25
0.50
0.75
1000M2
1000M3
1000M4
Dataset
Method
ClustalO
Muscle
Mafft−LNSI
PASTA
Mafft−Profile(Default)
UPP(Default)


0.75
500
FN tree error
0.0
0.2
0.4
0.6
0.8
1000M2
1000M3
1000M4
Dataset
Method
ClustalO
Muscle
Mafft−LNSI
PASTA
Mafft−Profile(Default)
UPP(Default)
True alignment

MSA and Phylogeny Estimation

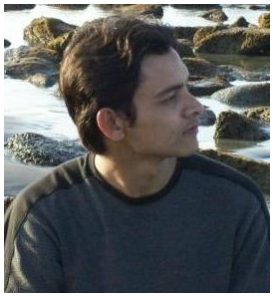
- MSA estimation impacts phylogenetic accuracy, and many datasets are hard to align:
  - Large datasets
  - Datasets with high rates of evolution
  - Datasets with fragmentary sequences
- Co-estimation methods (e.g., Bali-Phy) can be excellent but do not run on datasets with more than about 100 sequences.
- Other methods can be very good for small enough datasets with low enough rates of evolution (e.g., Prank and MAFFT)
- SATé (Liu et al., Science 2009 and Systematic Biology 2012) provides very good alignments and trees even for high rates of evolution and large datasets
- PASTA (RECOMB 2014) is a direct improvement on SATé (matches or improves accuracy on small datasets, better accuracy on large datasets, can analyze very large datasets – up to 200,000 sequences so far)
- UPP – different approach, highly robust to fragmentary sequences, similar to PASTA in accuracy but can analyze even larger datasets, highly parallelizable. (Not yet available)

SEPP, TIPP, and UPP

- SEPP: SATe-enabled Phylogenetic Placement (Mirarab, Nugyen and Warnow, PSB 2012)
- TIPP: Taxon identification and phylogenetic profiling (Nguyen, Mirarab, Liu, Pop, and Warnow, submitted)
- UPP: Ultra-large multiple sequence alignment using SEPP (Nguyen, Mirarab, Kumar, Wang, Guo, Kim, and Warnow, in preparation)

Warnow Laboratory





PhD students: Siavash Mirarab*, Nam Nguyen, and Md. S. Bayzid**
Undergrad: Keerthana Kumar
Lab Website: http://www.cs.utexas.edu/users/phylo

Funding: Guggenheim Foundation, Packard, NSF, Microsoft Research New England, David Bruton Jr. Centennial Professorship, TACC (Texas Advanced Computing Center), and the University of Alberta (Canada)

TACC and UTCS computational resources

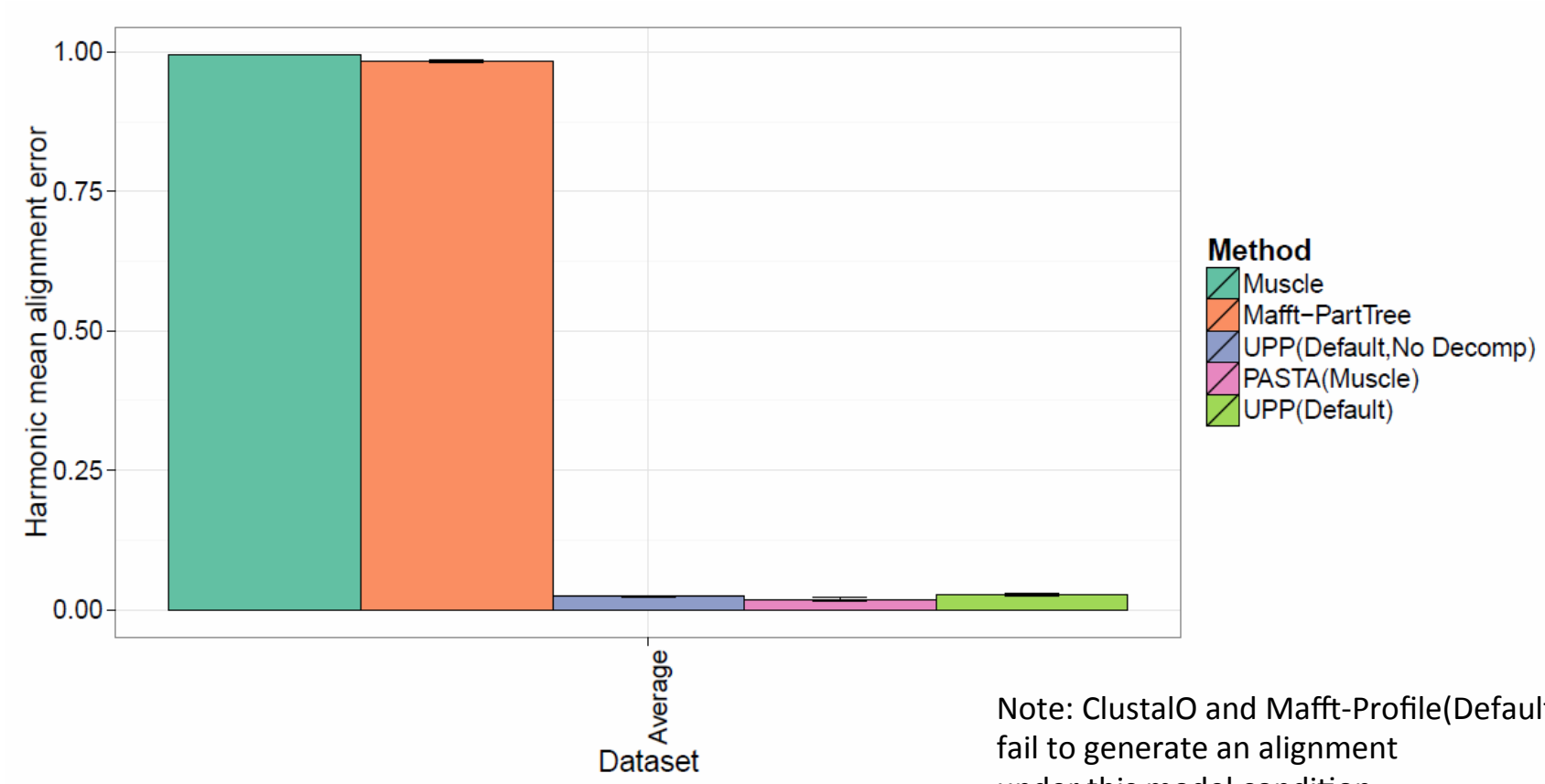
* Supported by HHMI Predoctoral Fellowship

** Supported by Fulbright Foundation Predoctoral Fellowship

Future Work

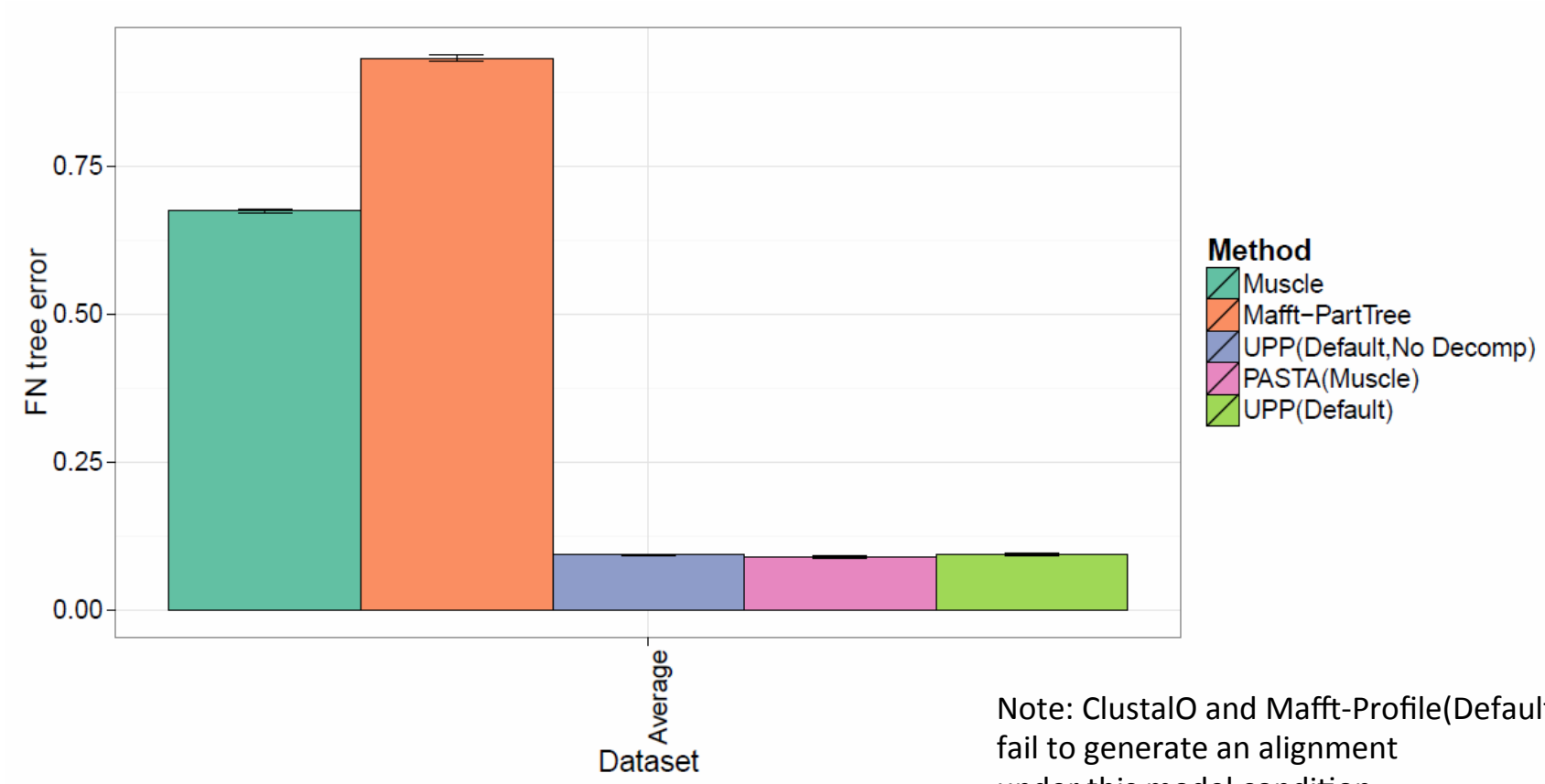
- Extending TIPP to non-marker genes
- Using the new HMM Family technique in SEPP and UPP
- Using external seed alignments in SEPP, TIPP, and UPP
- Boosting statistical co-estimation methods by using them in UPP for the backbone alignment and tree

Indelible 10K: Alignment error



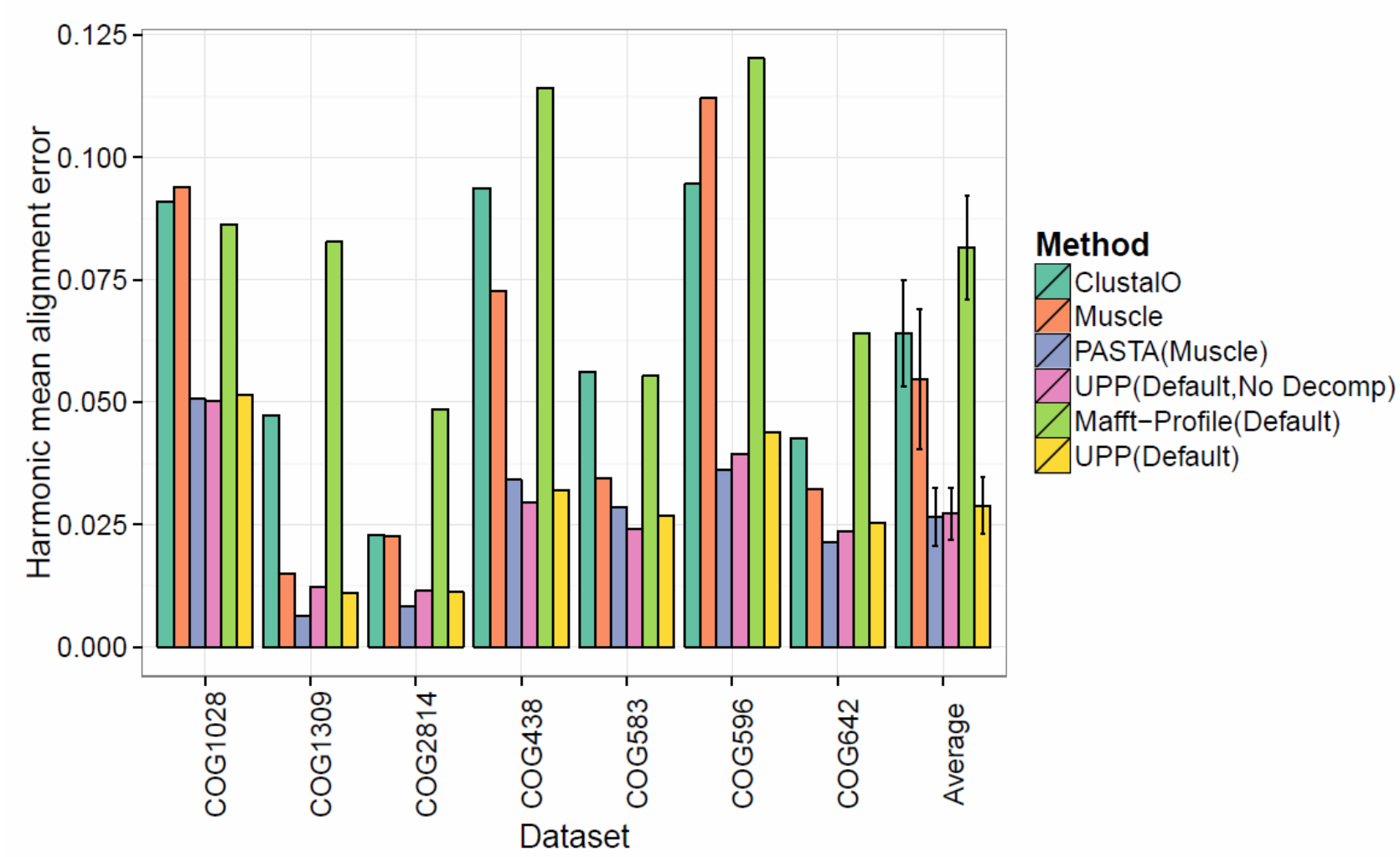
Note: ClustalO and Mafft-Profile(Default) fail to generate an alignment under this model condition.

Indelible 10K : Tree error

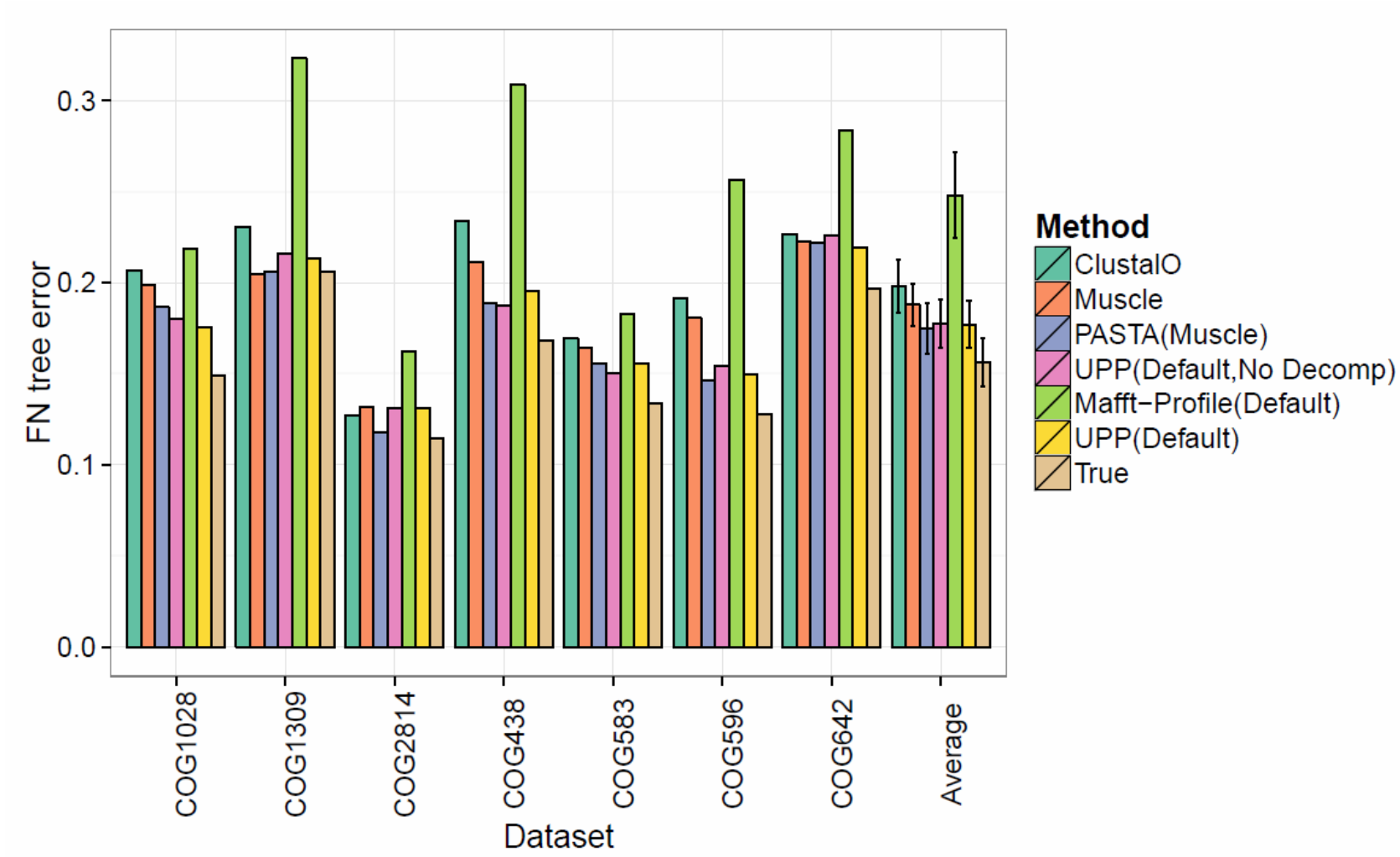


Note: ClustalO and Mafft-Profile(Default) fail to generate an alignment under this model condition.

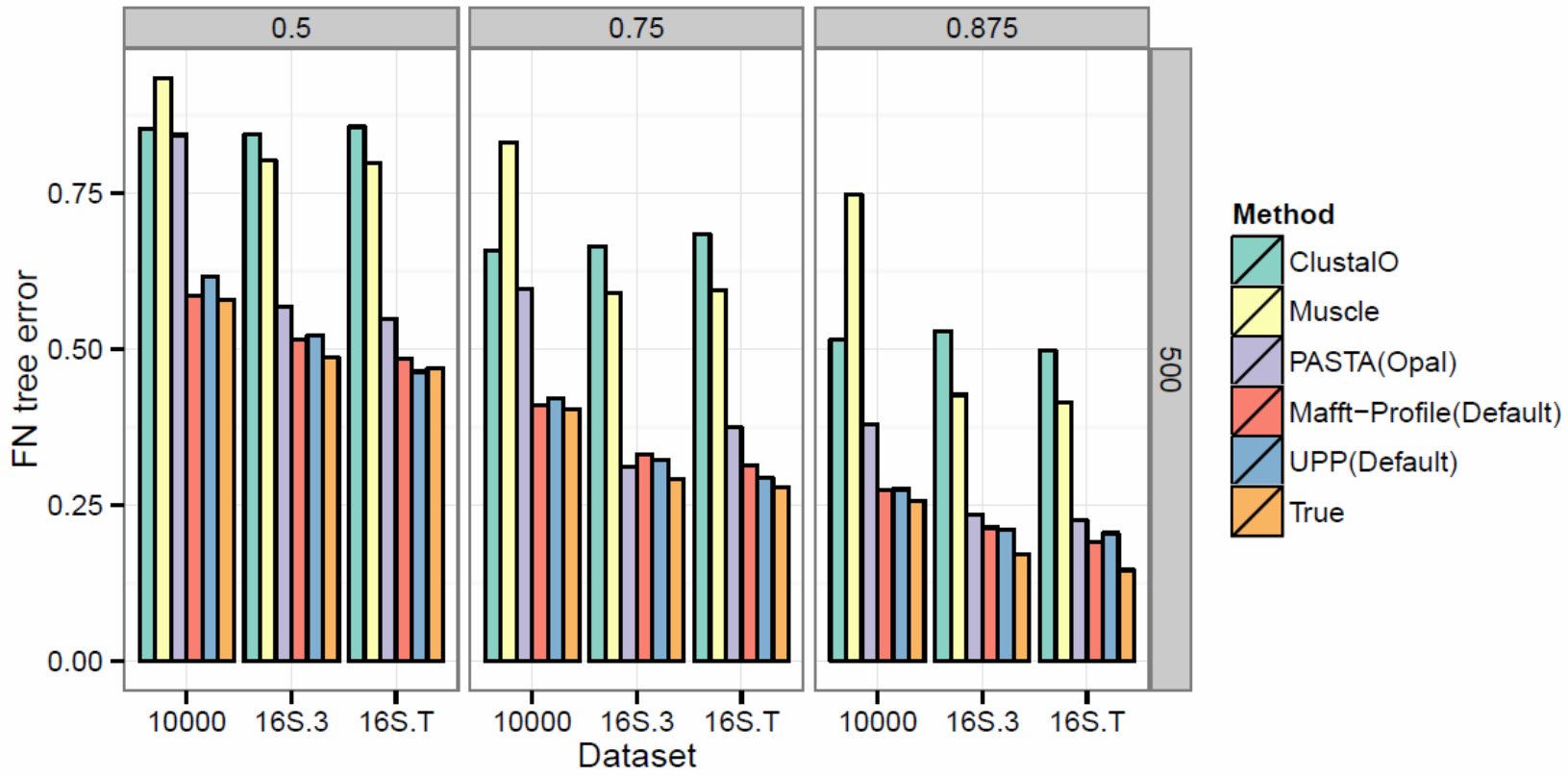
FastTree AA: Alignment error



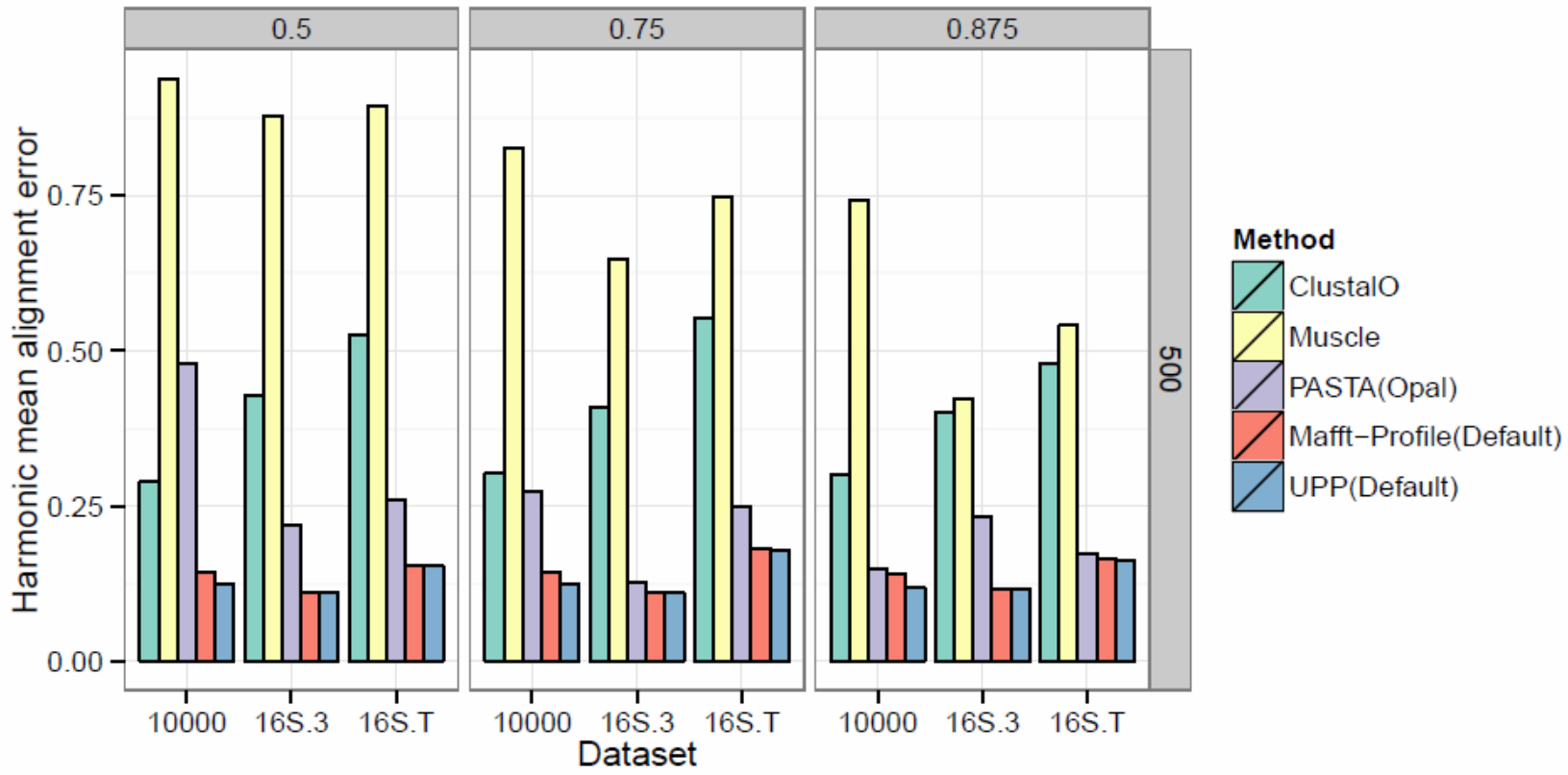
FastTree AA: Tree error



Large fragmentary: Tree error



Large fragmentary: Alignment error



10000-sequence dataset from RNASim
16S.3 and 16S.T from Gutell's Comparative Ribosomal Website (CRW)