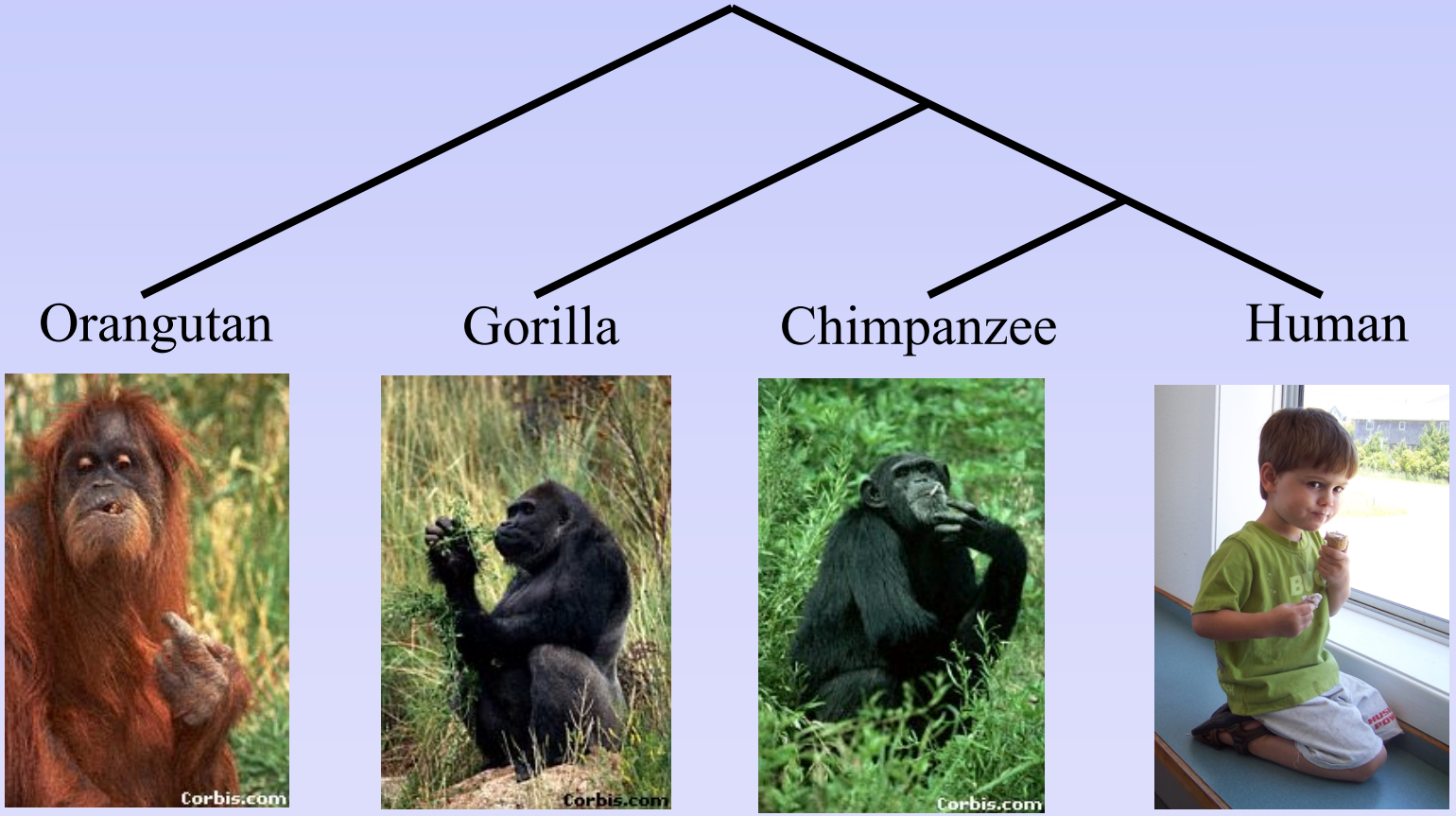
Methods for estimating ultra-large phylogenies and alignments

Tandy Warnow
Microsoft Research New England
The University of Texas at Austin

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

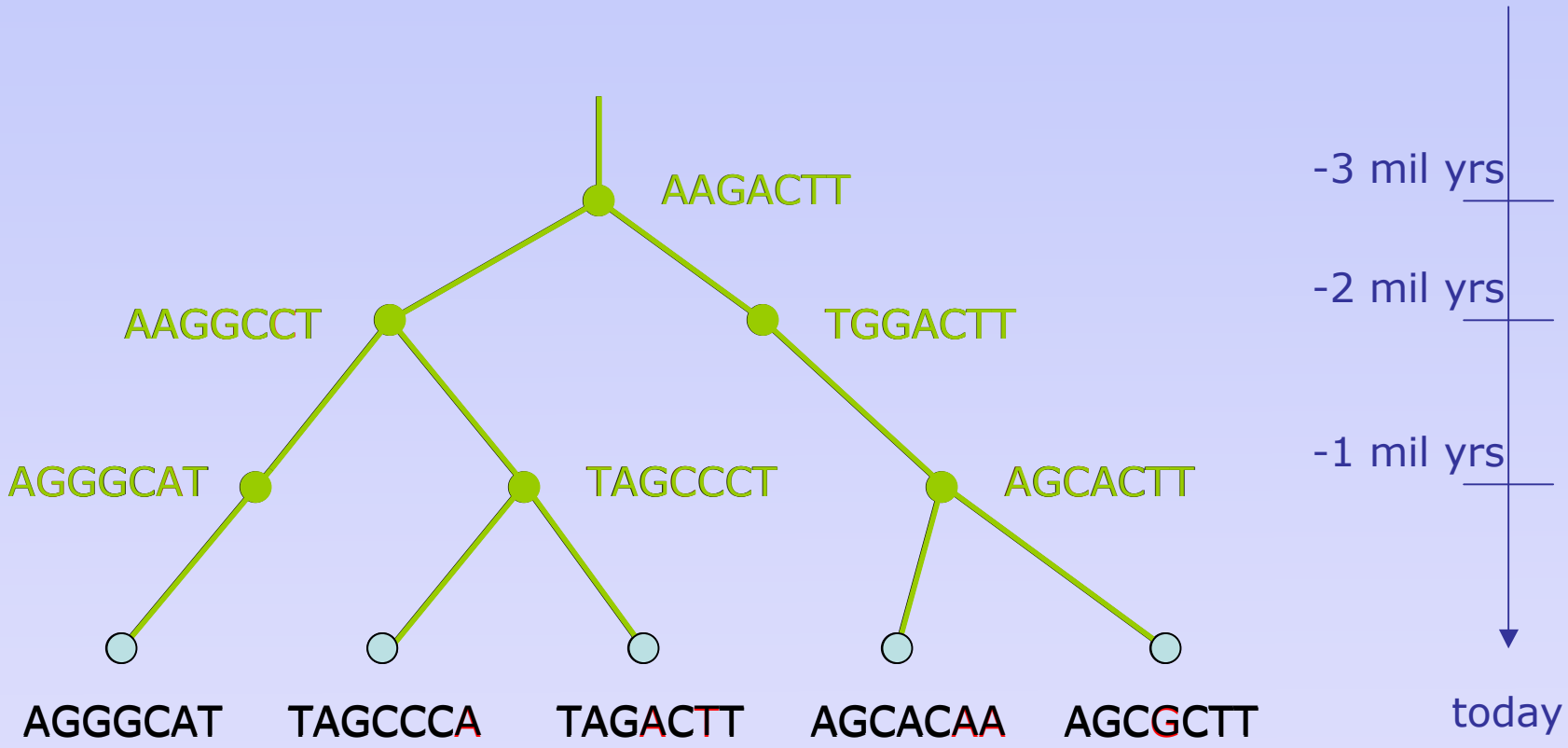
NP-hard optimization problems

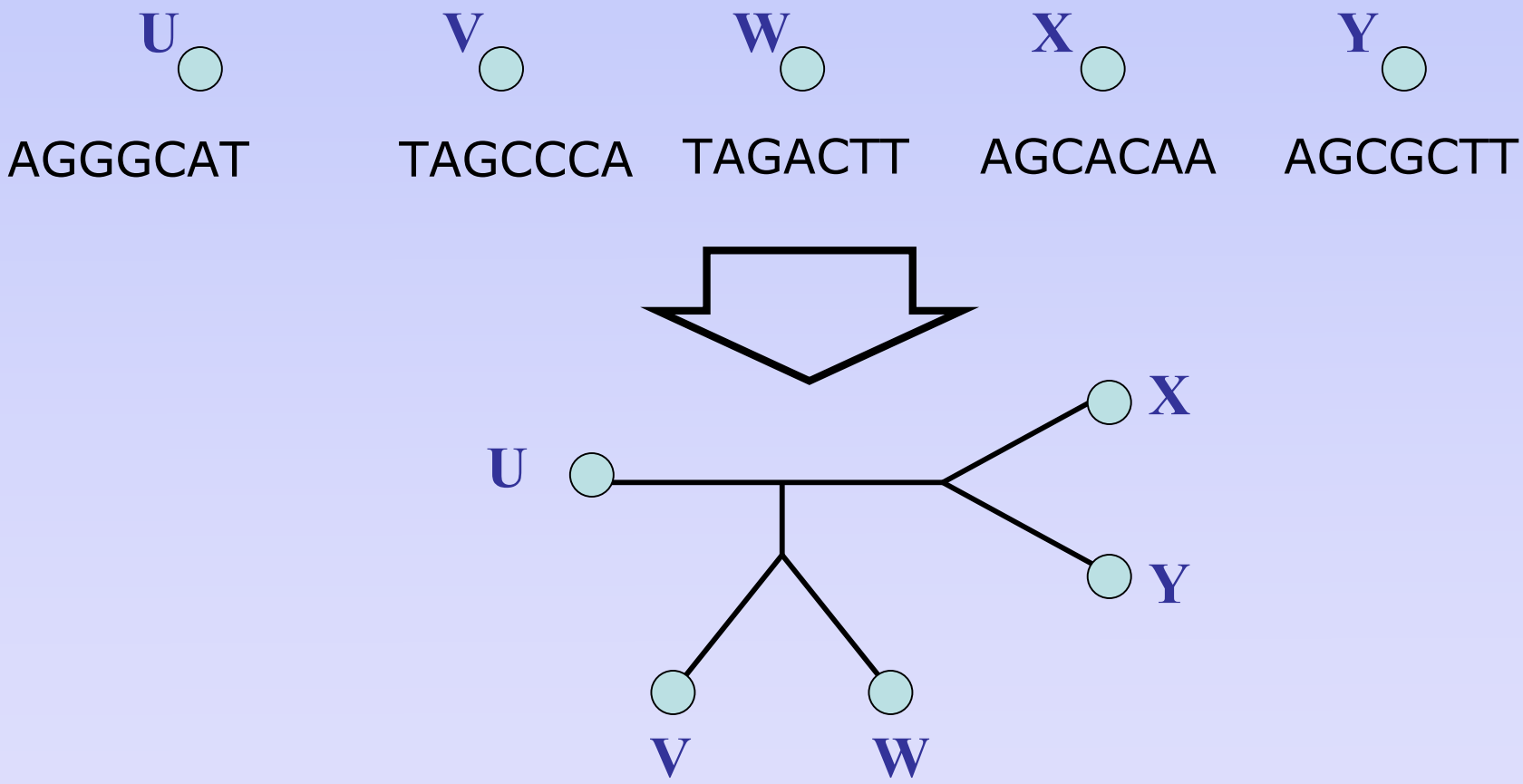
Millions of taxa

Important applications

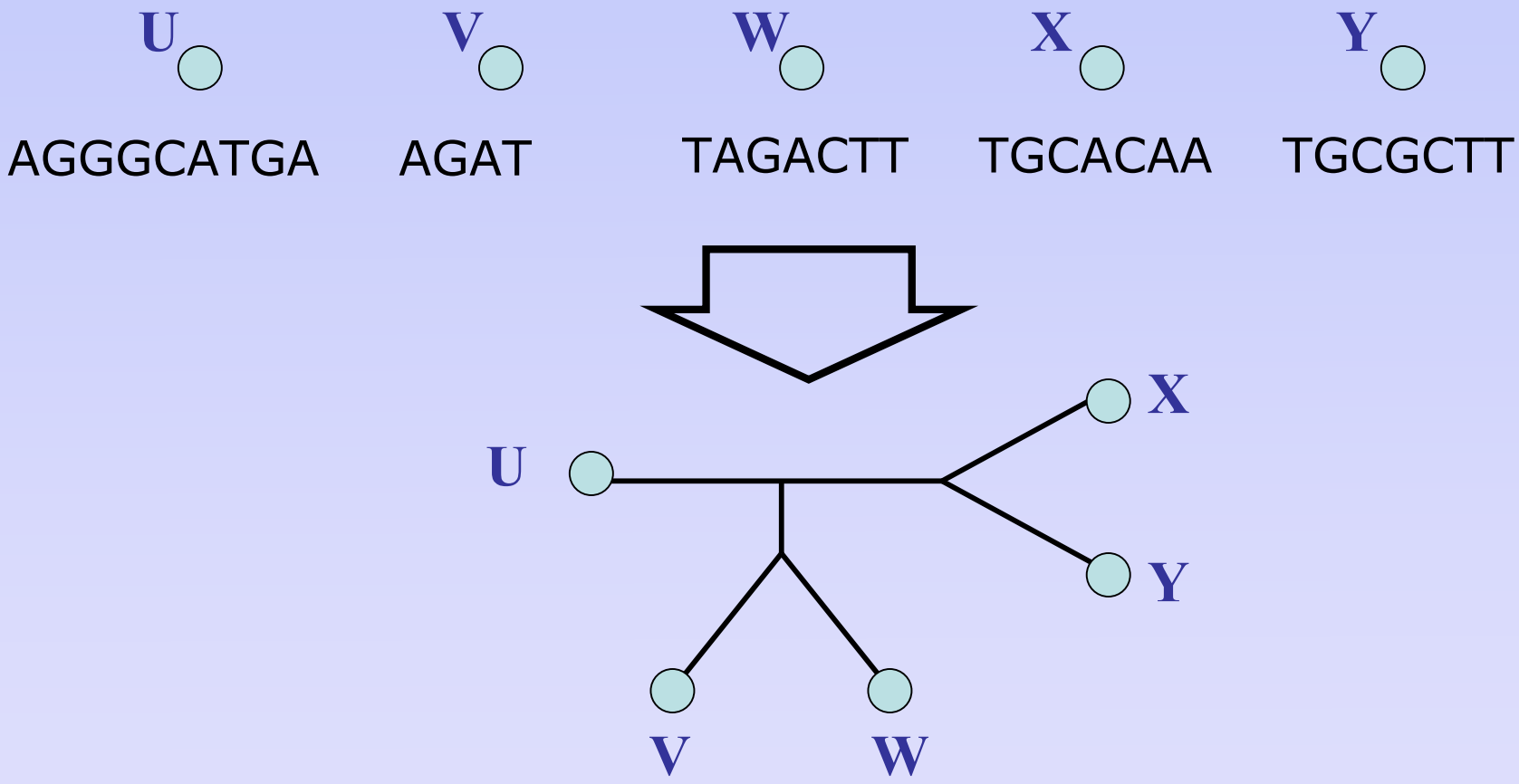
Current projects (e.g., iPlant) will attempt to estimate phylogenies with upwards of 500,000 species

DNA Sequence Evolution

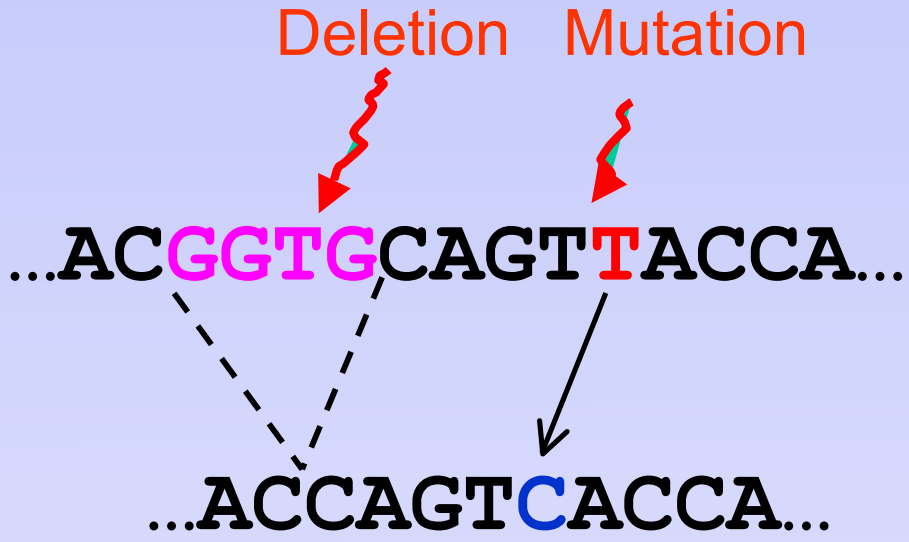


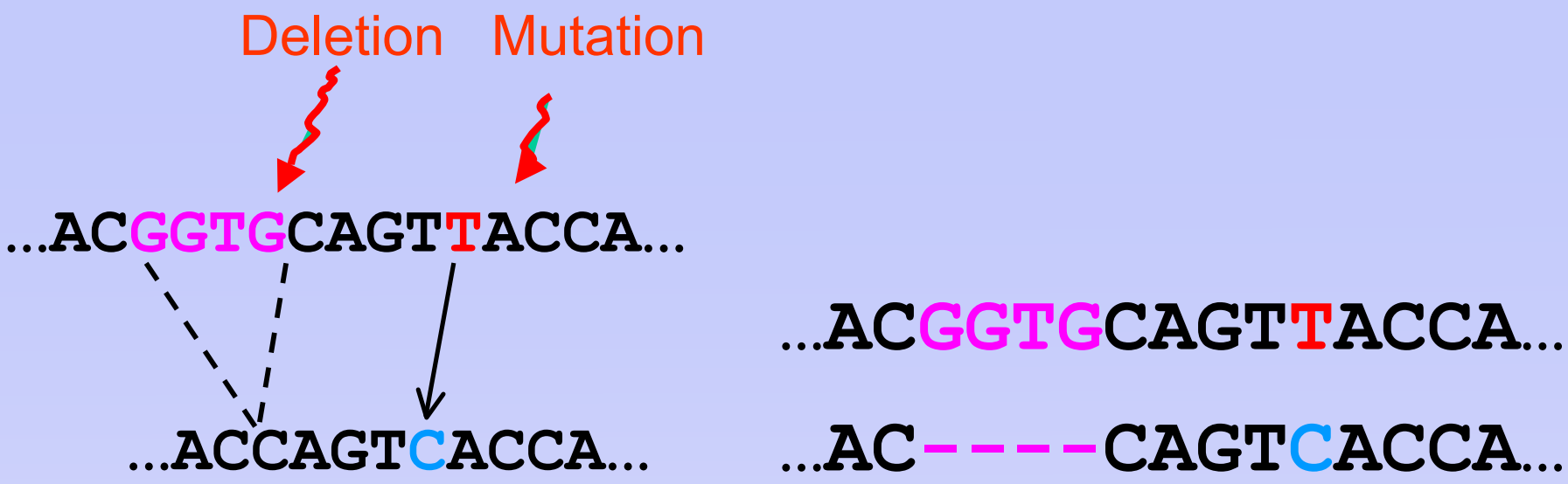

U
V
W
X
Y
AGGGCAT
TAGCCCA
TAGACTT
AGCACAA
AGCGCTT
X
U
Y
V
W

Consider real data:



Indels (insertions and deletions) also occur!





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
```



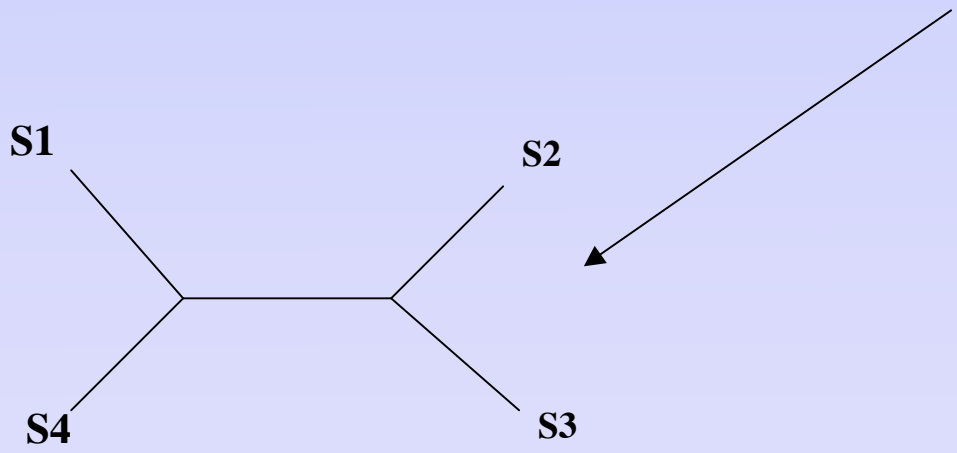
```
S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAC--GACCGC--
S3 = TAG-CT-------GACCGC--
S4 = -------TCAC--GACCGACA
```

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
```



Two-phase estimation

Alignment methods

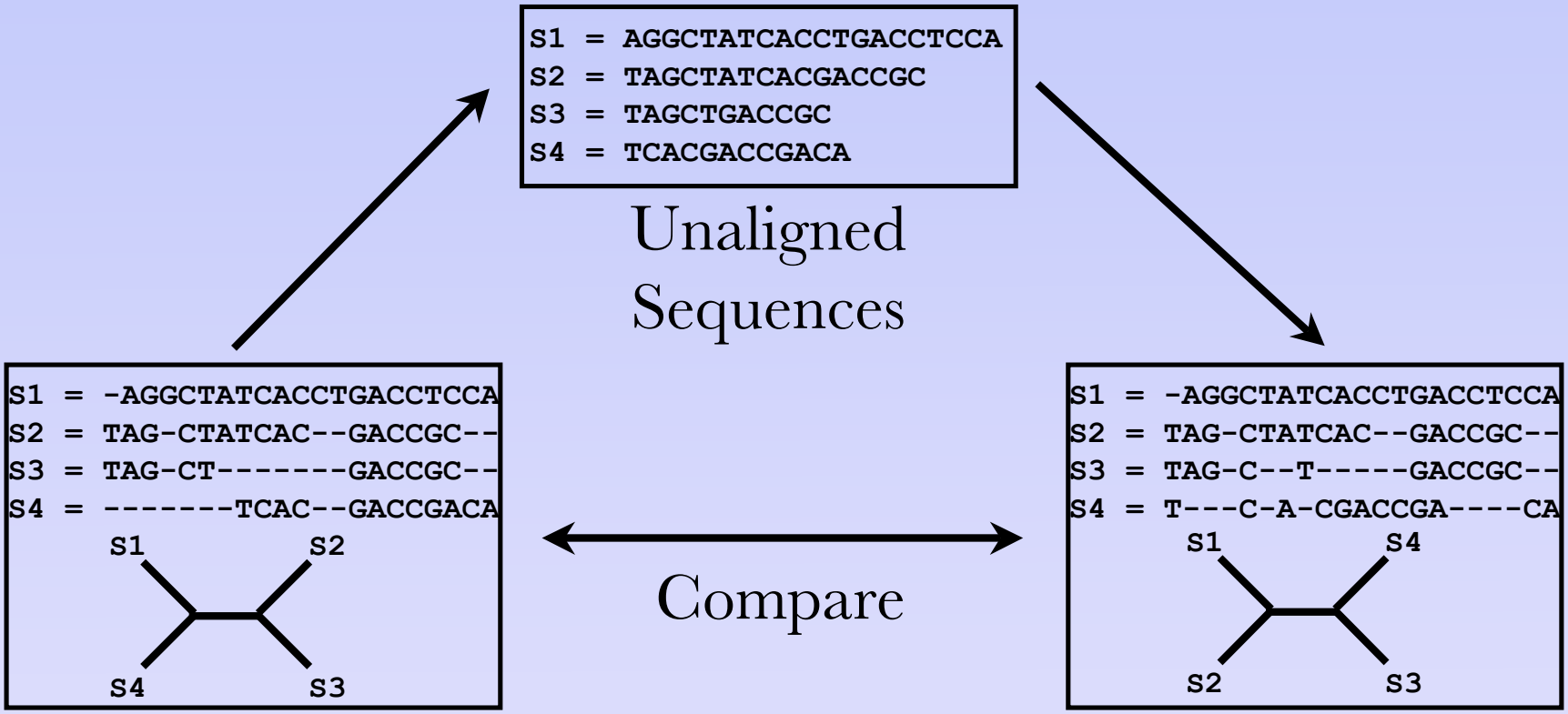
- Clustal
- POY (and POY*)
- Probcons (and Probtree)
- Probalign
- MAFFT
- Muscle
- Di-align
- T-Coffee
- Prank (PNAS 2005, Science 2008)
- Opal (ISMB and Bioinf. 2007)
- *FSA (PLoS Comp. Bio. 2009)*
- *Infernal (Bioinf. 2009)*
- Etc.

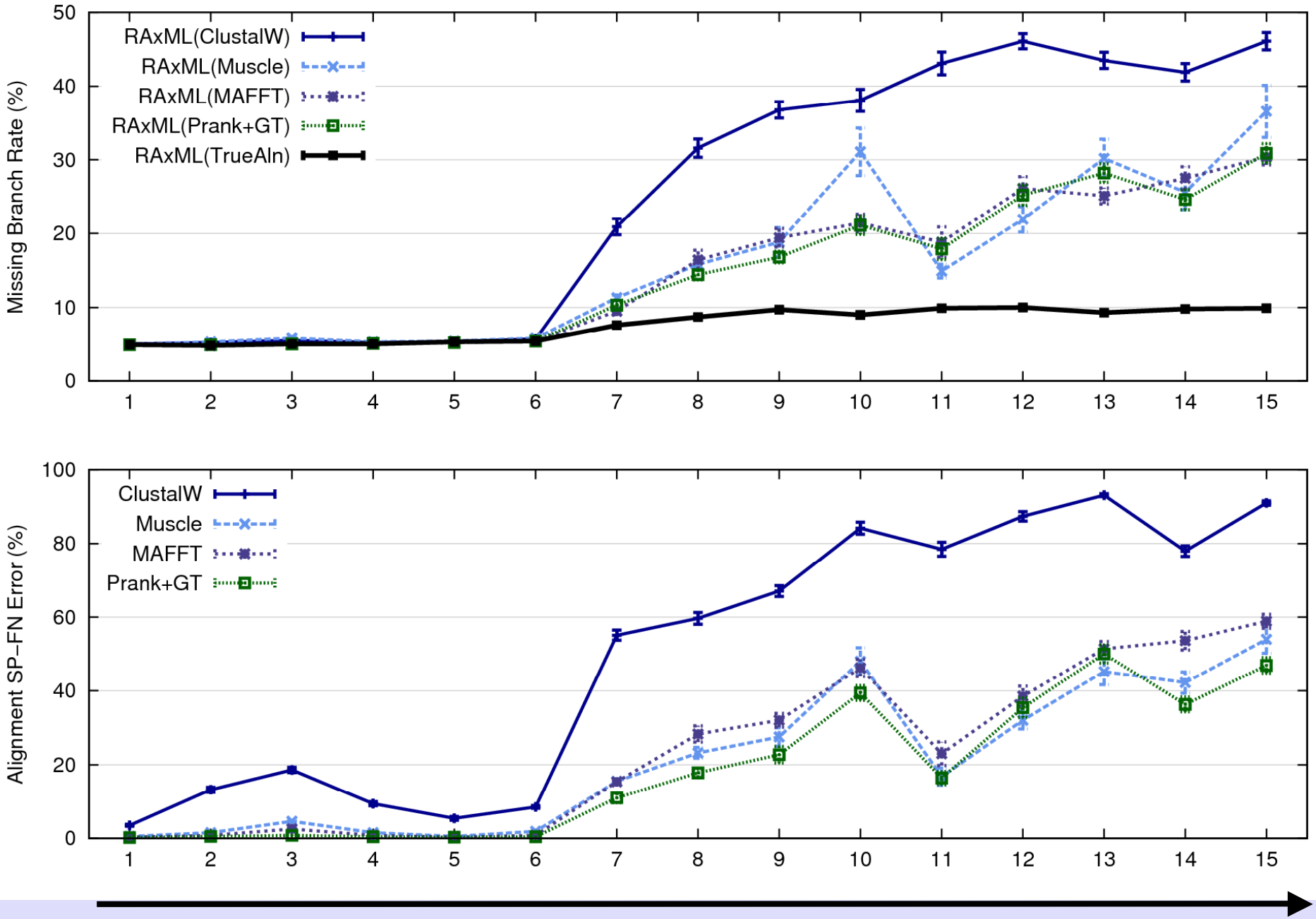
Phylogeny methods

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

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

Simulation Studies




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.
- *Biologists discard potentially useful markers* if they are difficult to align.

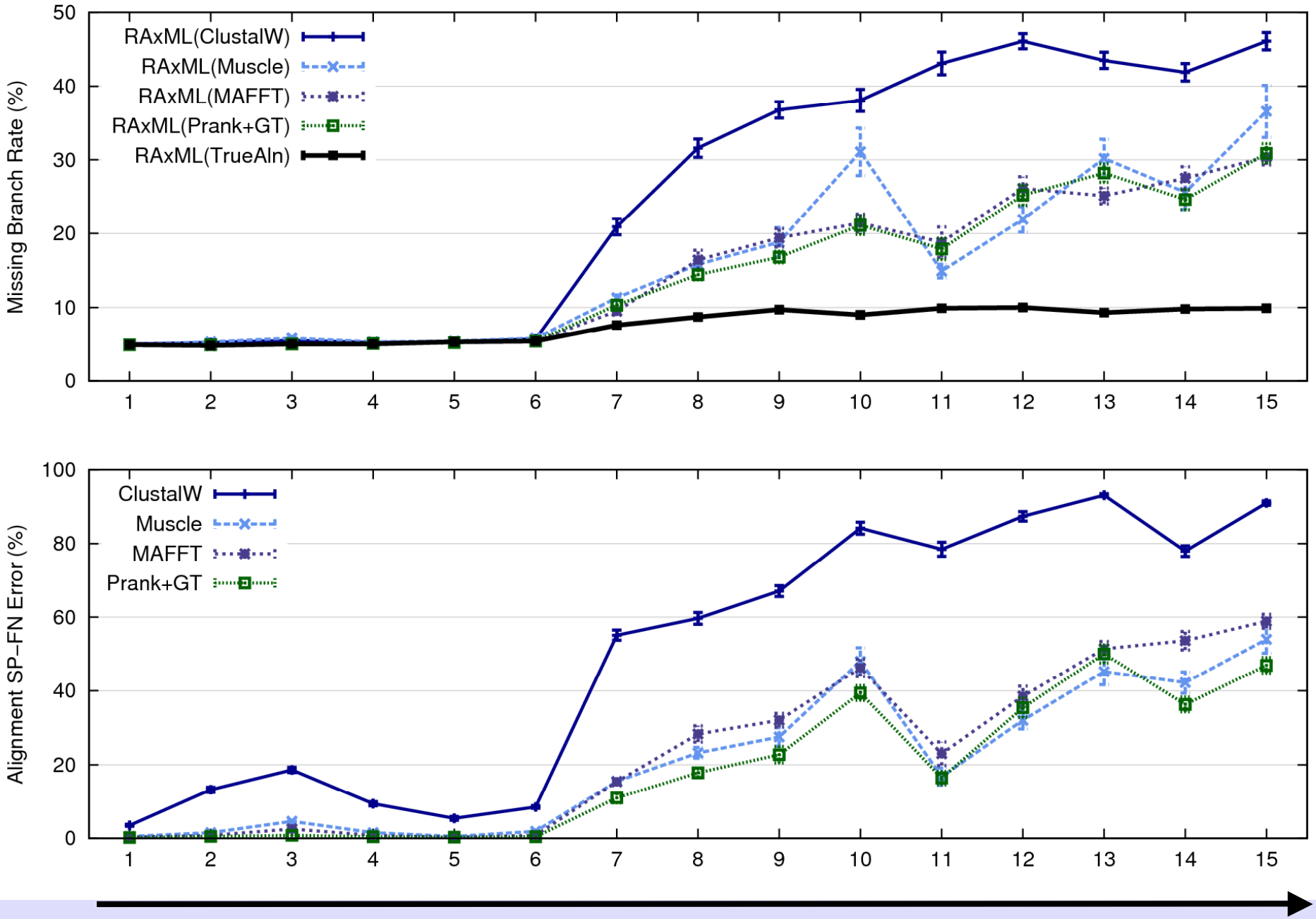
These issues seriously impact large-scale phylogeny estimation (and Tree of Life projects)

- SATé: Simultaneous Alignment and Tree Estimation (Liu et al., Science 2009, and Liu et al. Systematic Biology, in press)
- DACTAL: Divide-and-Conquer Trees without alignments (Nelesen et al., submitted)

Part 1: SATé

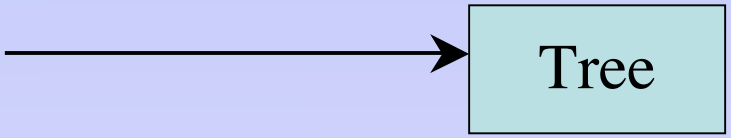
Liu, Nelesen, Raghavan, Linder, and Warnow, *Science*, 19 June 2009, pp. 1561-1564.

- Kansas SATé software developers: Mark Holder and Jiaye Yu
- Downloadable software for various platforms
- Easy-to-use GUI
- http://phylo.bio.ku.edu/software/sate/sate.html


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)

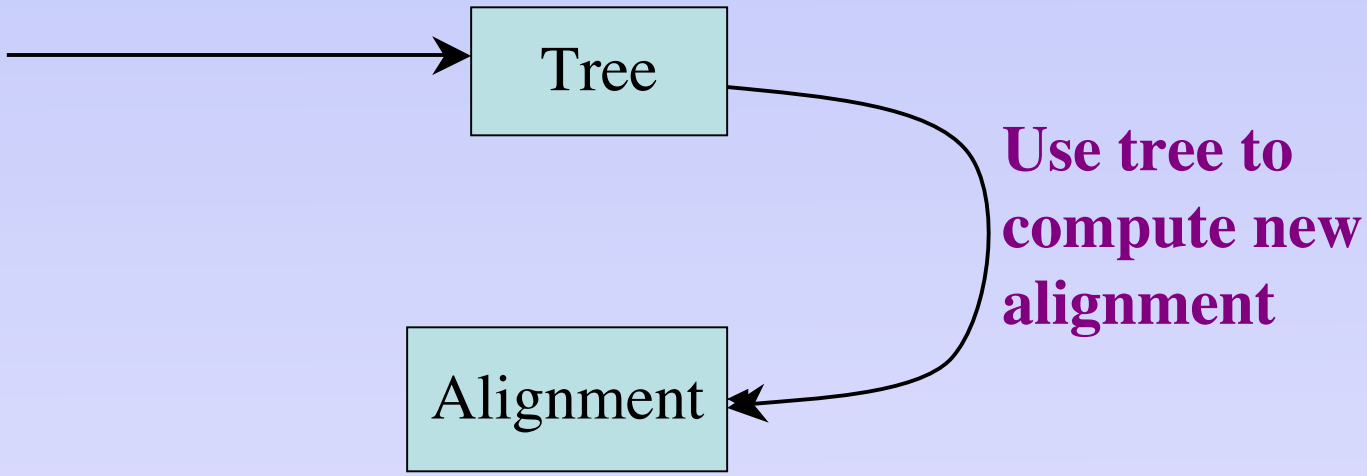
SATé Algorithm

Obtain initial alignment
and estimated ML tree



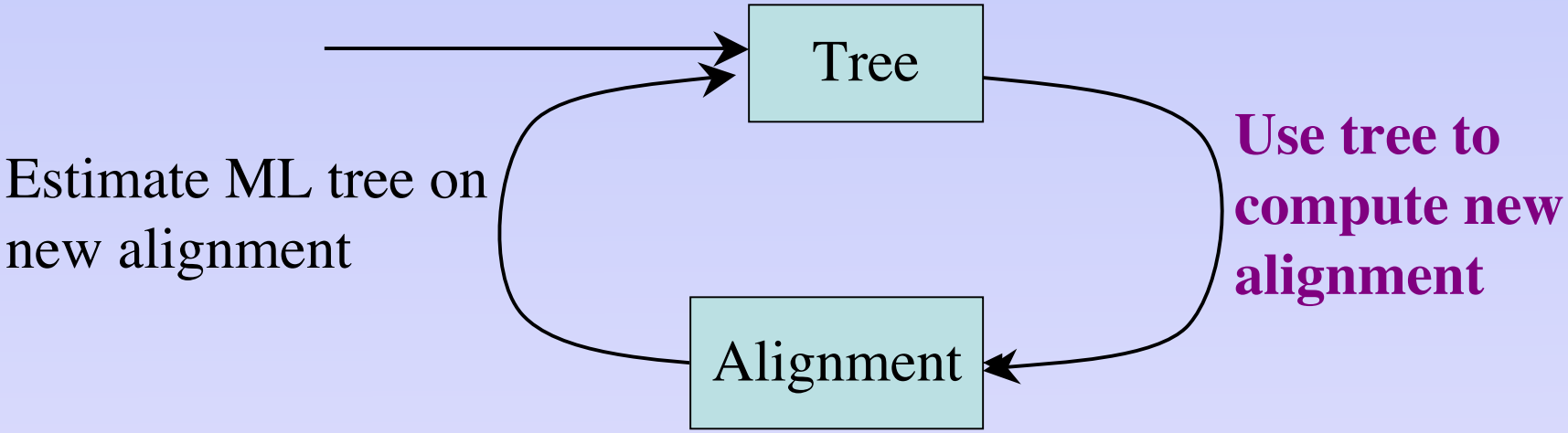
SATé Algorithm

Obtain initial alignment
and estimated ML tree



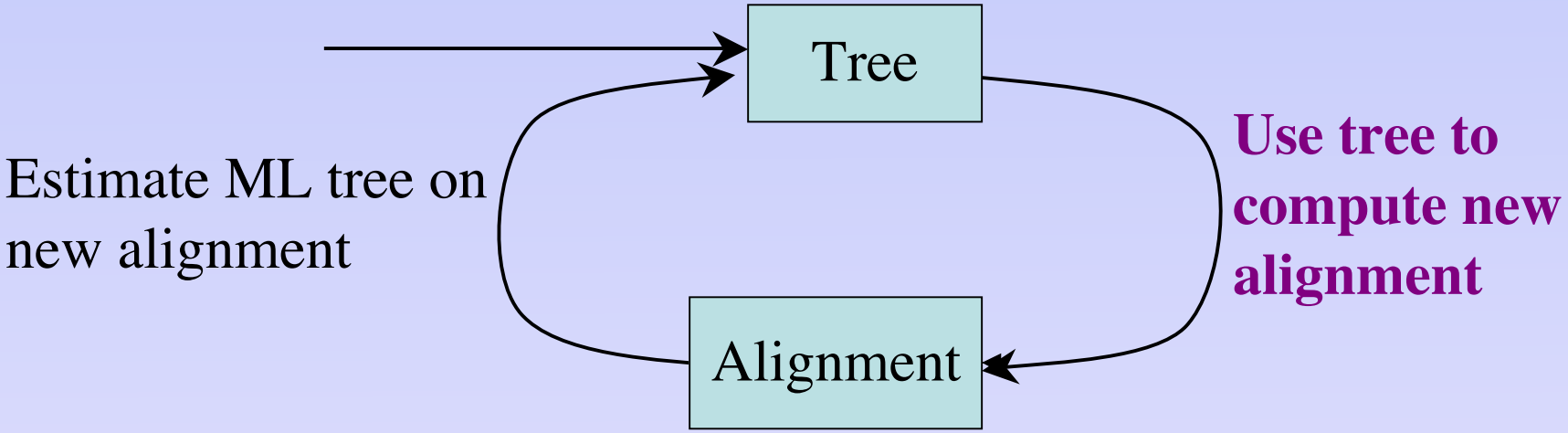
SATé Algorithm

Obtain initial alignment
and estimated ML tree



SATé Algorithm

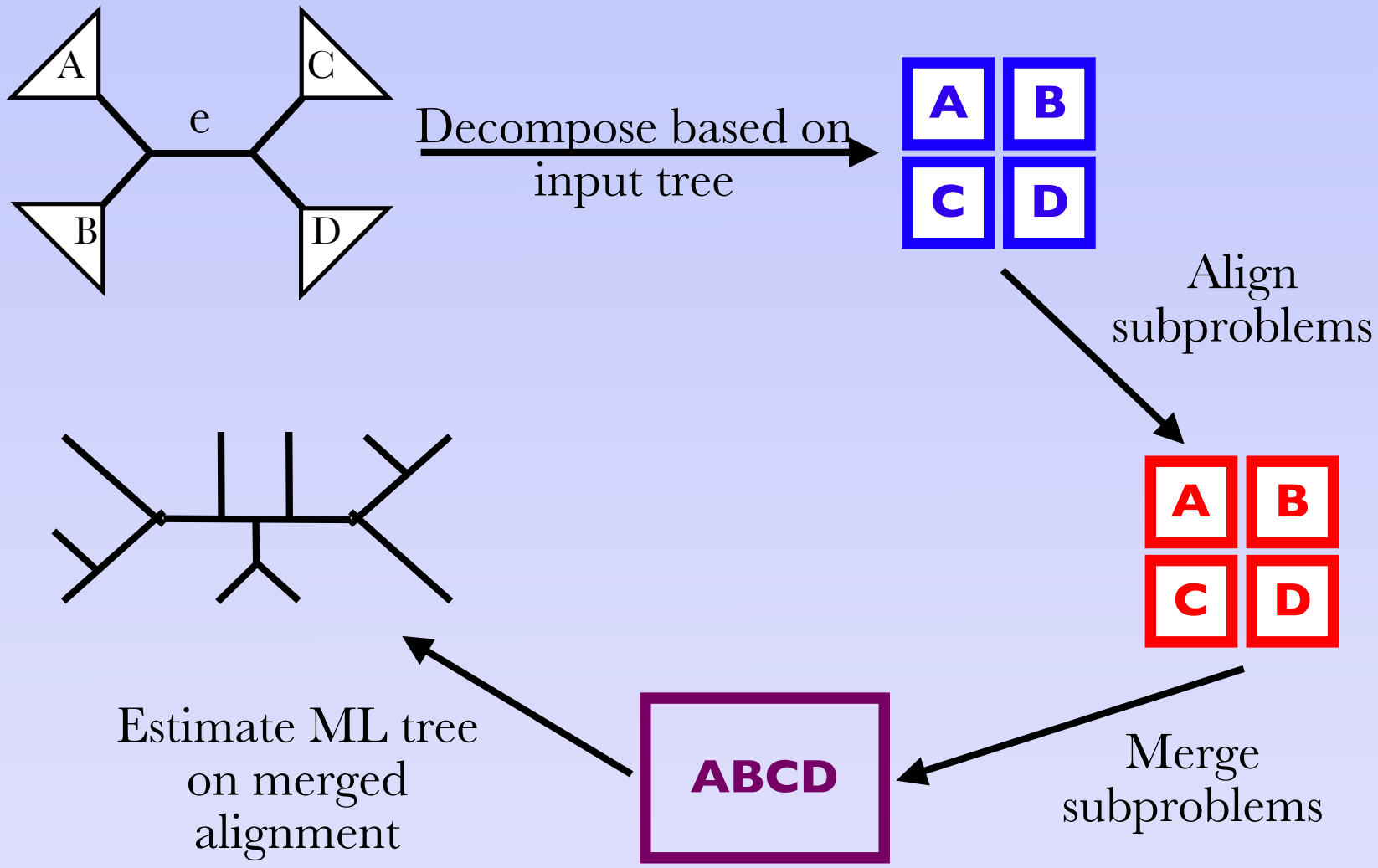
Obtain initial alignment
and estimated ML tree

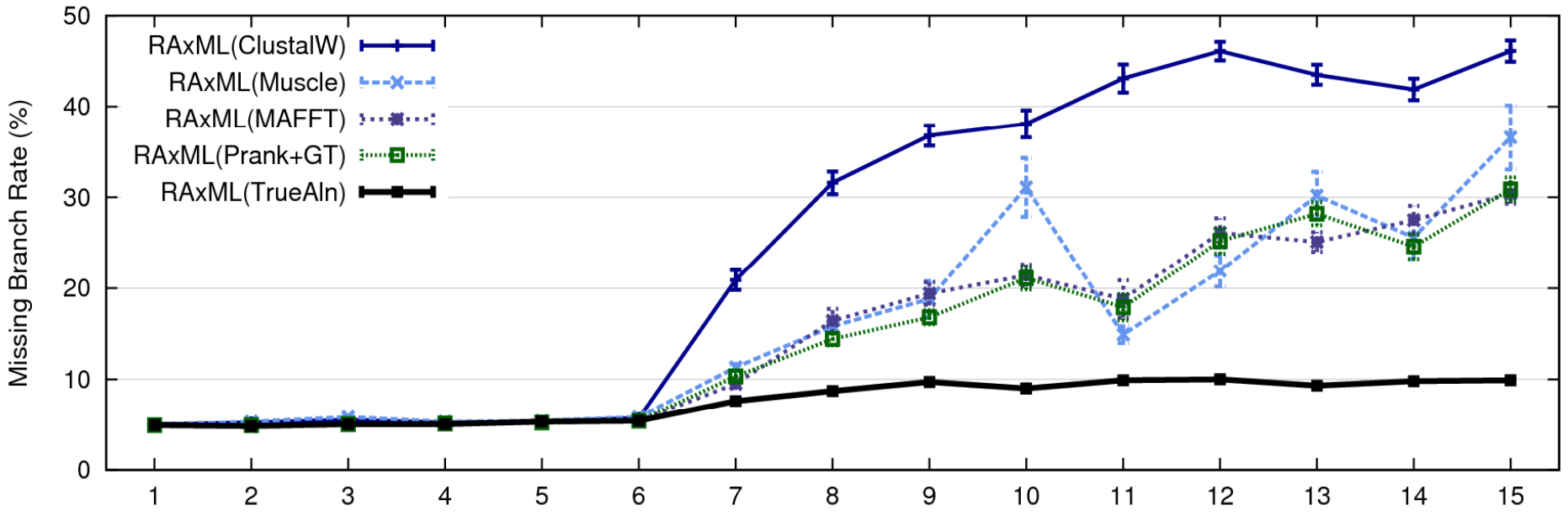


If new alignment/tree pair has worse ML score, realign using a different decomposition

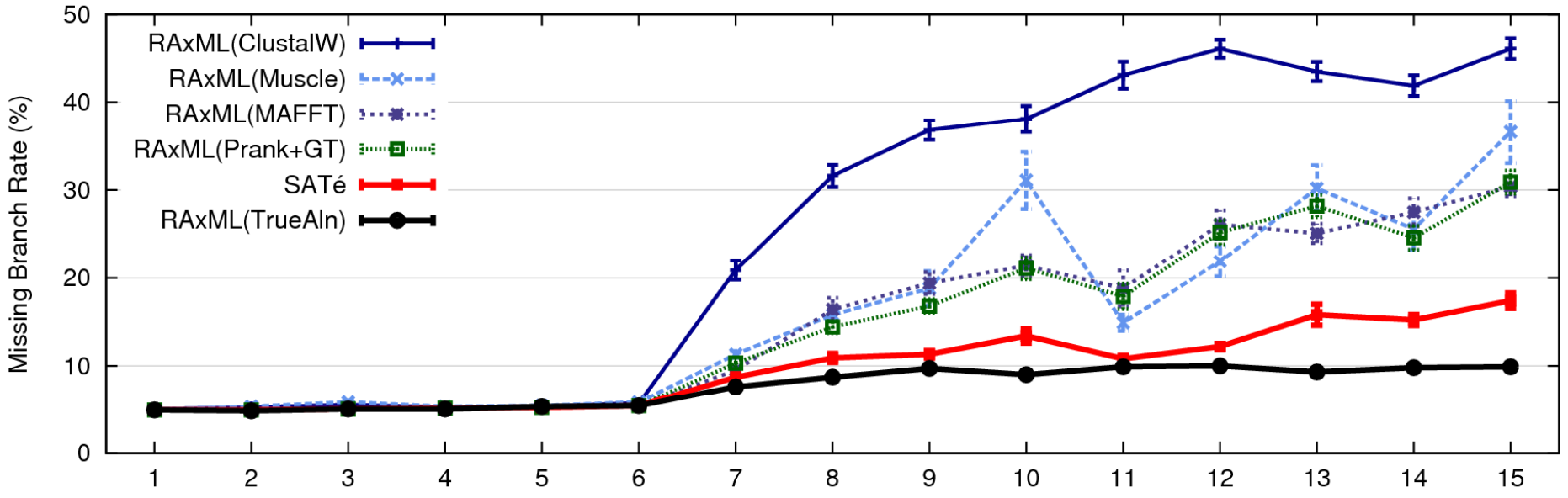
Repeat until termination condition (typically, 24 hours)

One SATé iteration (really 32 subsets)





1000 taxon models, ordered by difficulty



1000 taxon models, ordered by difficulty

24 hour SATé analysis, on desktop machines

(Similar improvements for biological datasets)

Why does SATé work well?

- We have proven that optimizing ML (treating gaps as missing data) is completely uninformative (all trees are optimal).

- Reconsidering why **SATé** works well led to a new divide-and-conquer strategy, **SATé-II,** with even better accuracy and speed.

Liu et al., Systematic Biology (in press)

Negative result

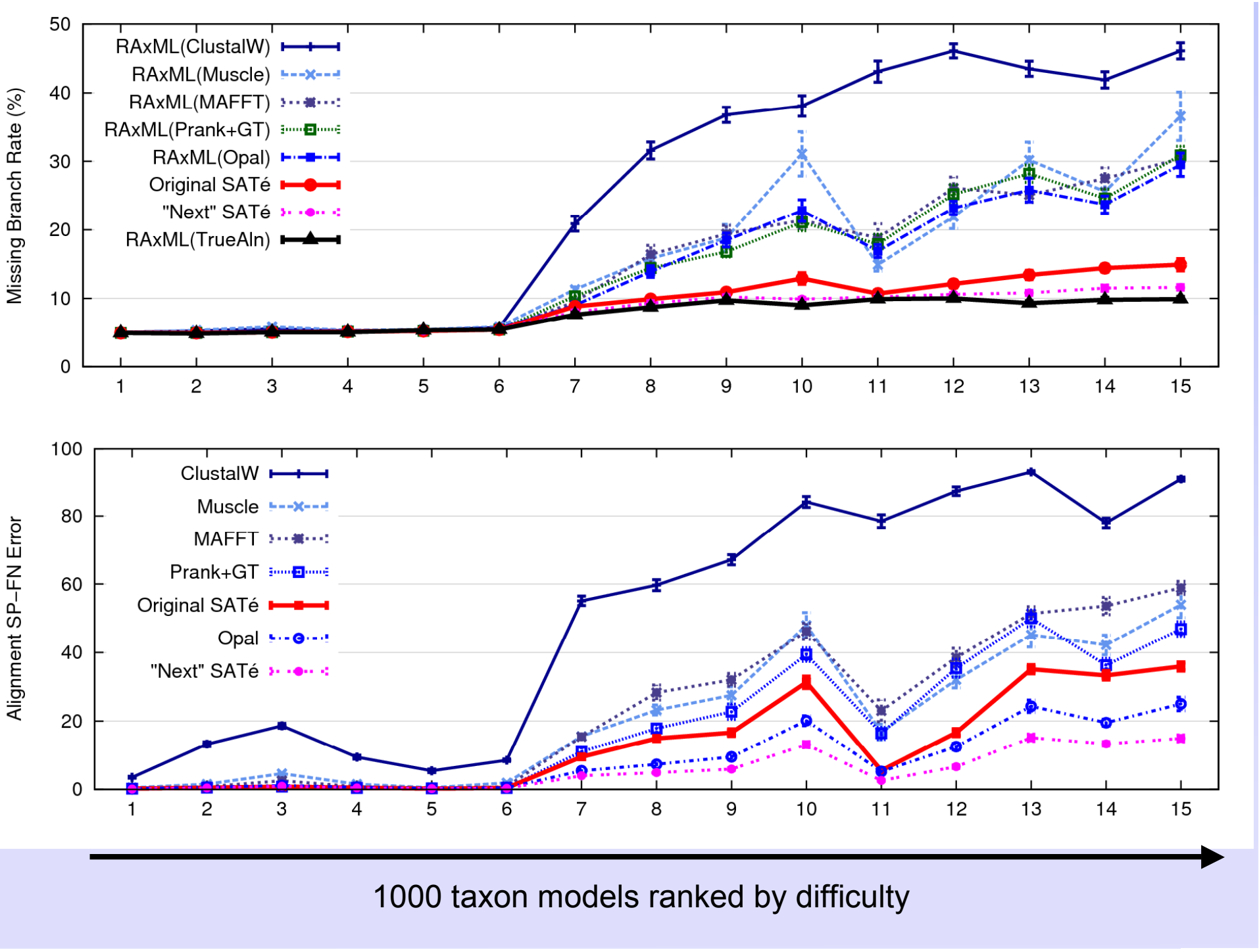
- Optimization problem: given sequence set S (unaligned), find alignment A and Jukes-Cantor model tree (T,θ) for S such that Pr(A|T,θ) is maximized (treating gaps as missing data).

- Theorem: For any input S of unaligned sequences and for all trees T there is an alignment A and set θ of Jukes-Cantor parameters on T s.t. Pr(A|T,θ) is maximum.

Thus, all trees T can optimize the maximum likelihood score.

Understanding SATé

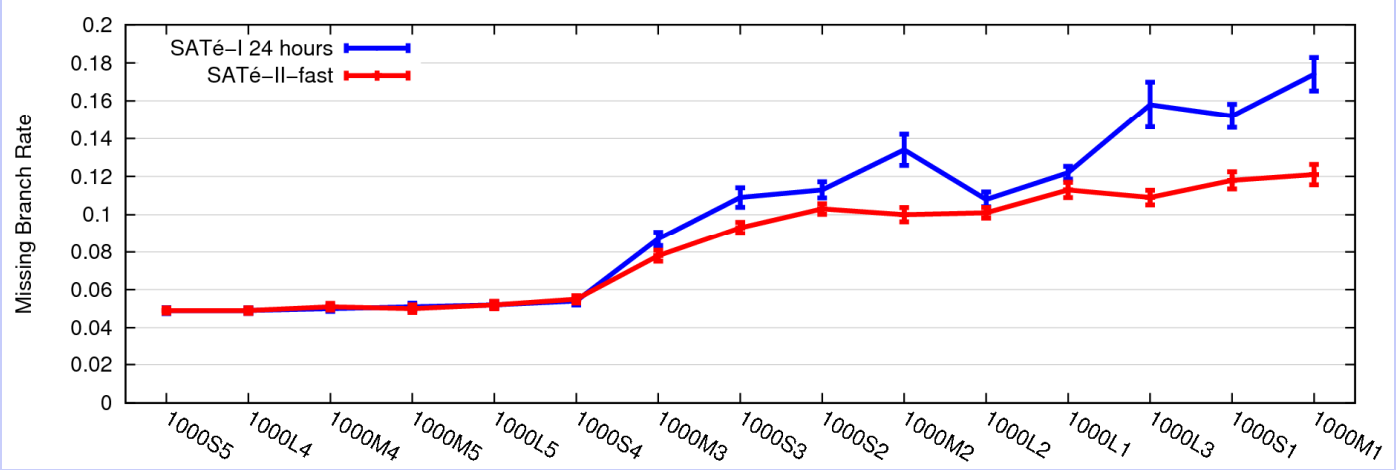
- Observations: (1) subsets of taxa that are small enough, closely related, and densely sampled are aligned more accurately than others.
- SATé-1 produces subsets that are closely related and densely sampled, but not small enough.
- SATé-2 (“next SATé”) changes the design to produce smaller subproblems.
- The next iteration starts with a more accurate tree. This leads to a better alignment, and a better tree.

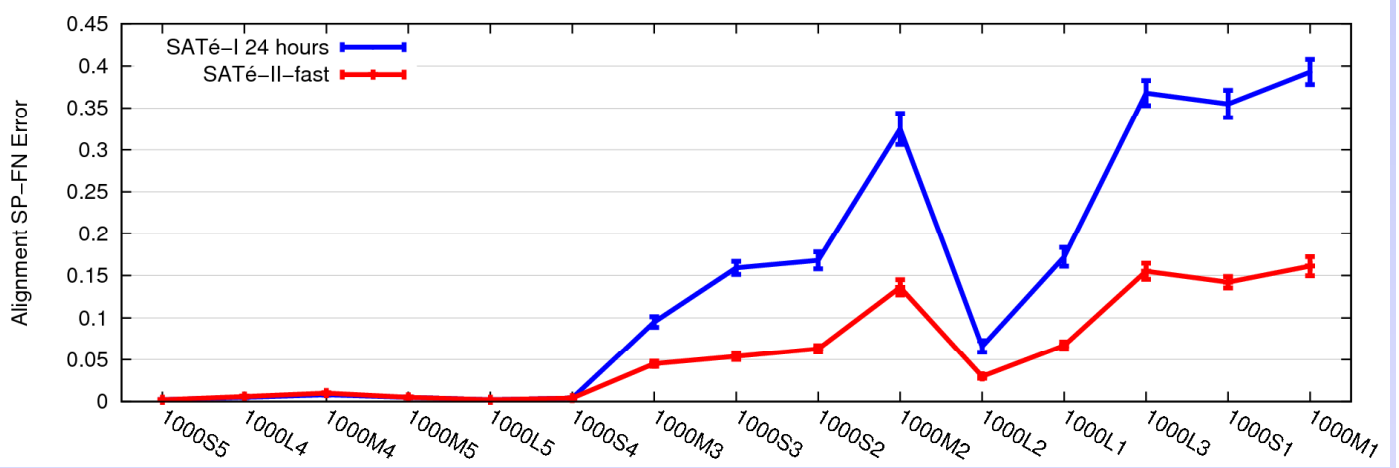

Missing Branch Rate (%)
RAxML(ClustalW)
RAxML(Muscle)
RAxML(MAFFT)
RAxML(Prank+GT)
RAxML(Opal)
Original SATé
"Next" SATé
RAxML(TrueAln)
Alignment SP–FN Error
ClustalW
Muscle
MAFFT
Prank+GT
Original SATé
Opal
"Next" SATé
1000 taxon models ranked by difficulty

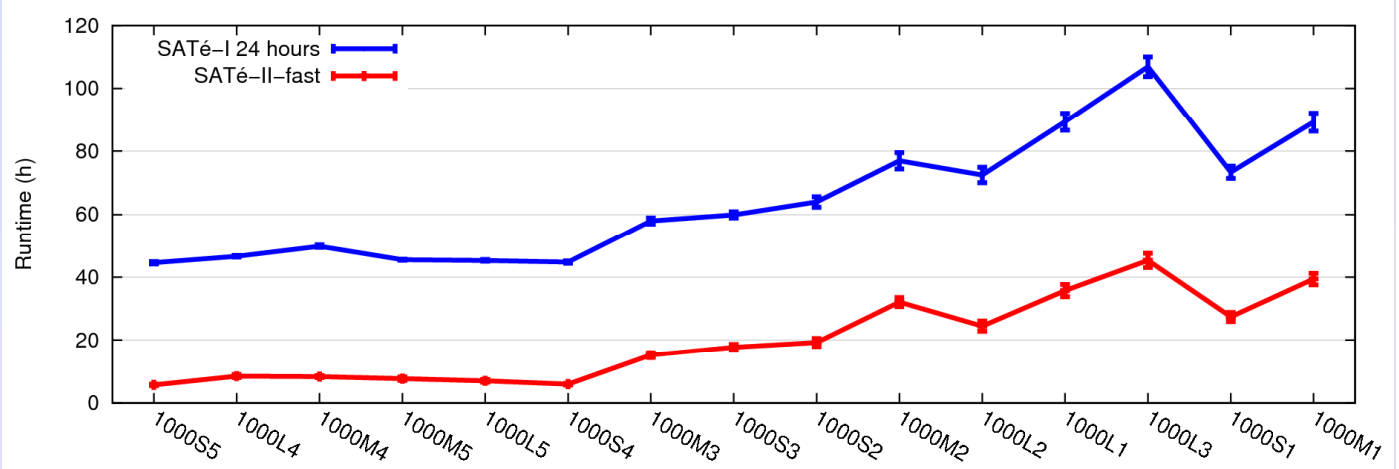
SATé-I vs. SATé-II

SATé-II

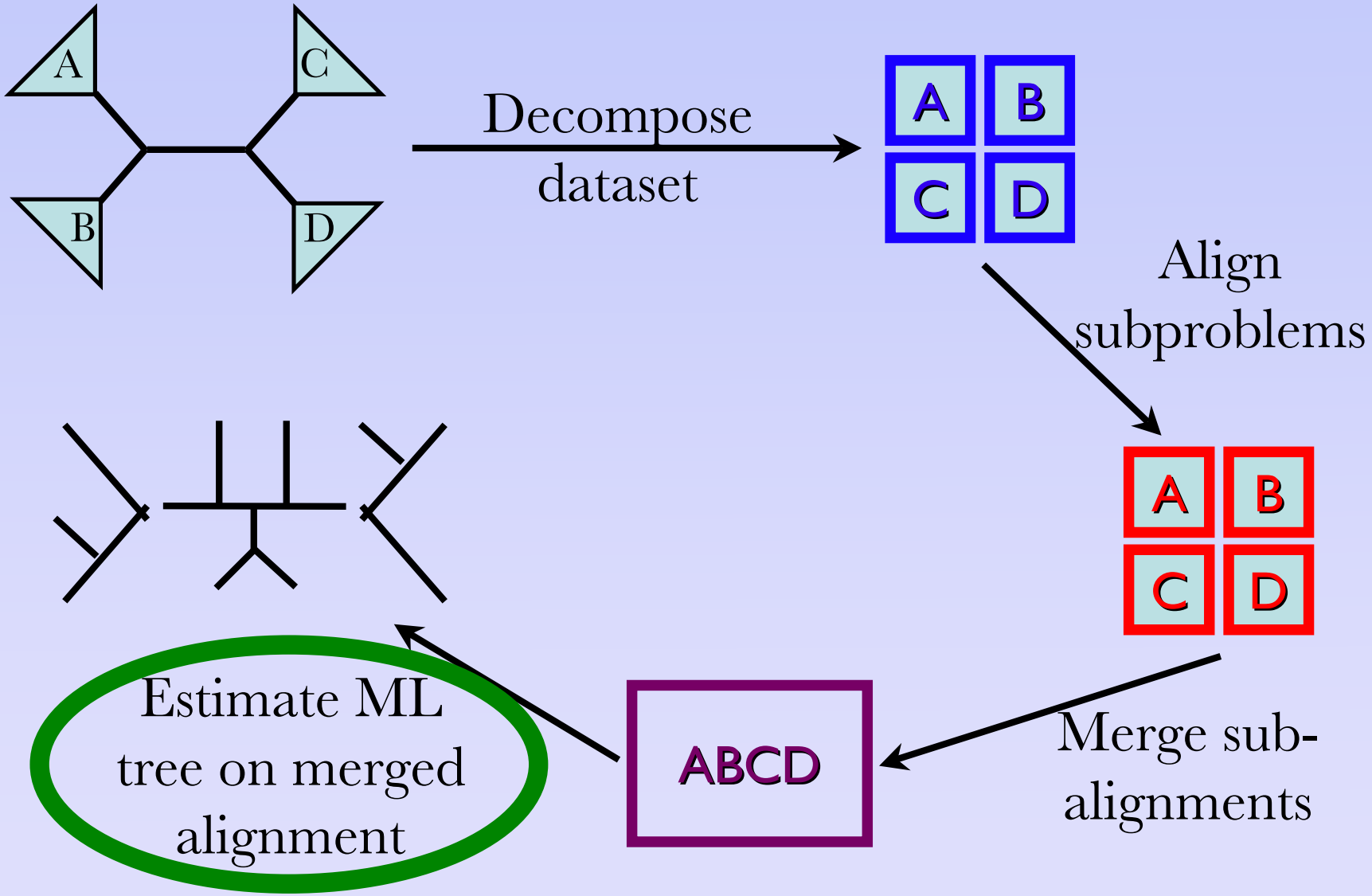
- **Faster** and more accurate than **SATé-I**
- Longer analyses or use of ML to select tree/alignment pair slightly better results







Limitations of SATé-I and -II



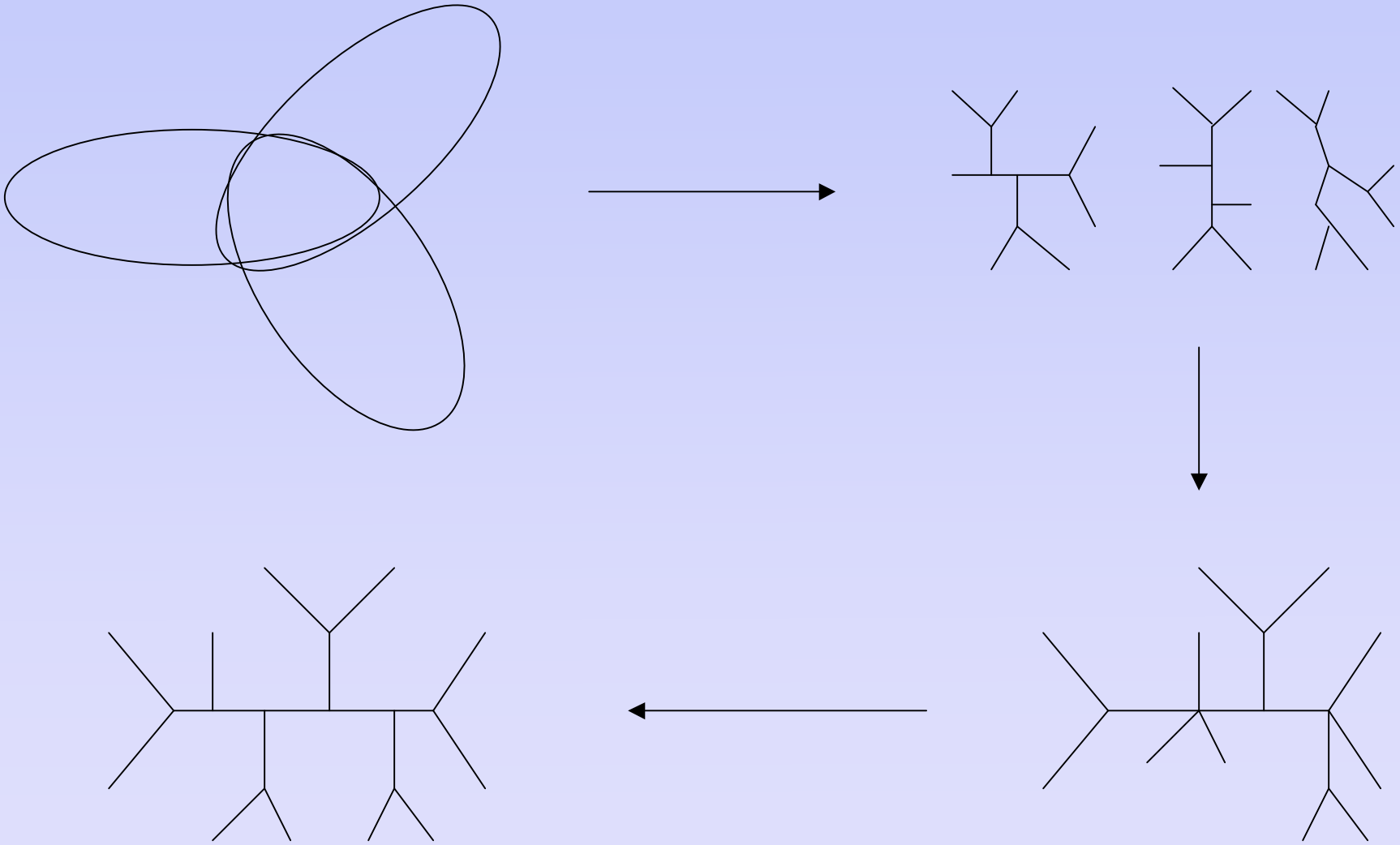
Part II: DACTAL

(Divide-And-Conquer Trees (without) ALignments)

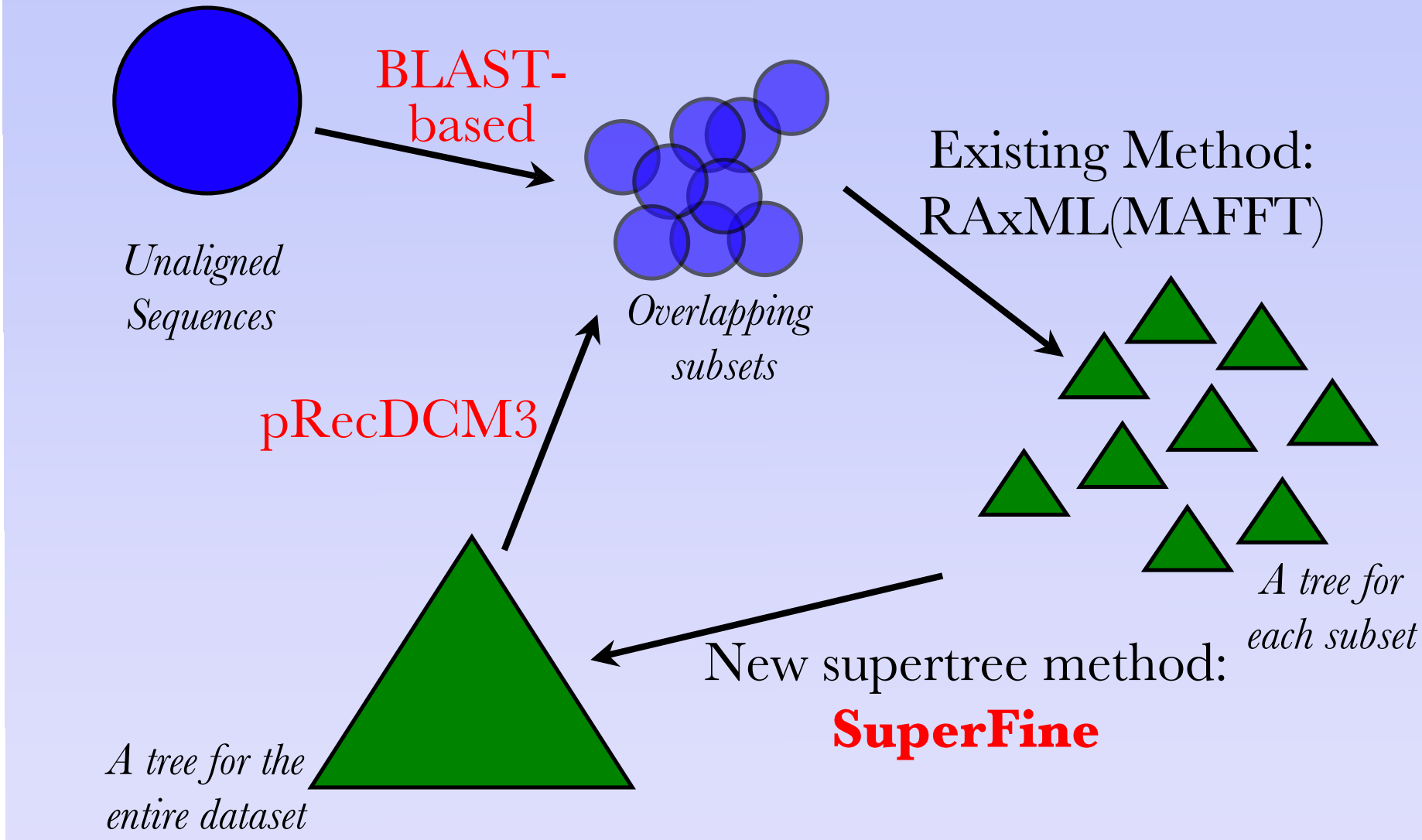
- Input: set S of unaligned sequences
- Output: tree on S (but no alignment)

(Nelesen, Liu, Wang, Linder, and Warnow, submitted)

Disk-Covering Methods (DCM) (starting in 1998)



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 outperforms SATé

- DACTAL faster and matches or improves upon accuracy of SATé for datasets with 1000 or more taxa

- The biggest gains are on the very large datasets

Average of 3 Largest CRW Datasets

CRW: Comparative RNA database, datasets 16S.B.ALL, 16S.T, and 16S.3

6,323 to 27,643 sequences

These datasets have curated alignments based on secondary structure

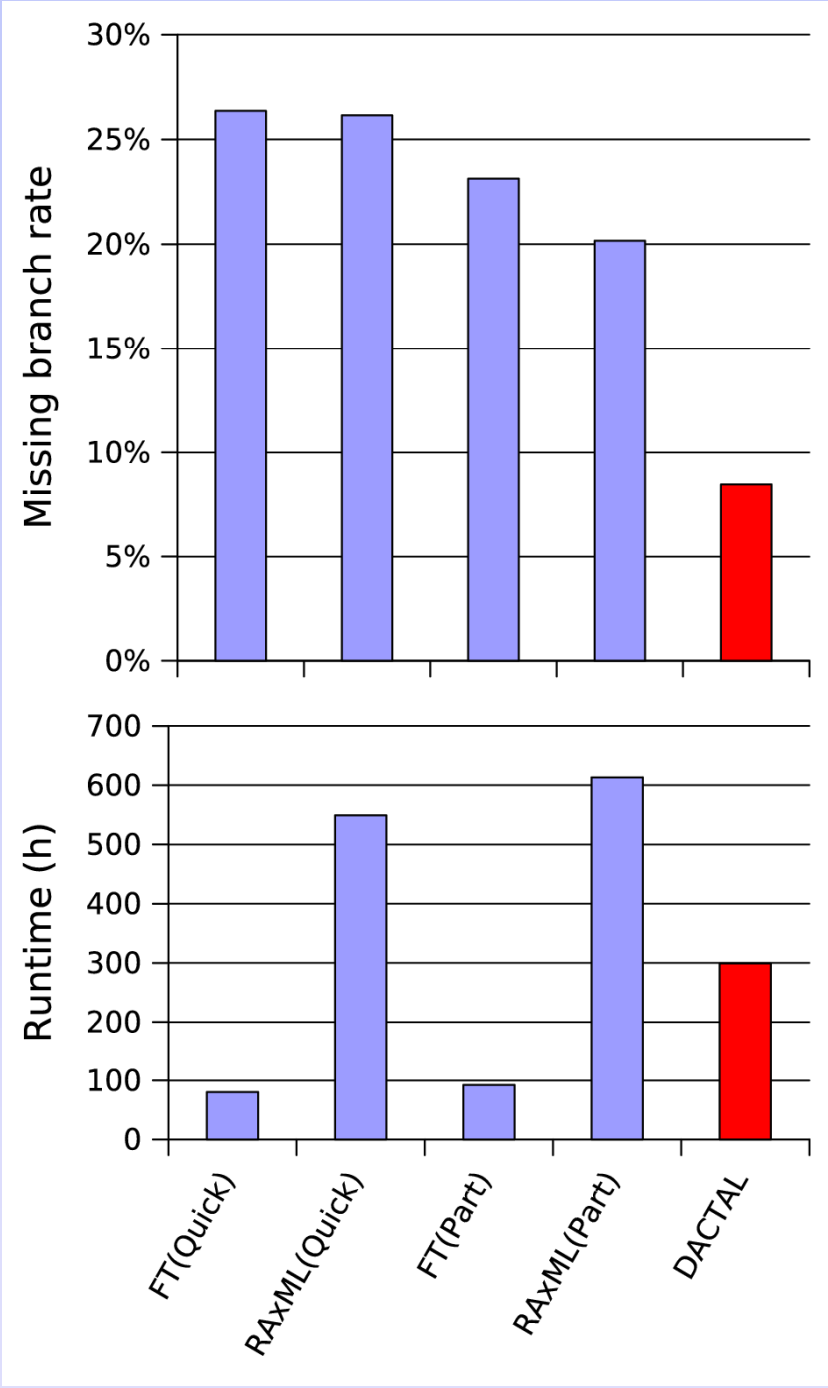
Reference trees are 75% RAxML bootstrap trees

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

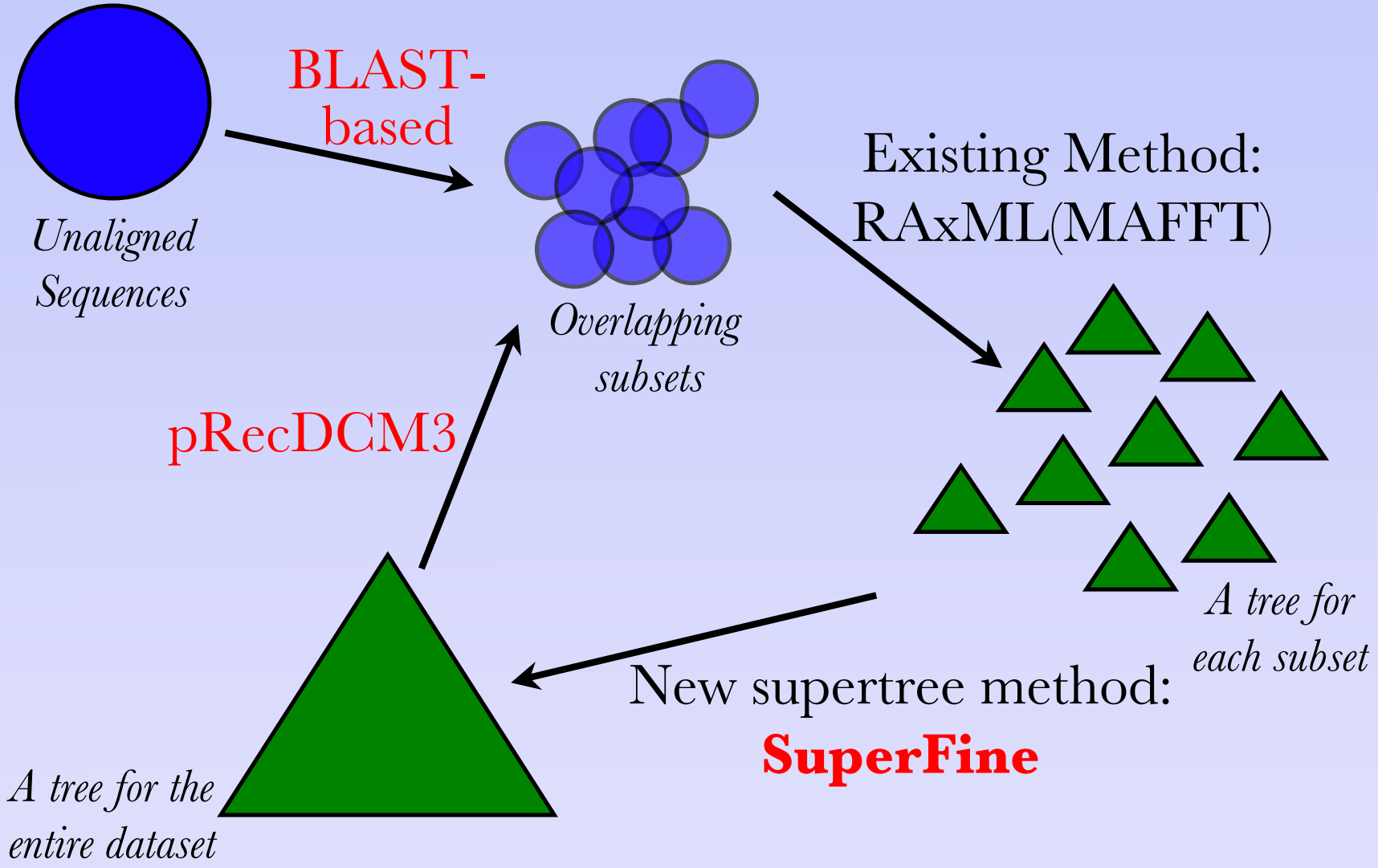
DACTAL is robust to starting trees

PartTree and Quicktree are the only MSA methods that run on all 3 datasets

FastTree (FT) and RAxML are ML methods



Why DACTAL performs well



Implications

- Standard alignment methods do not provide adequate accuracy on large datasets.
- When markers tend to yield poor alignments and trees, develop better methods - don’t throw out the data.
- Phylogenetic analyses of large (10,000 taxa and up) need new techniques.

Other Research Projects

- Supertree methods
- Faster maximum likelihood methods
- Datamining sets of trees and alignments
- Visualization of ultra-large trees and multiple sequence alignments
- Comparative genomics: whole genome phylogeny using gene order and content
- Estimating species trees from gene trees
- Reticulate phylogeny detection and estimation

Acknowledgments

- Microsoft Research New England
- National Science Foundation: Assembling the Tree of Life (ATOL), ITR, and IGERT grants (0733029, 0331453, 0114387)
- Current and former students: Kevin Liu, Luay Nakhleh, Serita Nelesen, Sindhu Raghavan, Jerry Sun, Rahul Suri, Shel Swenson, and Li-San Wang
- Collaborators
  - Randy Linder, Integrative Biology, UT-Austin
  - Mark Holder and Jiaye Yu (Kansas) for software at http://phylo.bio.ku.edu/software/sate/sate.html