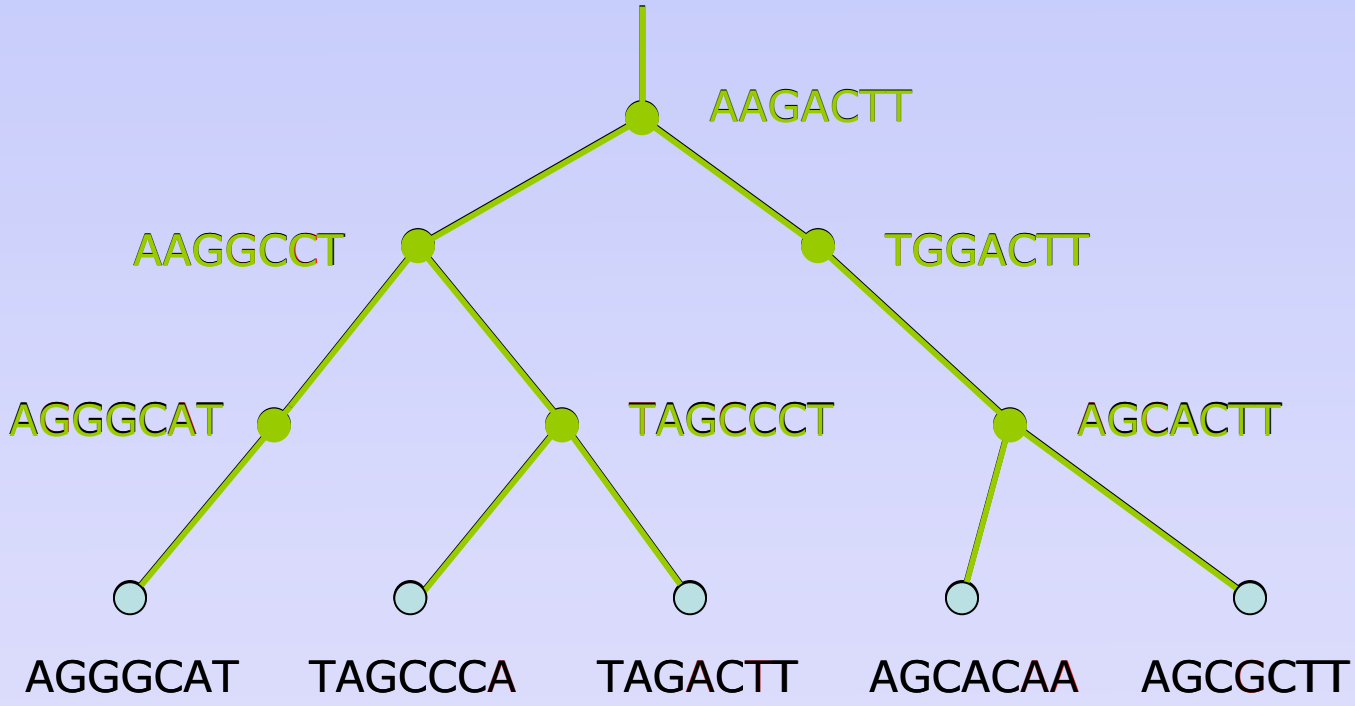
Fast and accurate large-scale co-estimation of alignments and trees

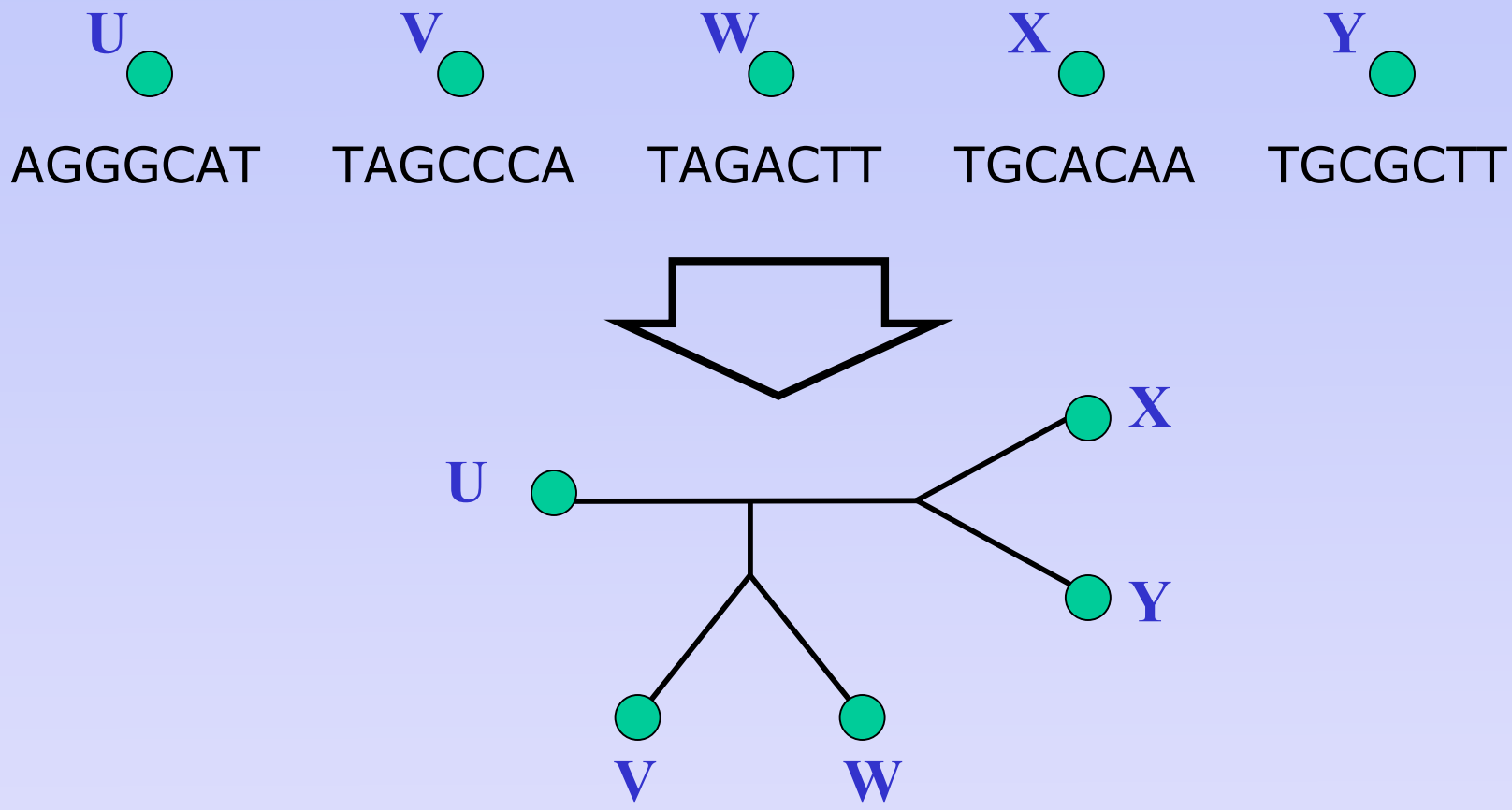
Tandy Warnow

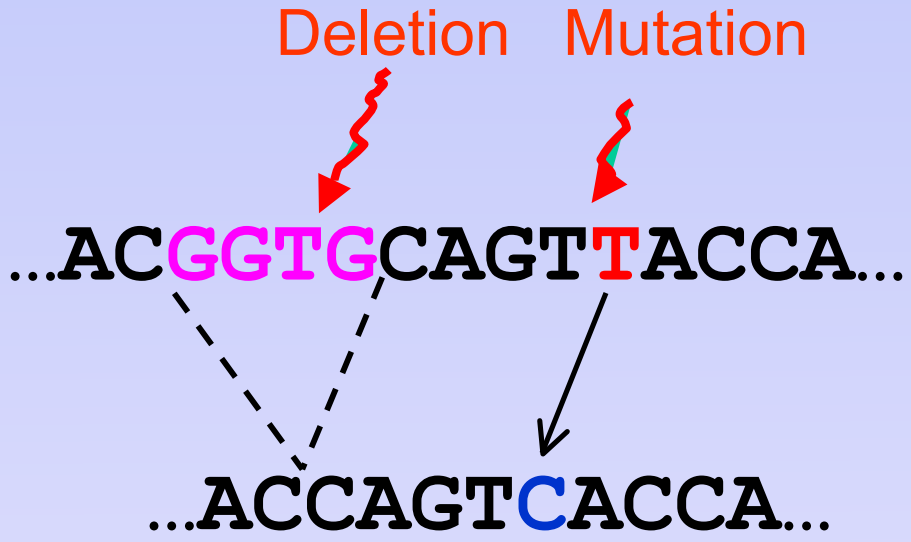
The University of Texas at Austin

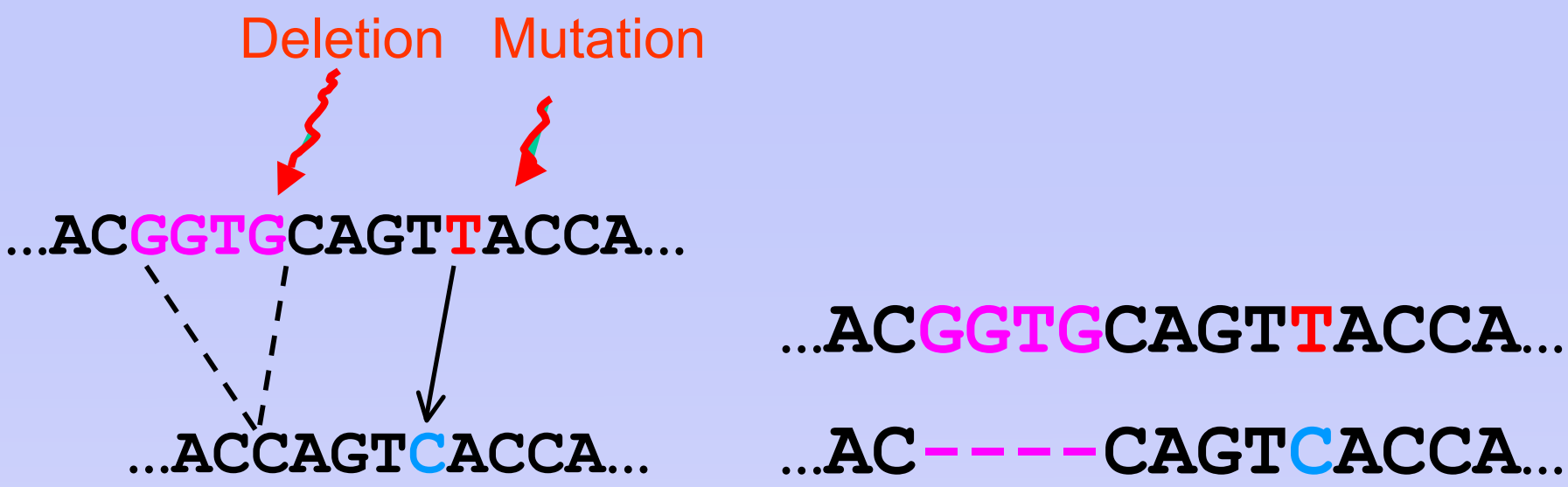
Joint work with K. Liu, S. Raghavan, S. Nelesen, and C.R. Linder

DNA Sequence Evolution




U
V
W
X
Y
AGGGCAT
TAGCCCA
TAGACTT
TGCACAA
TGCGCTT
X
U
Y
V
W


Deletion
Mutation
...ACGGTGCAGTTACCA...
...ACCAGTCACCA...



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: Multiple Sequence Alignment

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



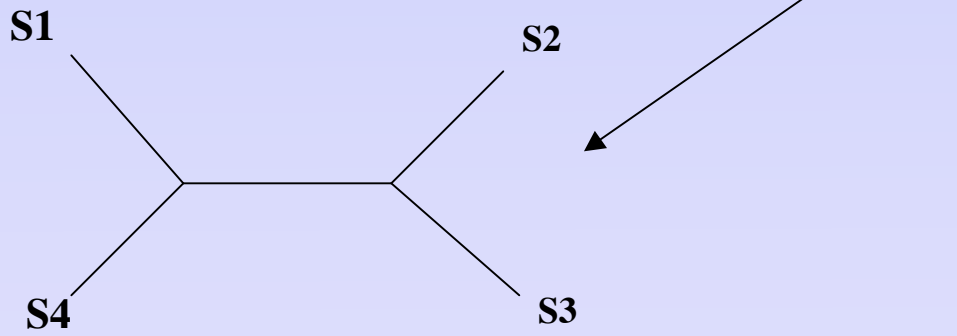
```
S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAC--GACCGC--
S3 = TAG-CT-------GACCGC--
S4 = -------TCAC--GACCGACA
```

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
```



Many methods

Alignment methods

- Clustal
- POY (and POY*)
- Probcons (and Probtree)
- MAFFT
- Prank
- Muscle
- Di-align
- T-Coffee
- Opal
- Etc.

Phylogeny methods

- Bayesian MCMC
- Maximum parsimony
- Maximum likelihood
- Neighbor joining
- FastME
- UPGMA
- Quartet puzzling
- Etc.

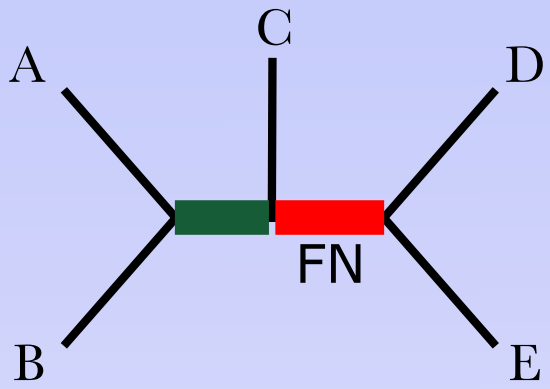
***RAxML**: best heuristic for large-scale ML optimization*

Question: How well do two-phase methods perform?

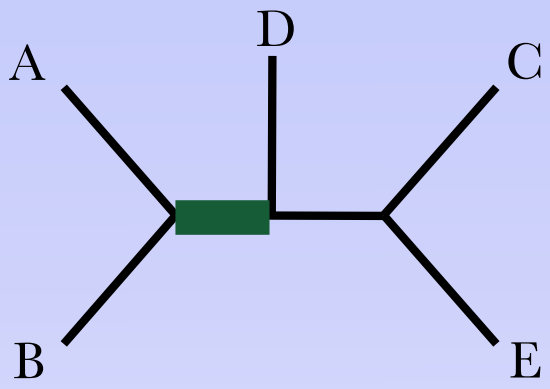
- ROSE simulation:
  - 1000, 500, and 100 sequences
  - Evolution with substitutions and indels
  - Varied gap lengths, rates of evolution
- Estimated alignments using leading methods
- Used RAxML to compute trees
- Recorded tree error (missing branch rate)
- Recorded alignment error

Liu et al., *Science* 2009

Quantifying Error

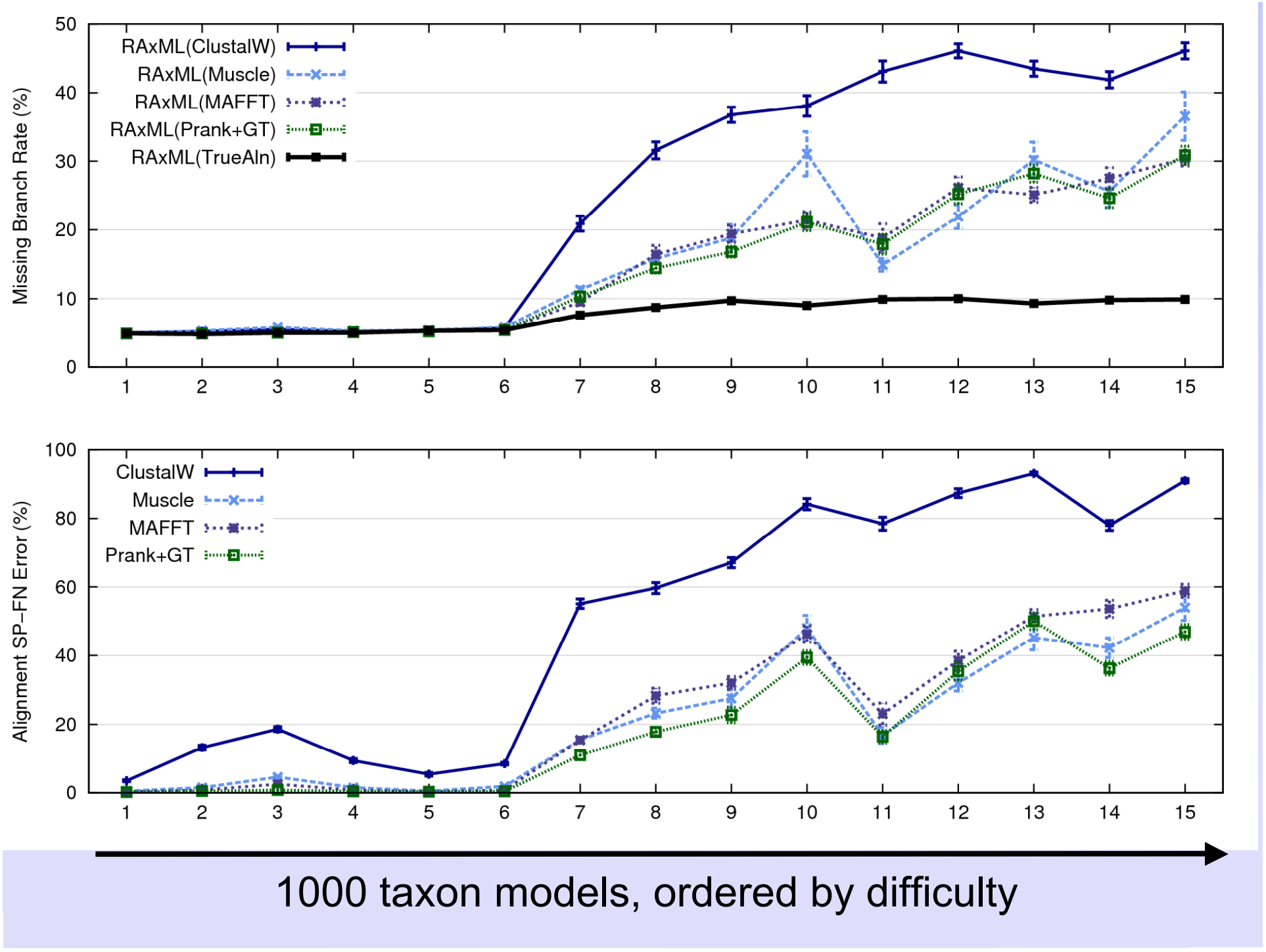


True Tree



Estimated Tree

False negative (FN) - aka “missing branch” :
An edge in the true tree missing from the estimated tree

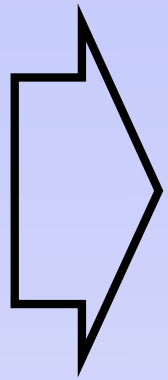

Missing Branch Rate (%)
RAxML(ClustalW)
RAxML(Muscle)
RAxML(MAFFT)
RAxML(Prank+GT)
RAxML(TrueAln)
Alignment SP–FN Error (%)
ClustalW
Muscle
MAFFT
Prank+GT
1000 taxon models, ordered by difficulty

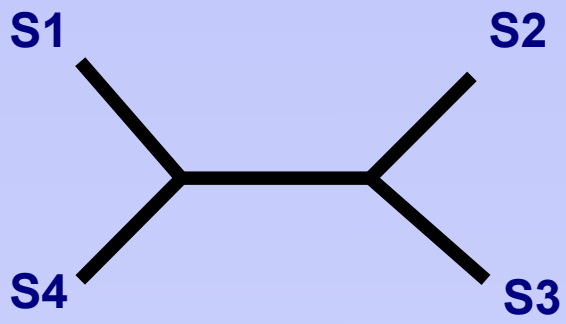
Problems with the two-phase approach

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

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

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





and

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

Statistical simultaneous estimation methods (BALiPhy, Alifritz, Statalign) are not scalable.

POY and related methods are not more accurate than standard two-phase methods.

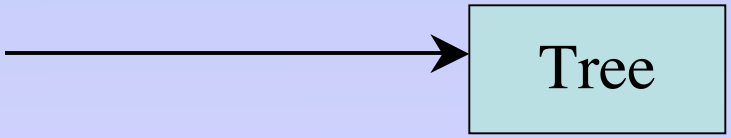
SATé:
(Simultaneous Alignment and Tree Estimation)

- Liu, Nelesen, Raghavan, Linder, and Warnow
- Strategy: iterate between alignments and trees, *re-aligning sequences on each tree using a novel divide-and-conquer approach*, and computing ML trees on the new alignments.
- Optimization criterion: alignment/tree pair that optimizes maximum likelihood under GTR+Gamma (RAxML GTRMIX, treating gaps as missing data).

- *Science*, 19 June 2009, pp. 1561-1564.

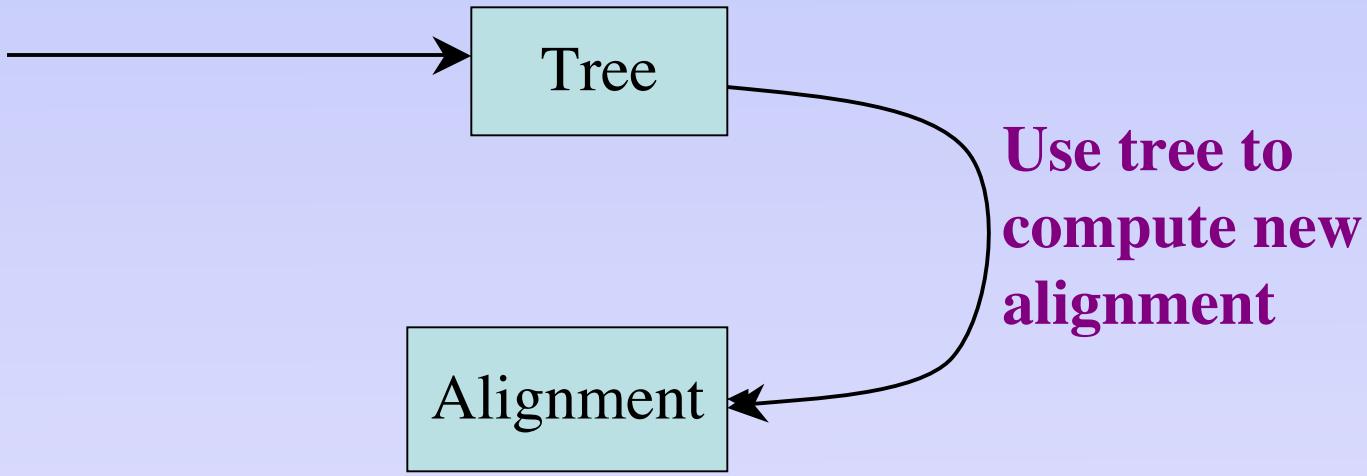
SATé Algorithm

Obtain initial alignment
and estimated ML tree



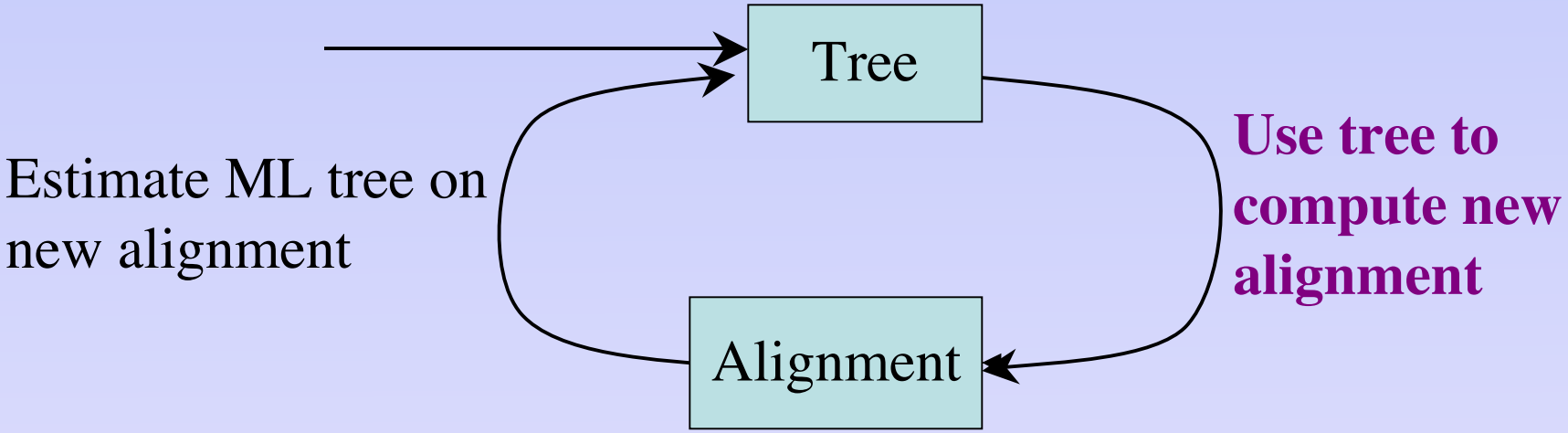
SATé Algorithm

Obtain initial alignment
and estimated ML tree



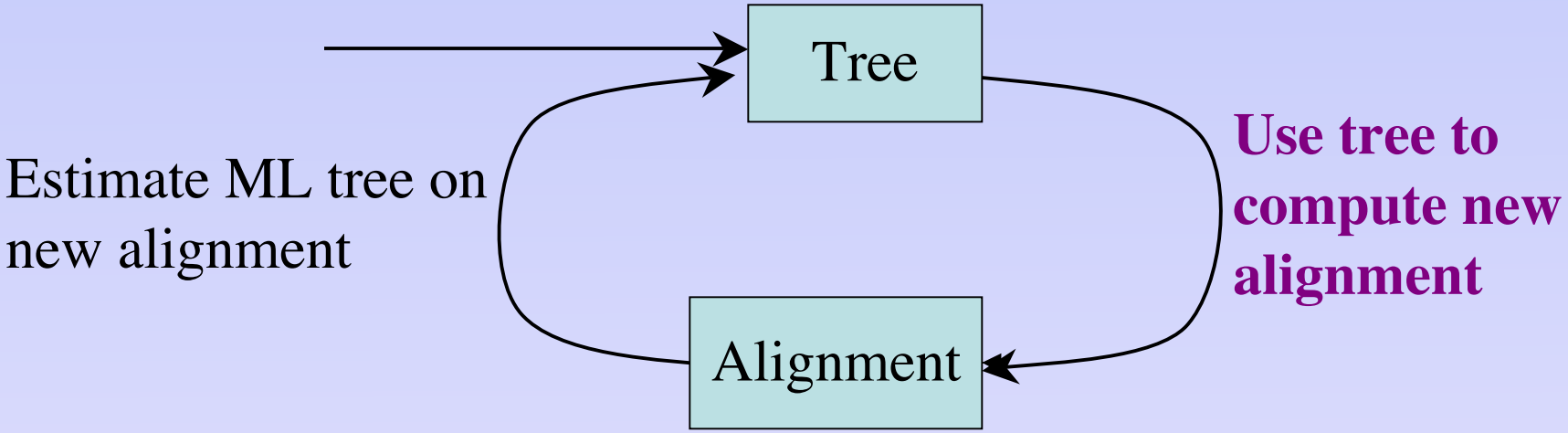
SATé Algorithm

Obtain initial alignment and estimated ML tree



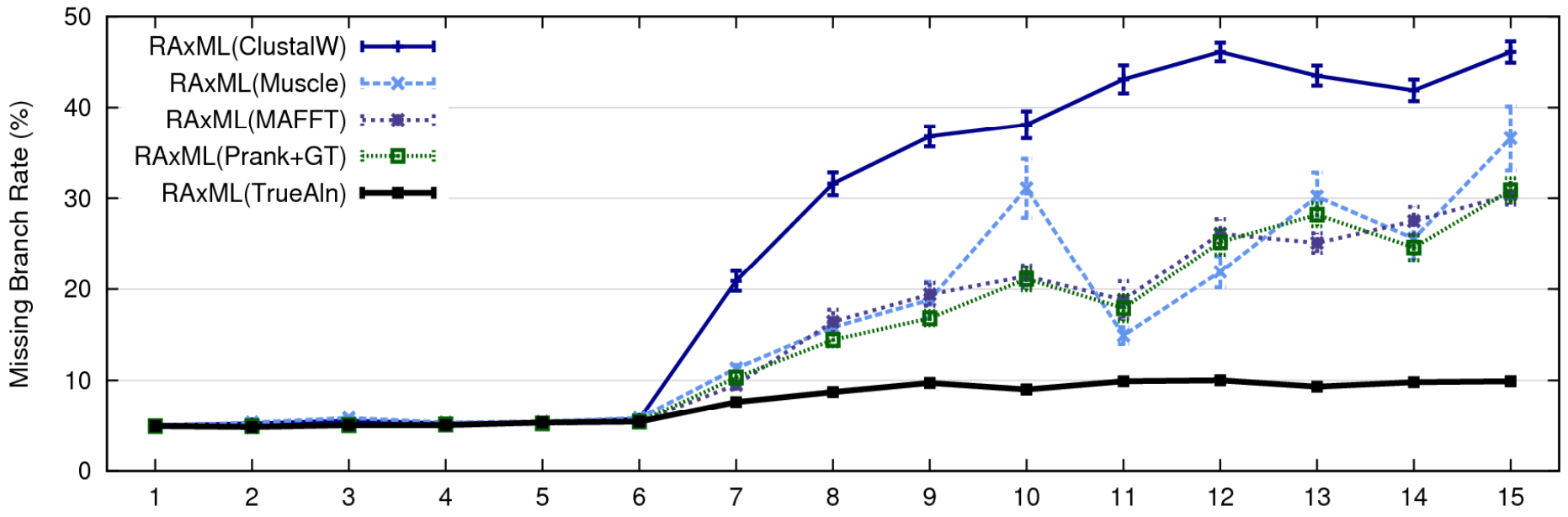
SATé Algorithm

Obtain initial alignment
and estimated ML tree

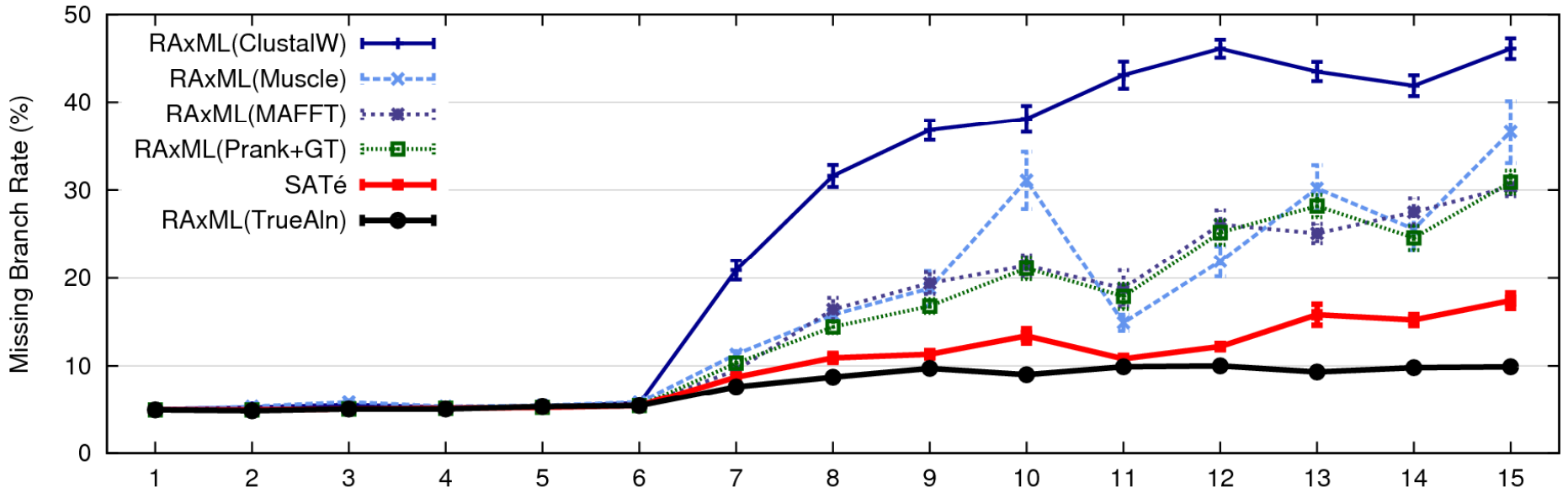


If new alignment/tree pair has worse ML score, realign using a different decomposition

Repeat until termination condition (typically, 24 hours)



1000 taxon models, ordered by difficulty



1000 taxon models, ordered by difficulty

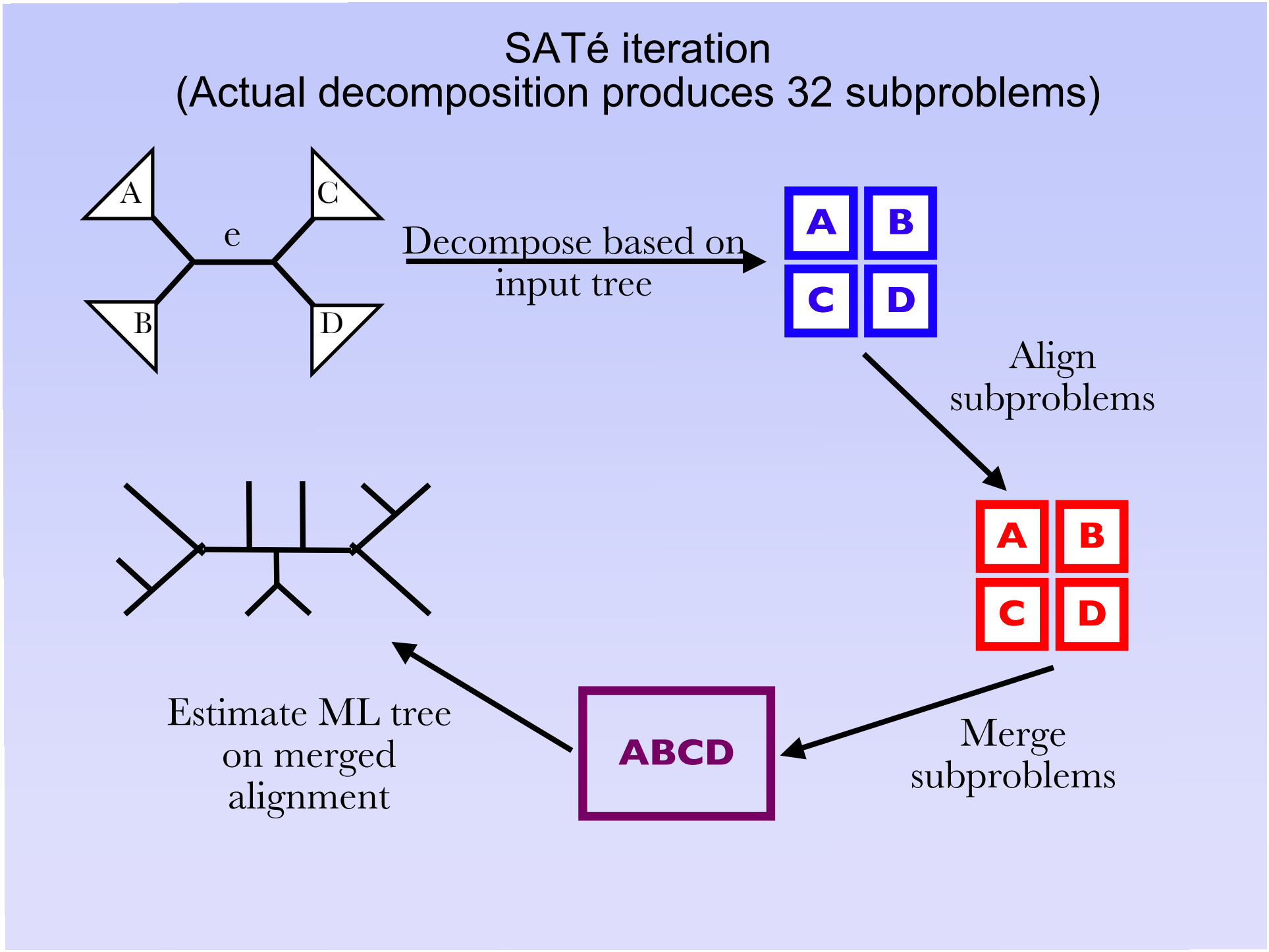
24 hour SATé analysis, on desktop machines, run sequentially (no parallelism)

Similar improvements for biological datasets

Why is SATé so accurate?

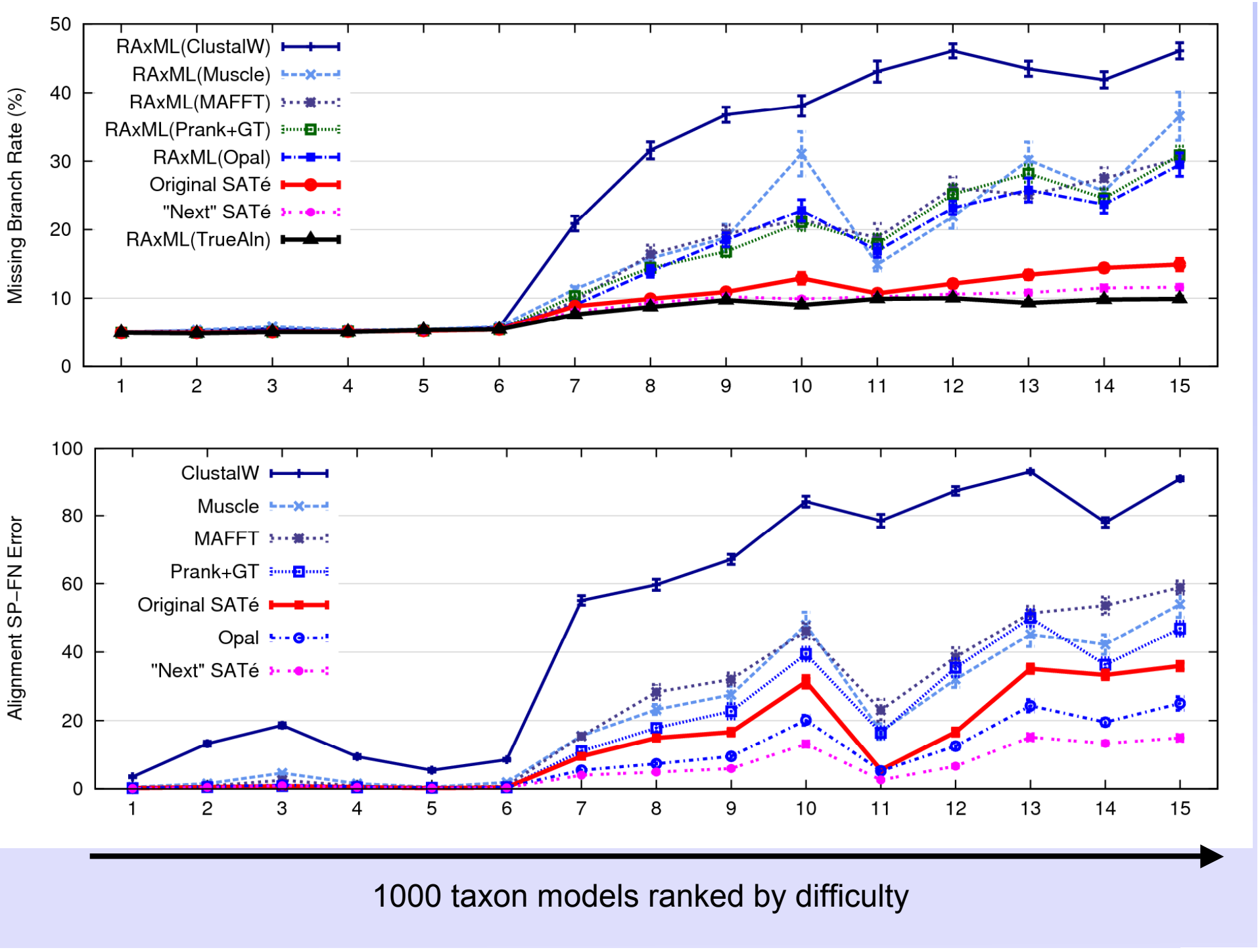
- *Not because it's good to optimize ML* under a model treating gaps as missing data - we prove that for this optimization problem, *all trees are optimal!*

- Instead, the key is the *divide-and-conquer* strategy (with ML giving a small but statistically significant additional improvement).


SATé iteration
(Actual decomposition produces 32 subproblems)
A
C
e
B
D
Decompose based on input tree
A
B
C
D
Align subproblems
A
B
C
D
Merge subproblems
ABCD
Estimate ML tree on merged alignment

SATé-II

- **SATé-II: different re-alignment strategy**, but same general algorithm design.
- Fast version gives highly accurate results in just a few hours on datasets with 1000 taxa.


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

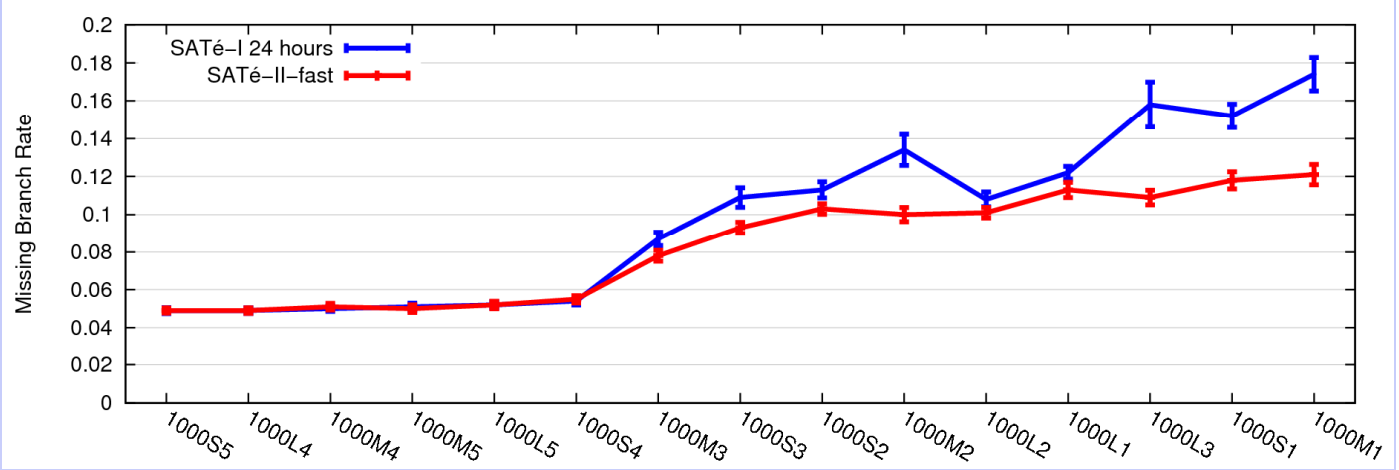
SATé-I

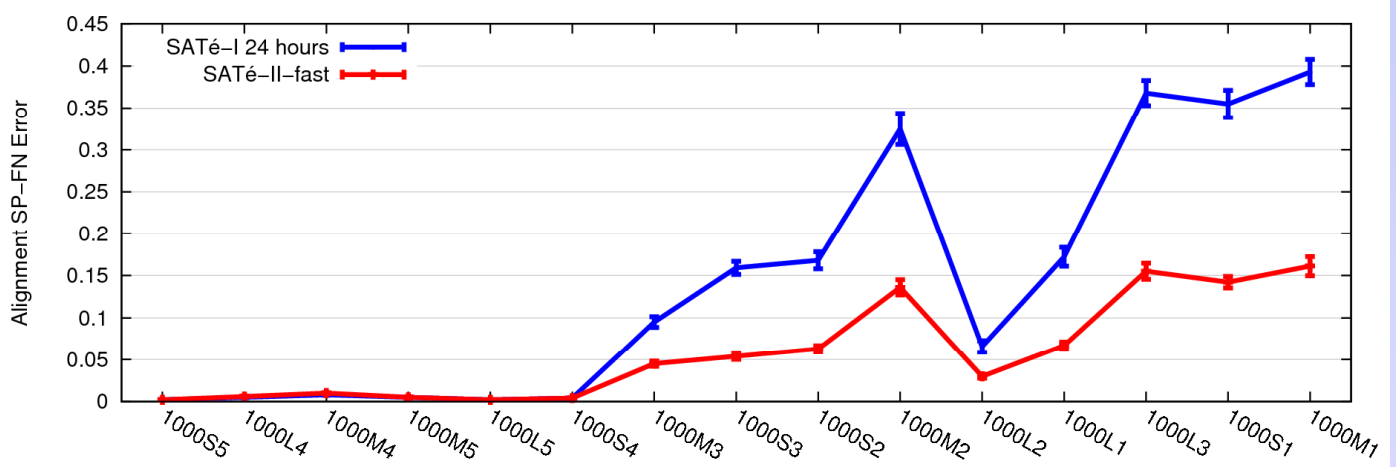
vs.

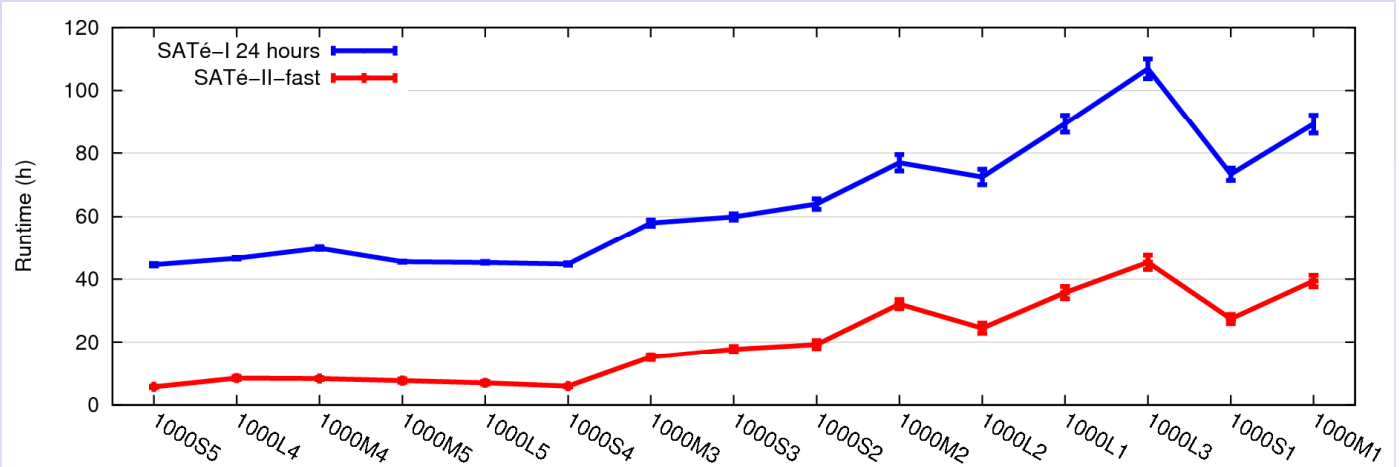
SATé-II

SATé-II

- **Faster** and more accurate than **SATé-I**
- Longer analyses or use of ML to select tree/alignment pair slightly better results







Summary

- SATé-I and SATé-II produce more accurate trees and alignments than standard two-phase methods.
- SATé-II is faster and more accurate than SATé-I.
- Why these methods work well is not the use of ML (treating gaps as missing data).
- Better results would potentially be obtained if statistical co-estimation of alignments and trees, under a model in which indels are events. However, such approaches are not yet feasible for large datasets.

Acknowledgments

- National Science Foundation: Assembling the Tree of Life (ATOL), ITR, and IGERT grants (0733029, 0331453, 0114387)

- Mark Holder and Jiaye Yu (Univ of Kansas) downloadable software, with biologist-friendly GUI, available at

  http://phylo.bio.ku.edu/software/sate/sate.html