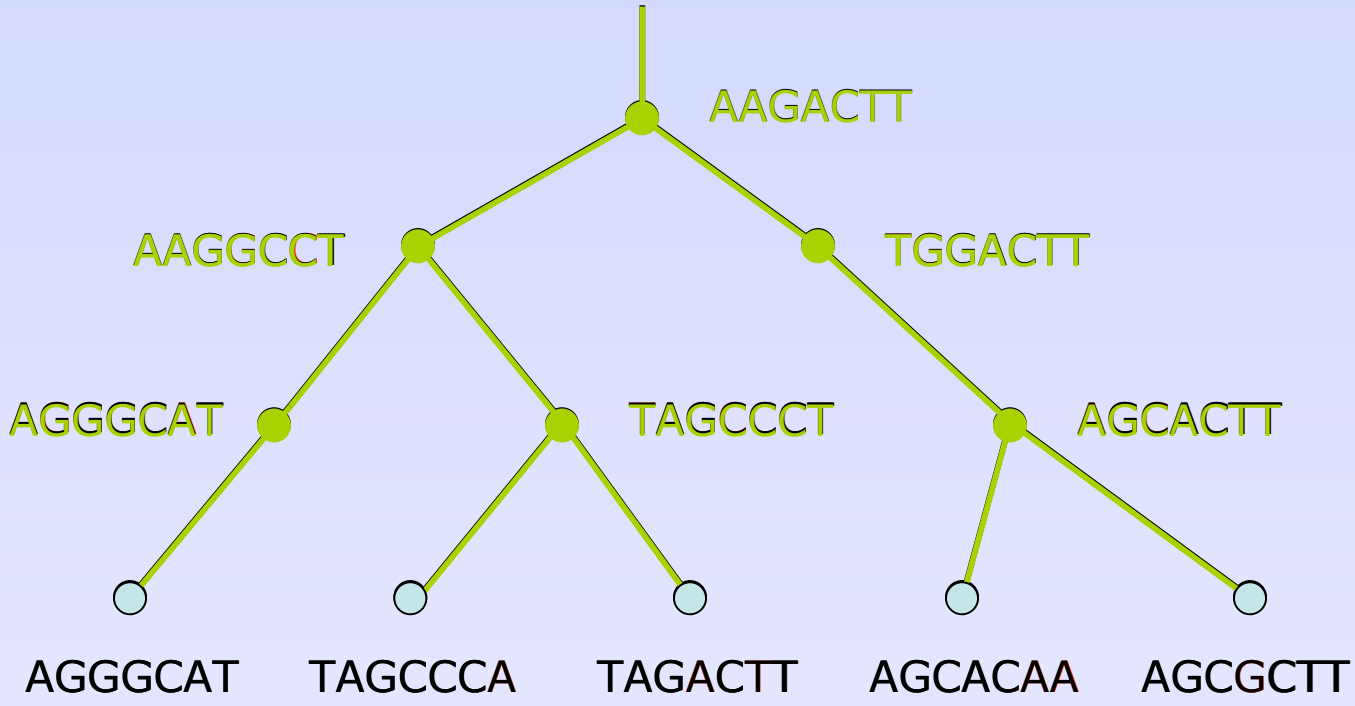
New methods for simultaneous estimation of trees and alignments

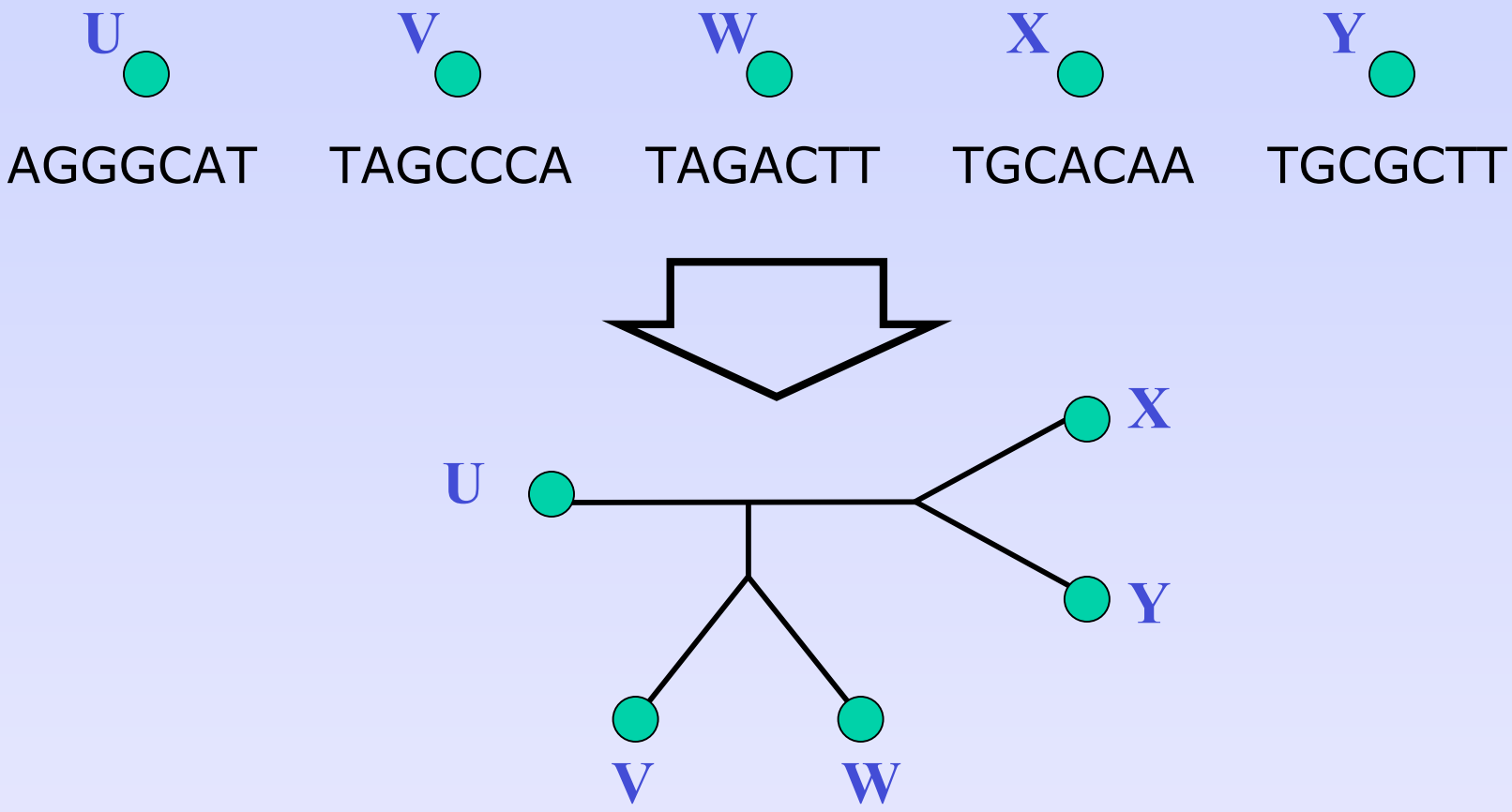
Tandy Warnow

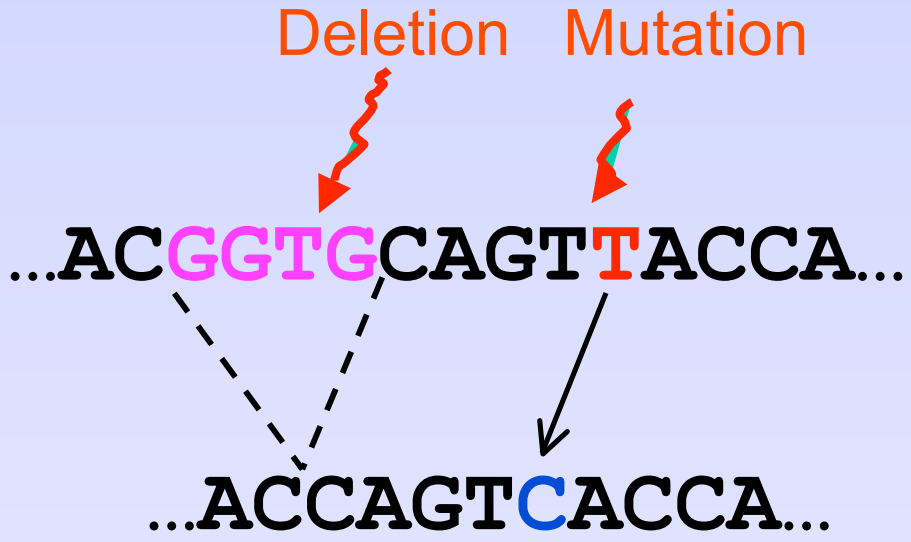
The University of Texas at Austin

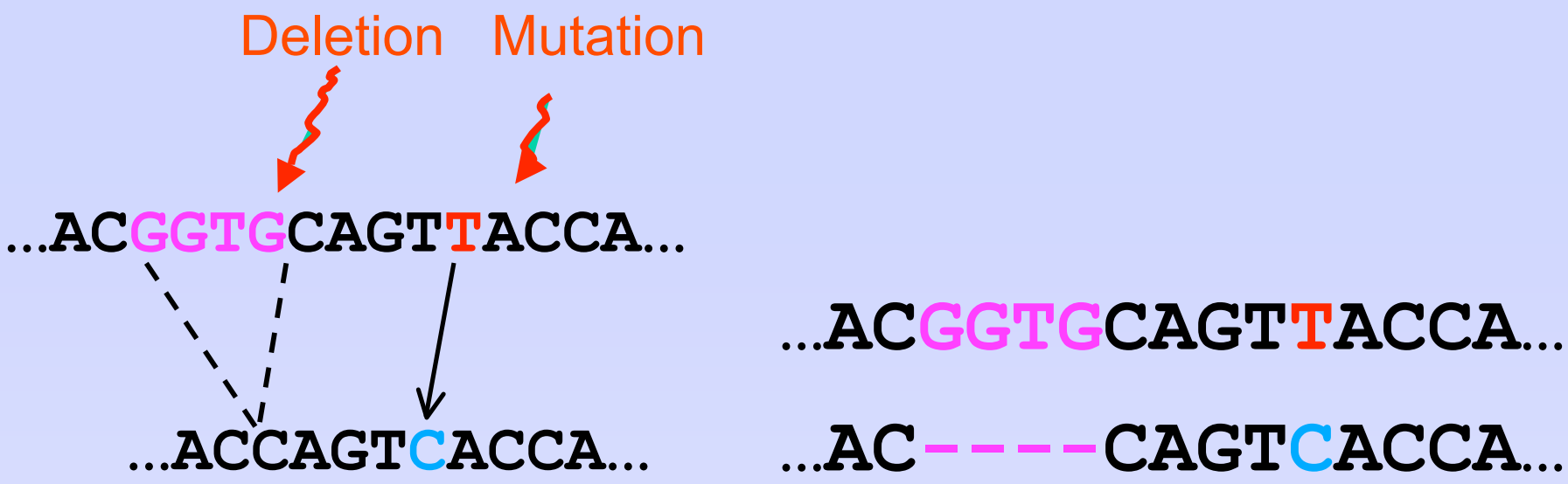
Joint work with K. Liu, S. Raghavan, S. Nelesen, and C.R. Linder

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 in the true phylogeny

Input: unaligned sequences

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

Phase 1: Multiple Sequence Alignment

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



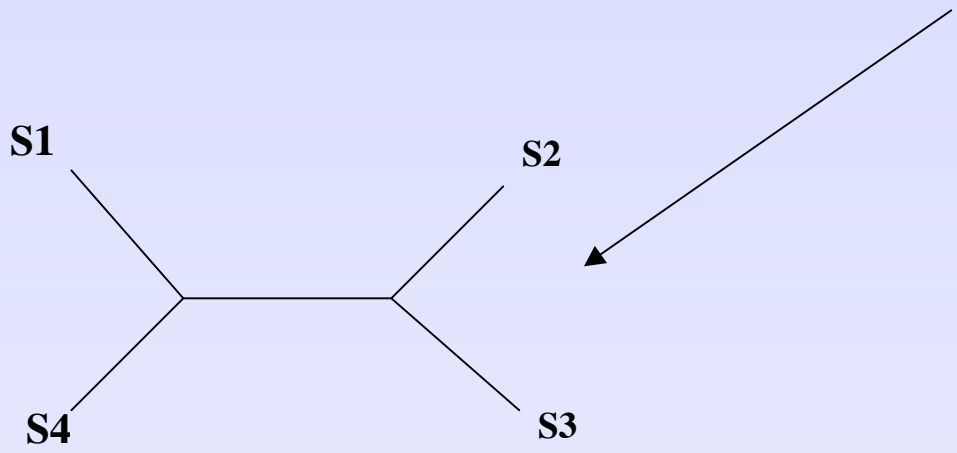
```
S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAC--GACCGC--
S3 = TAG-CT-------GACCGC--
S4 = -------TCAC--GACCGACA
```

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

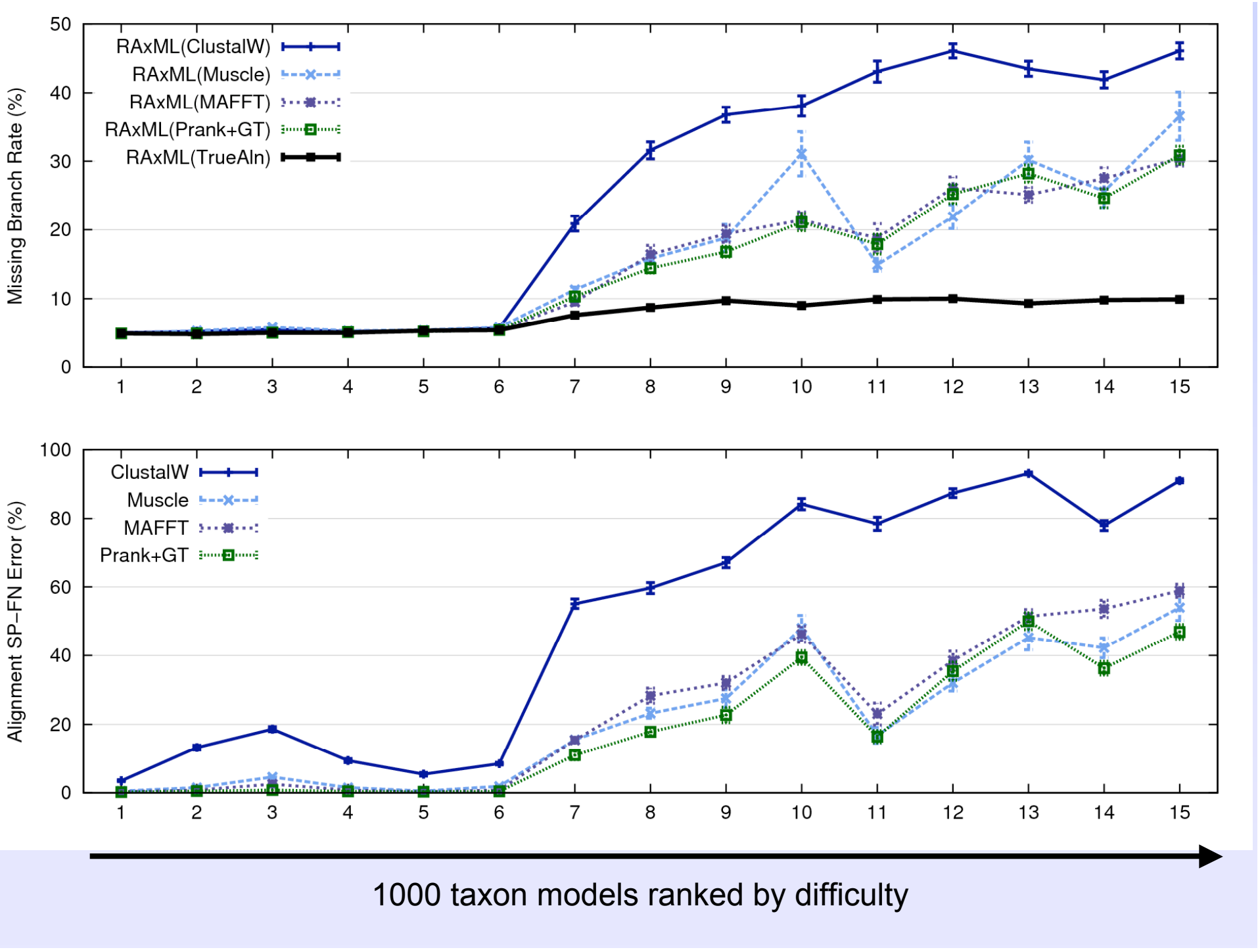
- Clustal
- POY (and POY*)
- Probcons (and Probtree)
- MAFFT
- Prank
- Muscle
- Di-align
- T-Coffee
- Opal
- Etc.

Phylogeny methods

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

Simulation study

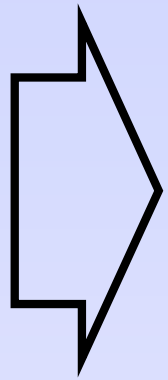
- ROSE simulation:
  - 1000, 500, and 100 sequences
  - Evolution with substitutions and indels
  - Varied gap lengths, rates of evolution
- Computed alignments
- Used RAxML to compute trees
- Recorded tree error (missing branch rate)
- Recorded alignment error (SP-FN)

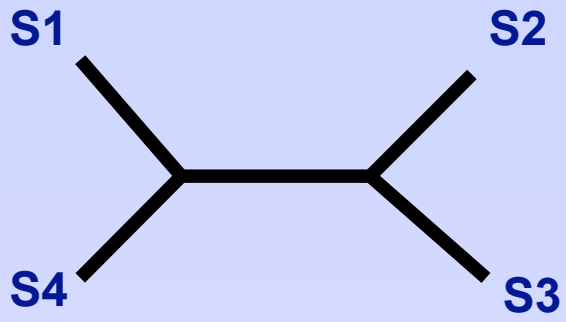

Missing Branch Rate (%)
RAxML(ClustalW)
RAxML(Muscle)
RAxML(MAFFT)
RAxML(Prank+GT)
RAxML(TrueAln)
Alignment SP–FN Error (%)
ClustalW
Muscle
MAFFT
Prank+GT
1000 taxon models ranked by difficulty

Problems with the two-phase approach

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

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





and

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

Current simultaneous estimation methods are not scalable.

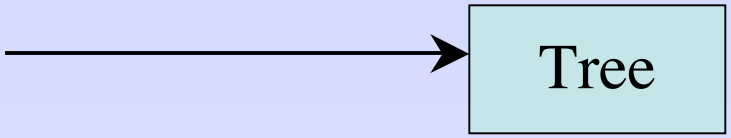
SATé:
(Simultaneous Alignment and Tree Estimation)

- Developers: Liu, Nelesen, Raghavan, Linder, and Warnow
- Search strategy: search through tree space, and *realigns sequences on each tree using a novel divide-and-conquer approach*.
- 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.

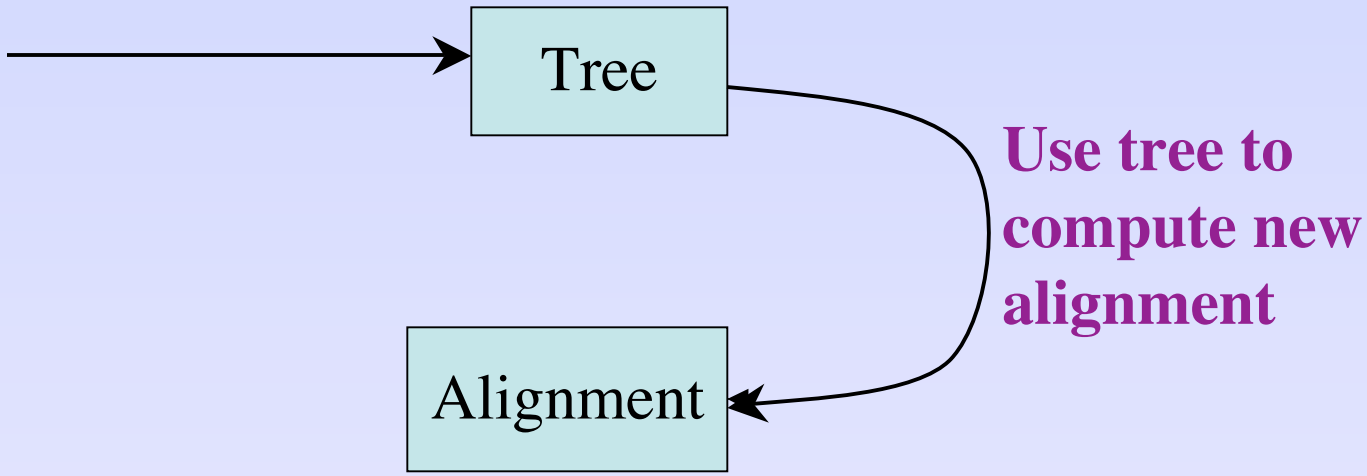
SATé Algorithm

Obtain initial alignment
and estimated ML tree



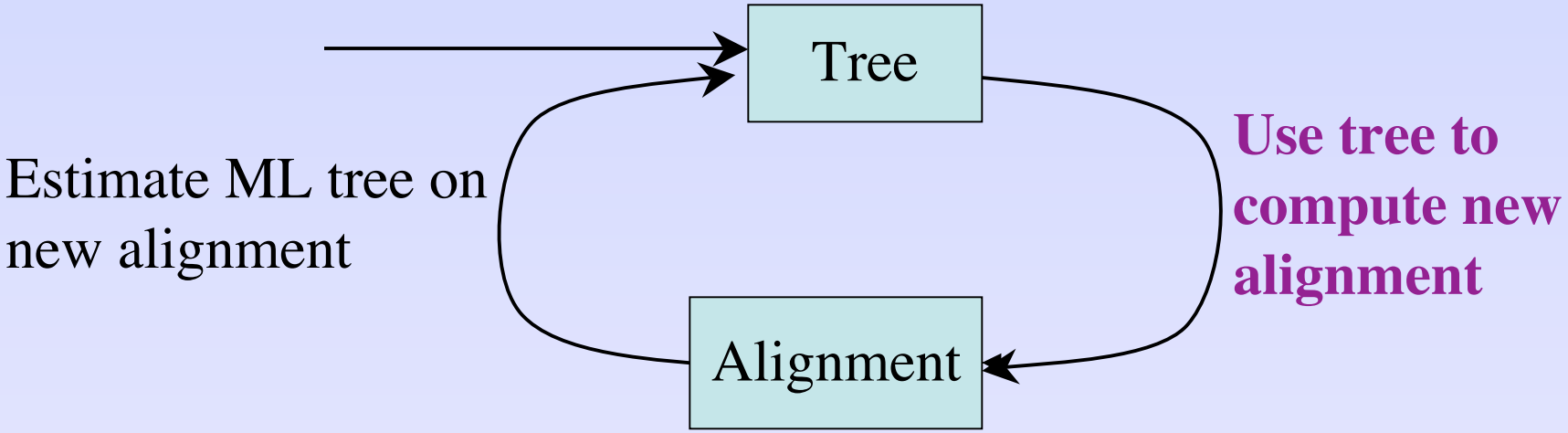
SATé Algorithm

Obtain initial alignment
and estimated ML tree

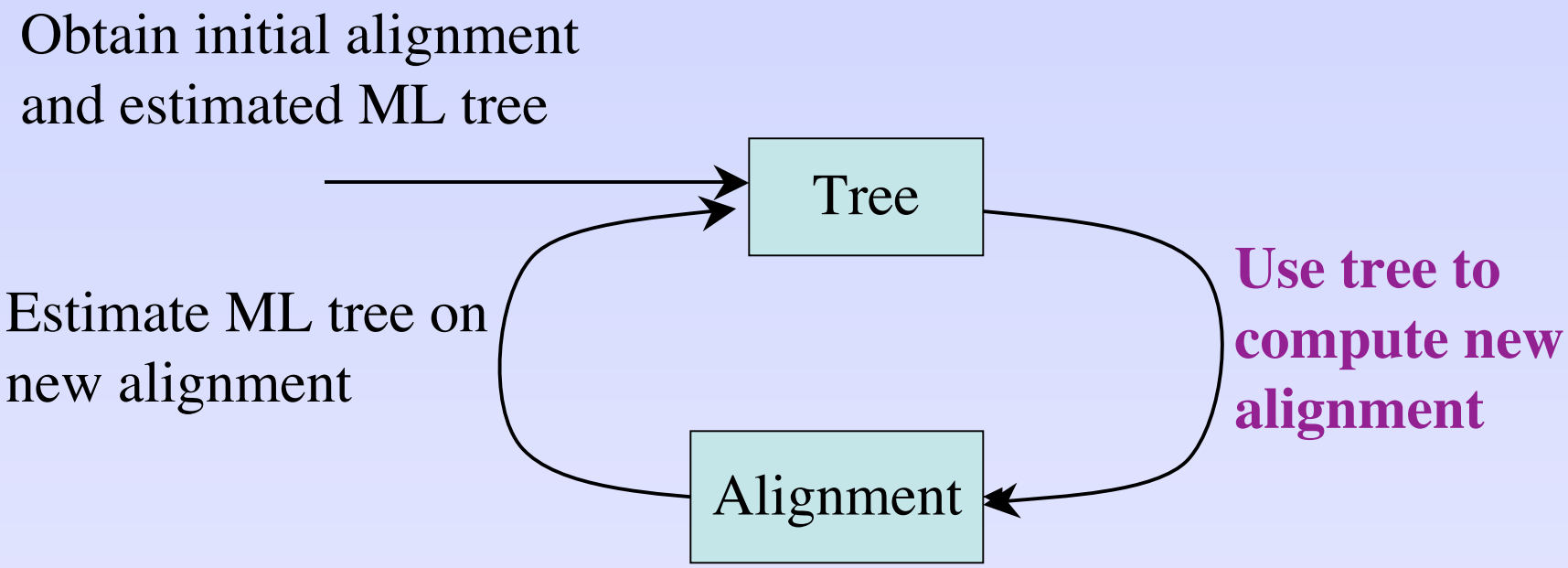


SATé Algorithm

Obtain initial alignment
and estimated ML tree

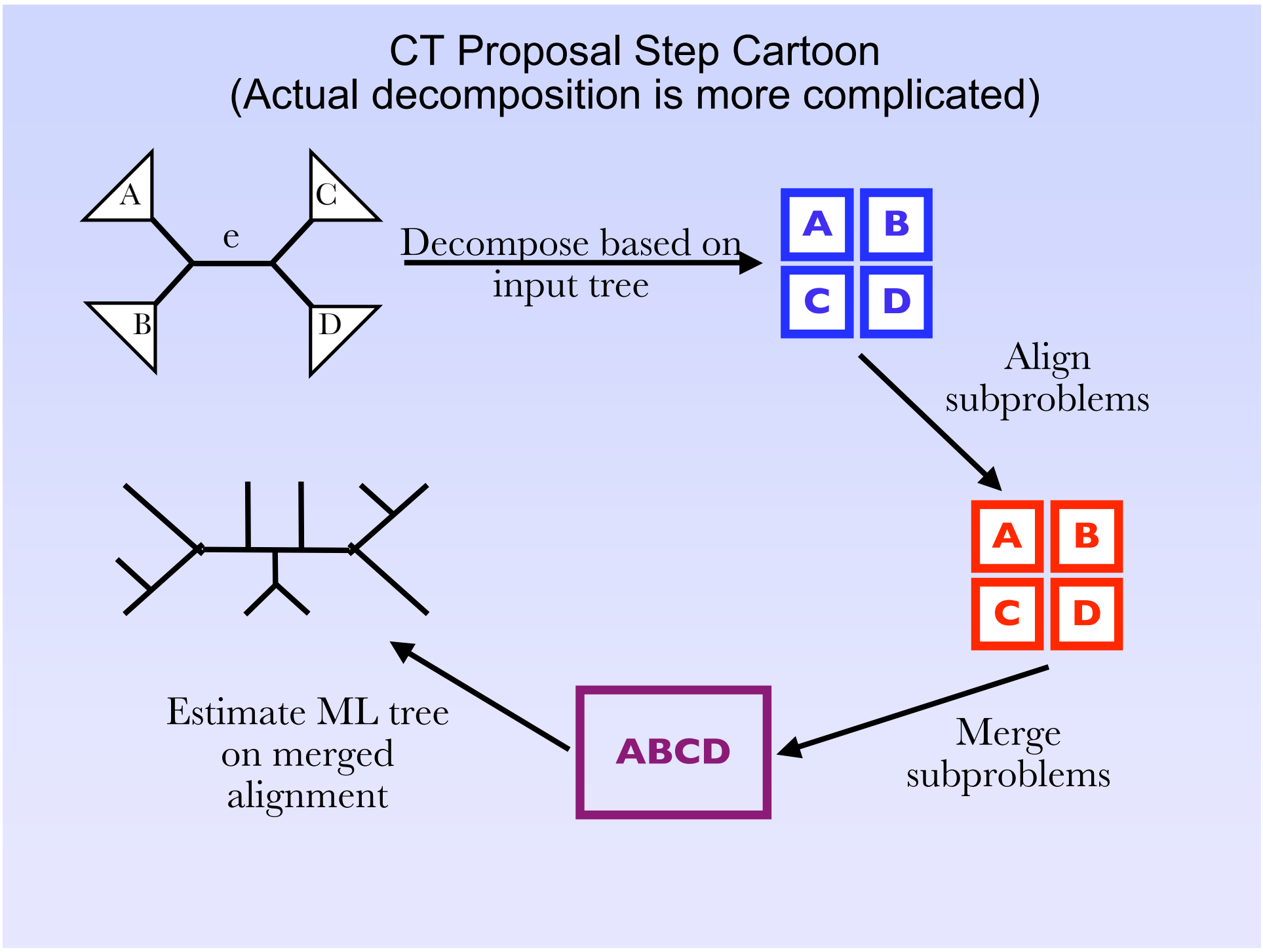


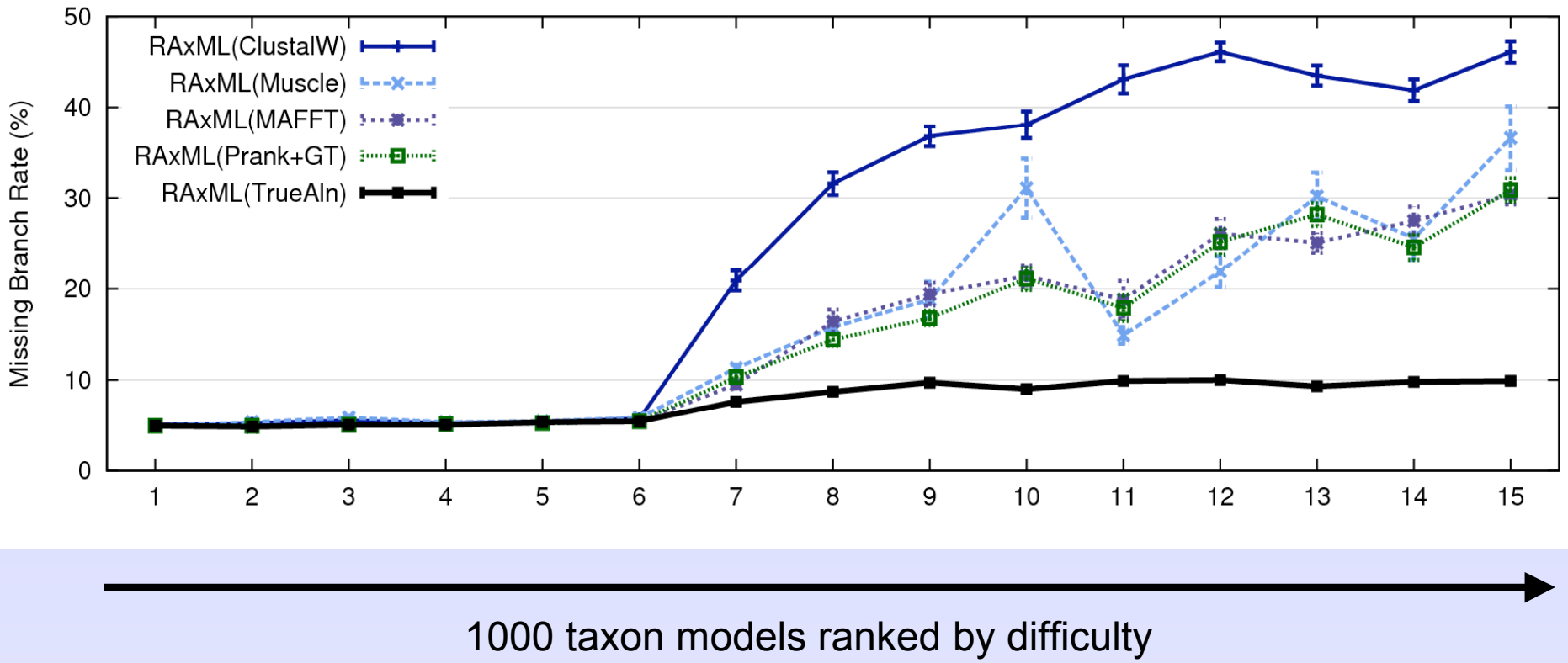
SATé Algorithm



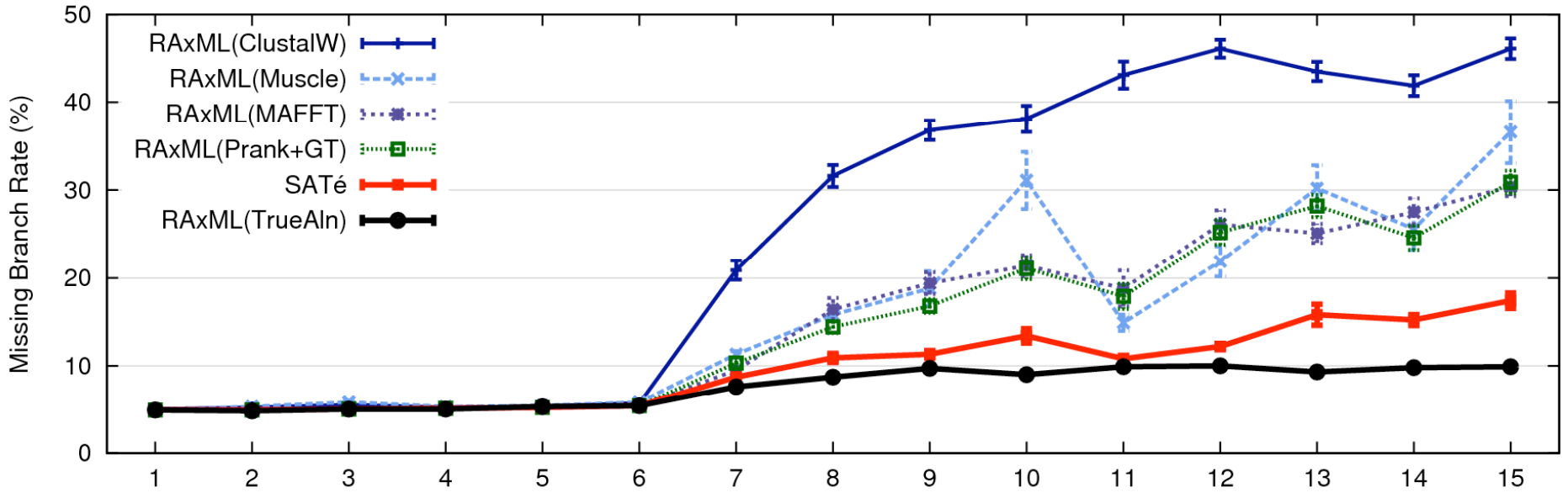
If new alignment/tree pair has worse ML score, realign using a different decomposition

Repeat until termination condition (typically, 24 hours)


CT Proposal Step Cartoon
(Actual decomposition is more complicated)
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



1000 taxon models ranked by difficulty



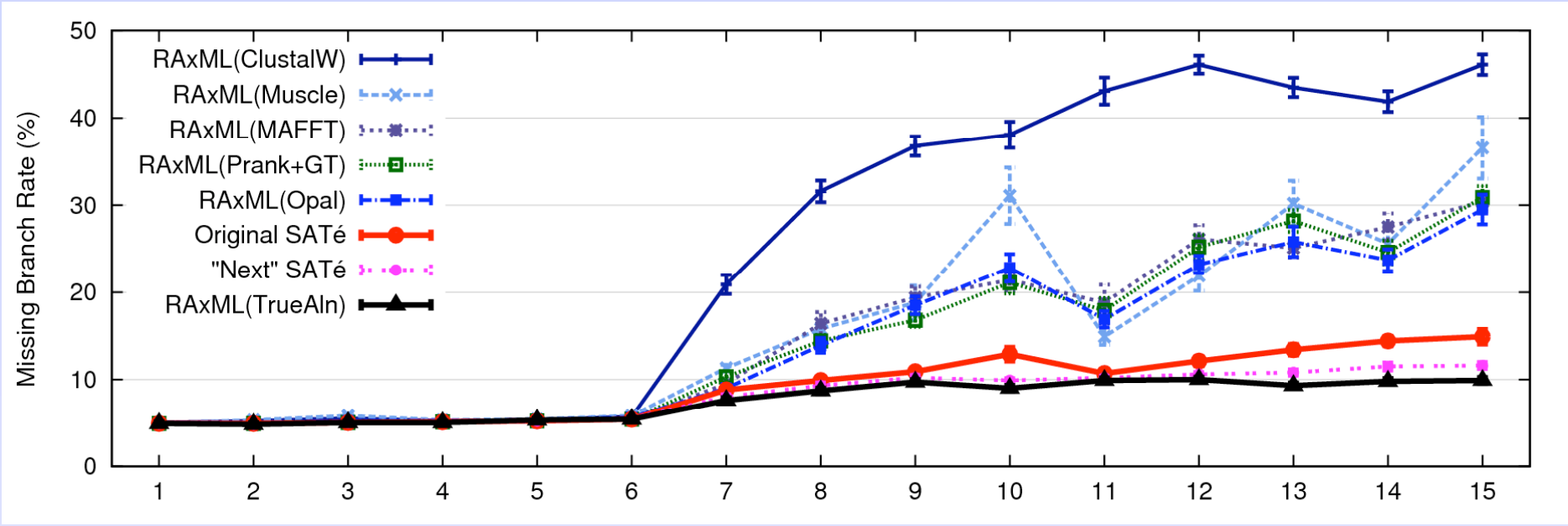
1000 taxon models ranked by difficulty

24 hour analysis, on desktop machines

“Next” SATé

Same basic strategy, but:

- ***Changed the technique to decompose into subproblems***
- Use ***Opal*** (Wheeler and Kececioglu, 2006) instead of Muscle to merge alignments

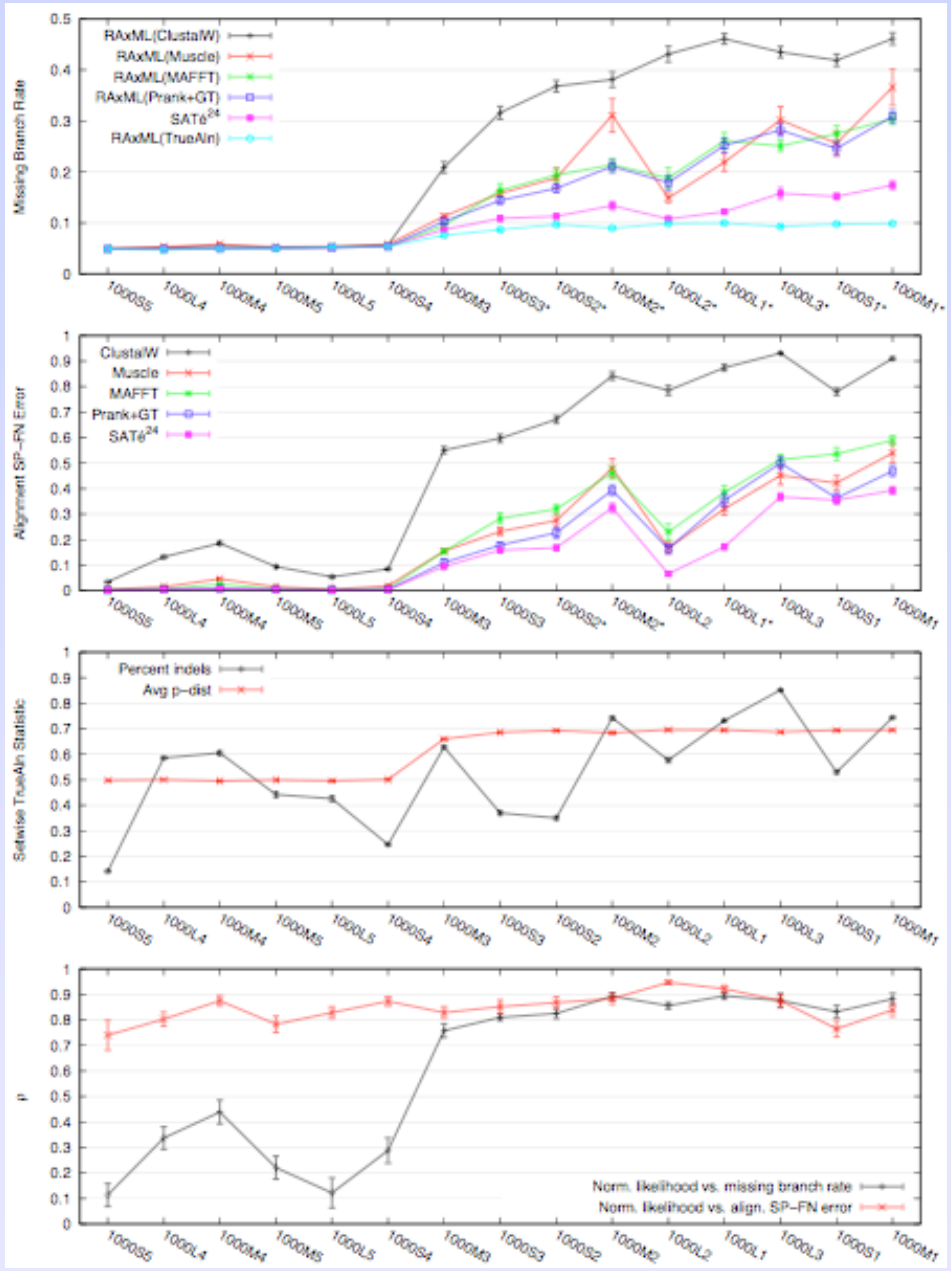


Liu et al., in preparation

Biological datasets

- ML analyses of curated alignments
  - 8 rDNA datasets produced by Robin Gutell
  - Early Bird ATOL project
  - Datasets from UT faculty
- Compared alignments and trees to the curated alignment and to the reference tree (bootstrap RAxML tree on the curated alignment)
- Typically, SATé produced trees closer to the reference trees than the other methods

Why does SATé perform well?


Missing Branch Rate
RAxML(ClustalW)
RAxML(Muscle)
RAxML(MAFFT)
RAxML(Prank+GT)
SATé24
RAxML(TrueAln)
1000S5 1000L4 1000M4 1000M5 1000L5 1000S4 1000M3 1000S3* 1000S2* 1000M2* 1000L2* 1000L1* 1000L3* 1000S1* 1000M1*
Alignment SP–FN Error
ClustalW
Muscle
MAFFT
Prank+GT
SATé24
1000S5 1000L4 1000M4 1000M5 1000L5 1000S4 1000M3 1000S3 1000S2* 1000M2* 1000L2 1000L1* 1000L3 1000S1 1000M1
Setwise TrueAln Statistic
Percent indels
Avg p–dist
1000S5 1000L4 1000M4 1000M5 1000L5 1000S4 1000M3 1000S3 1000S2 1000M2 1000L2 1000L1 1000L3 1000S1 1000M1
ρ
Norm. likelihood vs. missing branch rate
Norm. likelihood vs. align. SP–FN error
1000S5 1000L4 1000M4 1000M5 1000L5 1000S4 1000M3 1000S3 1000S2 1000M2 1000L2 1000L1 1000L3 1000S1 1000M1

Why does SATé perform well?

Answer: *not because we optimize ML (in which we treat gaps as missing data)!*

- Using a different re-alignment technique, Alexis Stamatakis has demonstrated that optimizing ML scores (treating gaps as missing data) can produce very bad alignments and trees.

- But SATé produces highly accurate trees and alignments.
  - It seems likely that the SATé re-alignment techniques do not produce problematic alignments - these rely upon alignment methods, MAFFT and Opal/Muscle, which have reasonable gap treatments.
  - In a sense, the use of ML within SATé is secondary: ML is used to select among reasonable alignments, not to generate the alignments.
  - Understanding why SATé works well is an interesting research question.

Conclusions

- SATé produces trees and alignments that improve upon the best two-phase methods for hard-to-align datasets, and can do so in reasonable time frames (at most a few days) on desktop computers.
- Improvements are underway.
- Better results would likely be obtained by using indels within the ML model. However, scalability of such methods is essential.

Acknowledgments

- National Science Foundation

- The Program for Evolutionary Dynamics at Harvard

- Collaborators: Randy Linder, Kevin Liu, Serita Nelesen, and Sindhu Raghavan