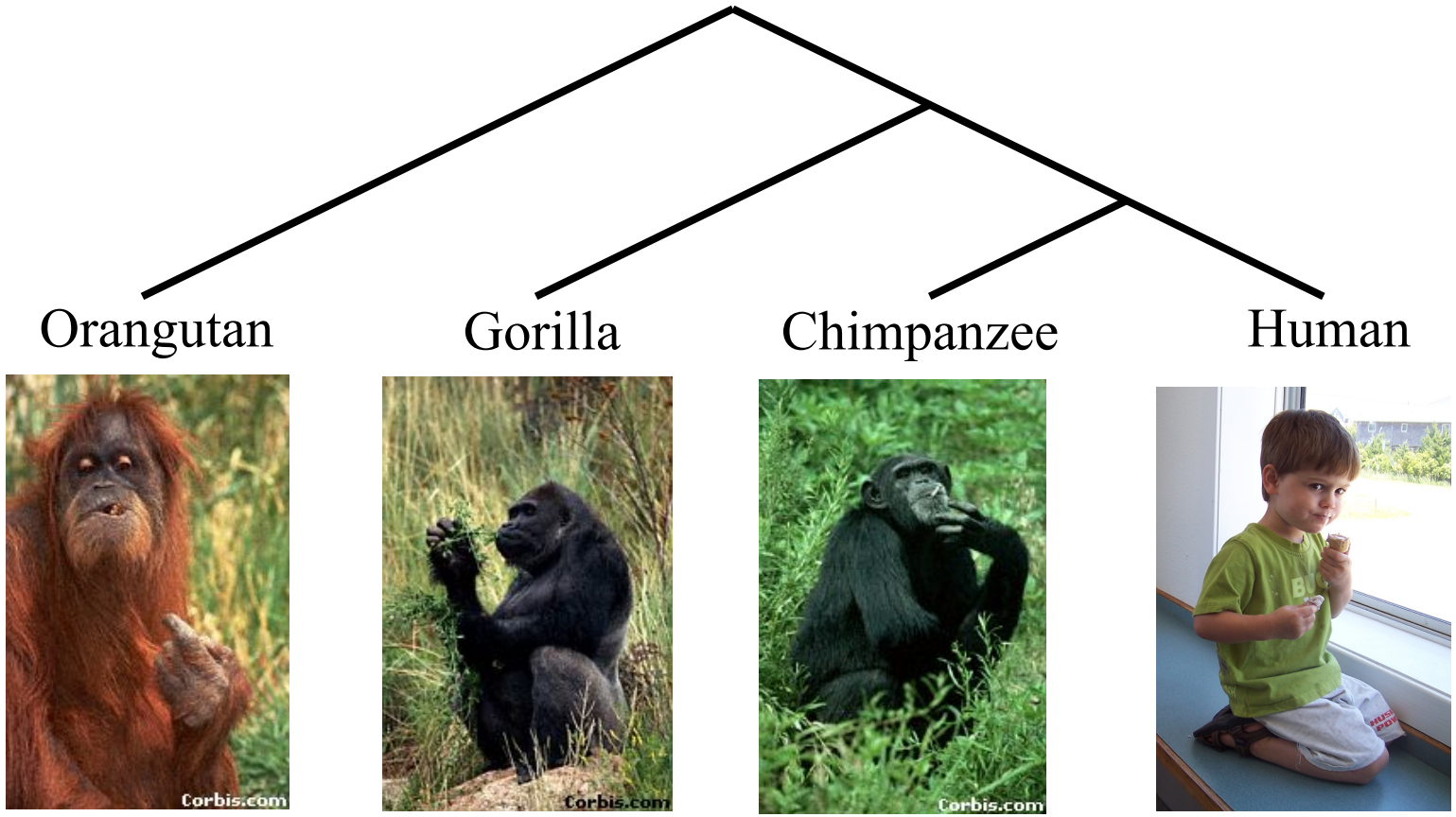
Ultra-large phylogeny estimation

Tandy Warnow
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

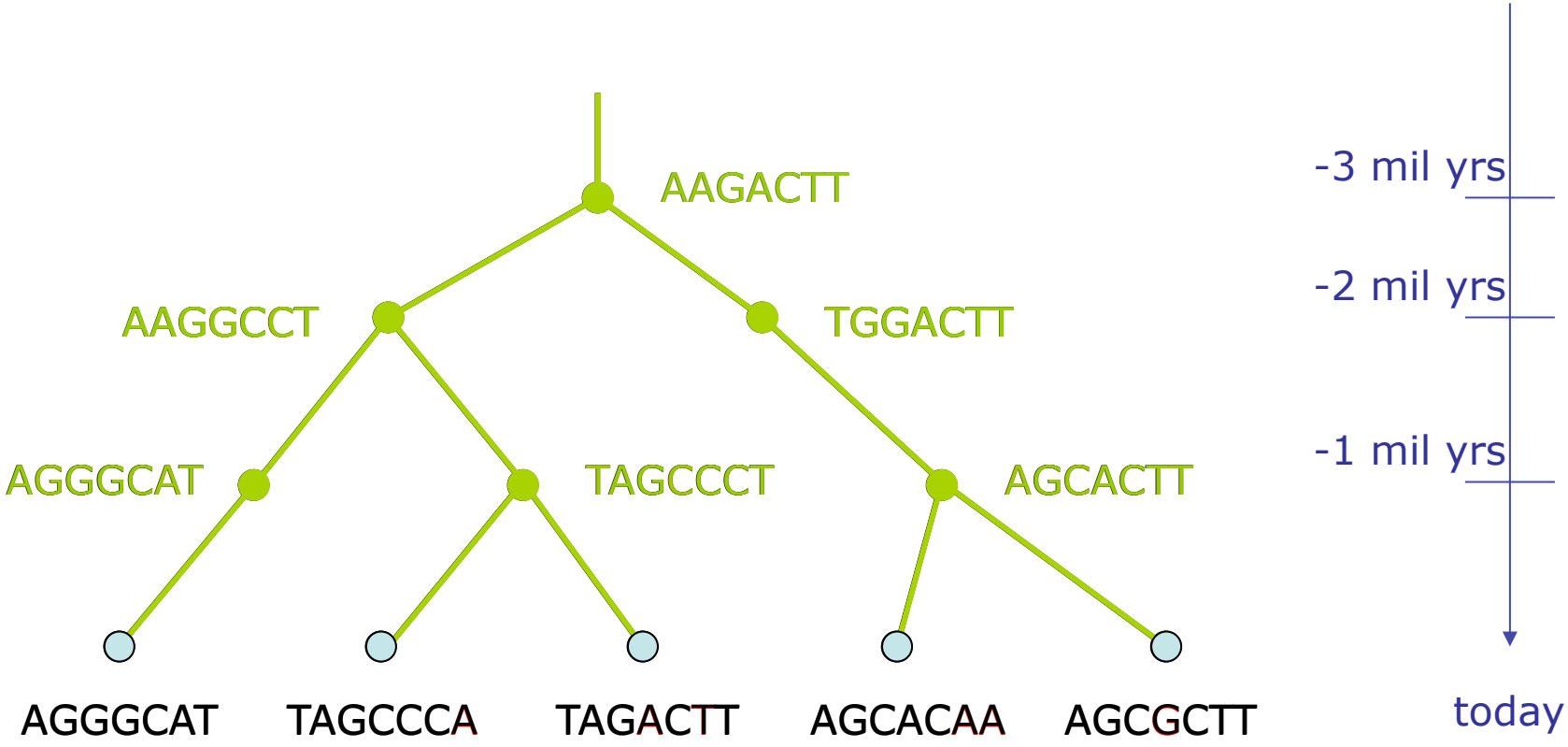
NP-hard optimization problems
Graph-theory
Stochastic models of evolution
Statistical methods
Statistical performance issues

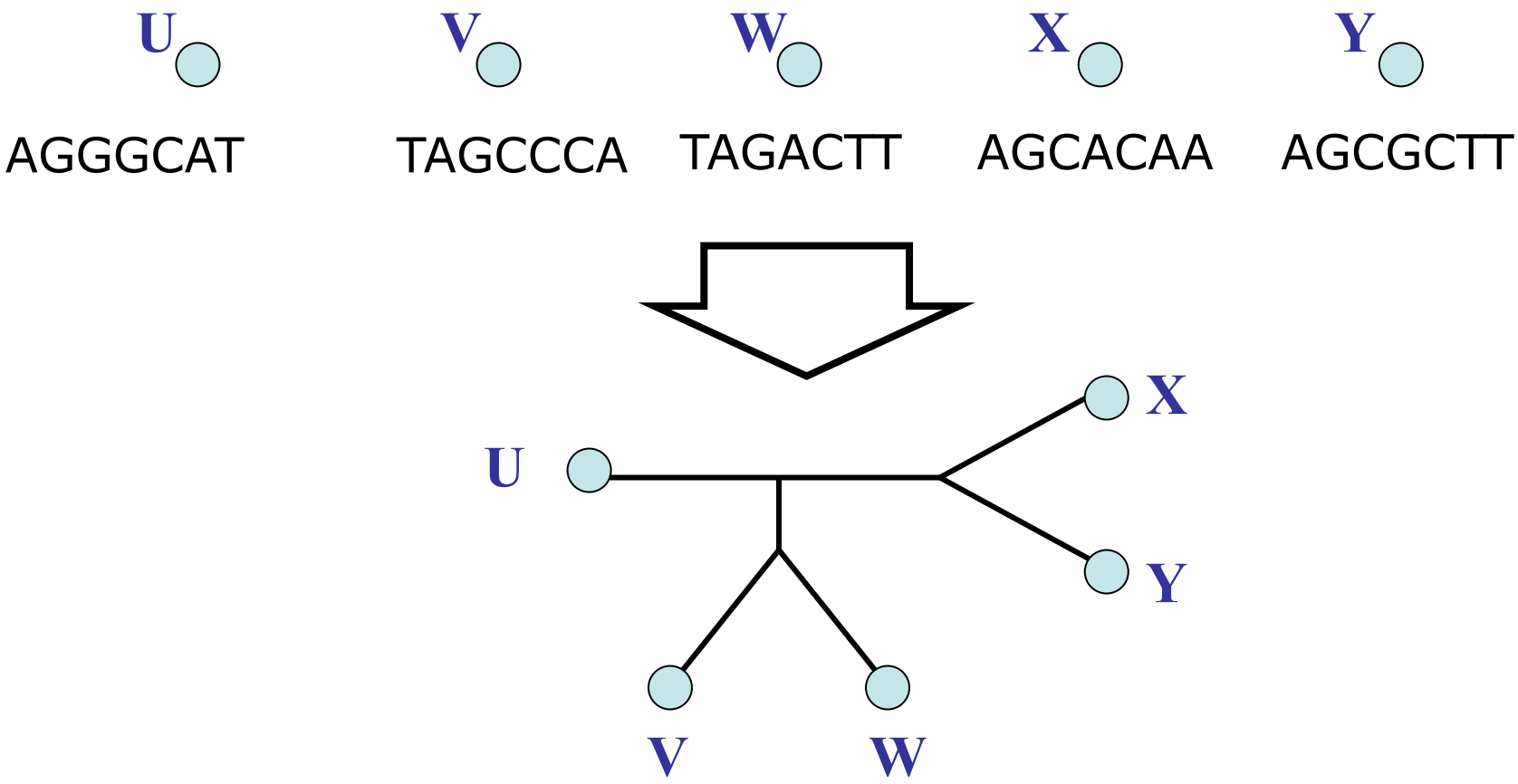
Millions of taxa

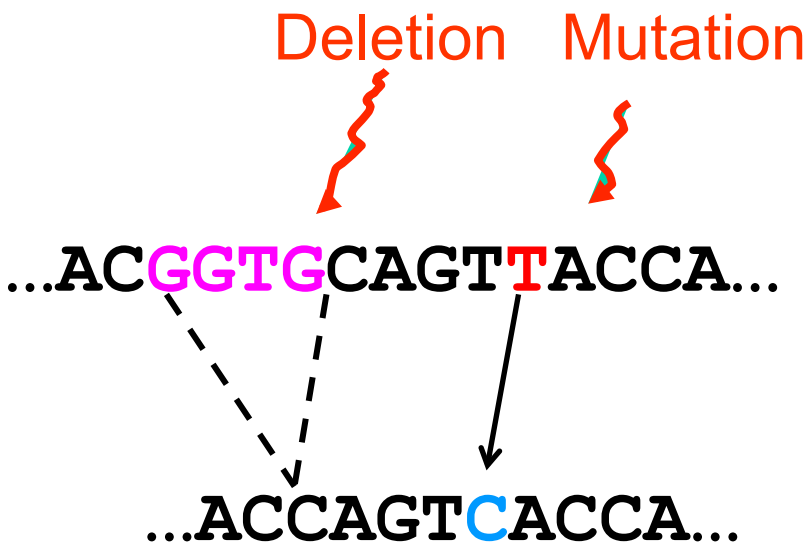
Important applications

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



...ACGGTGCAGTTACCA...

...AC----CAGTCACCA...

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 (MSA)

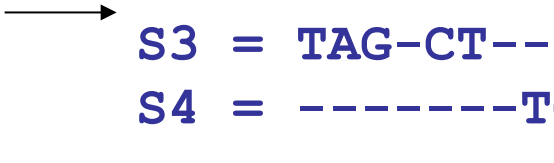
```
S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC
S3 = TAGCTGACCGC
S4 = TCACGACCGACA
```



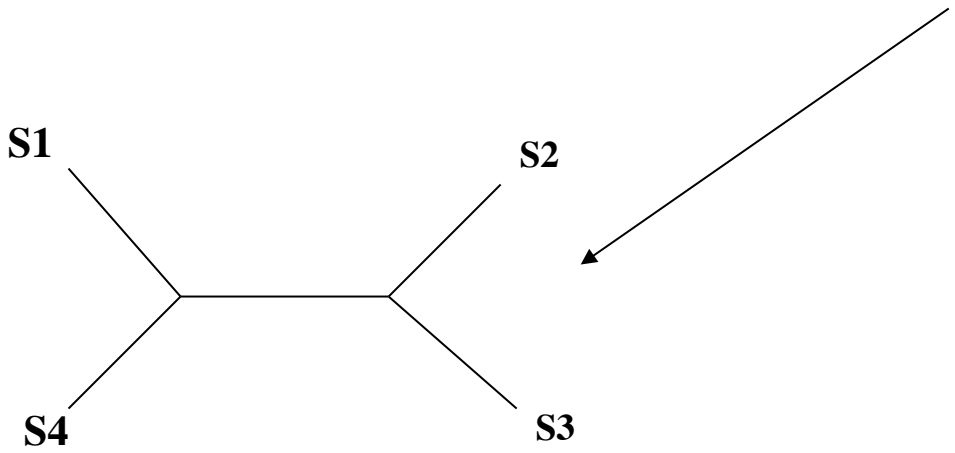
```
S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAC--GACCGC--
S3 = TAG-CT-------GACCGC--
S4 = -------TCAC--GACCGACA
```

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
```



Major Challenges

- Current phylogenetic datasets contain hundreds to thousands of taxa, with multiple genes. Future datasets will be substantially larger.
- Poor MSAs result in poor trees, and standard MSA methods are poor on large datasets.
- Statistical estimation methods produce better results, but are computationally intensive.

Large datasets beyond the scope of standard approaches.

Phylogenetic “boosters” (meta-methods)

Goal: improve accuracy, speed, robustness, or theoretical guarantees of base methods

Examples:

- DCM-boosting for distance-based methods
- DCM-boosting for heuristics for NP-hard problems
- SATé-boosting for alignment methods
- SuperFine-boosting for supertree methods
- DACTAL-boosting for all methods

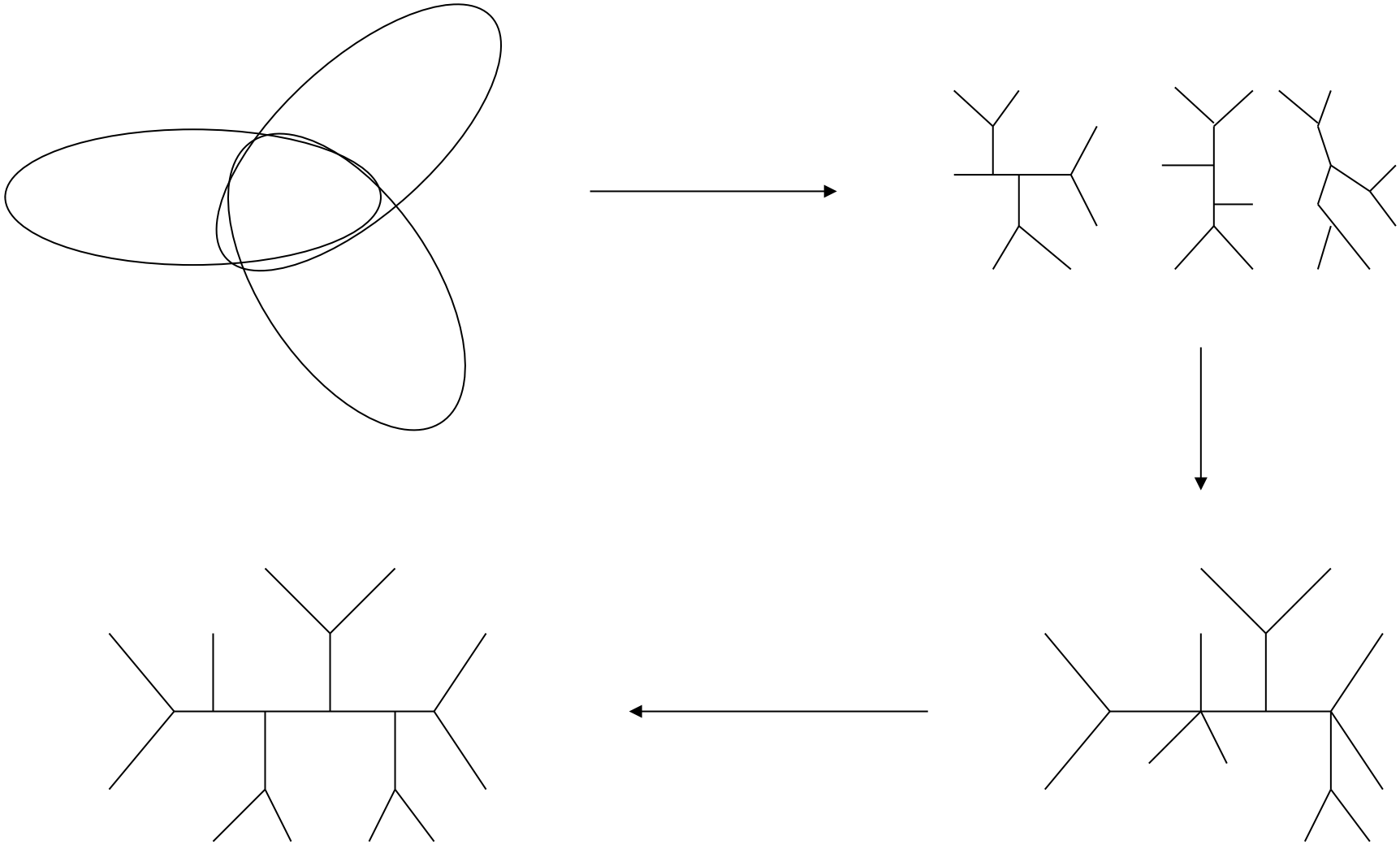
Phylogenetic “boosters” (meta-methods)

Goal: improve accuracy, speed, robustness, or theoretical guarantees of base methods

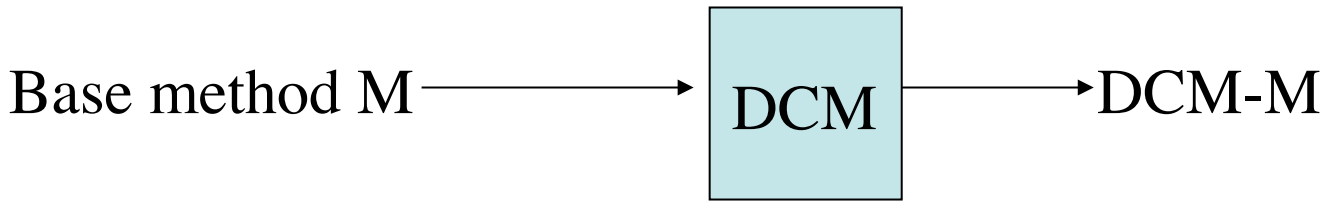
Examples:

- DCM-boosting for distance-based methods
- DCM-boosting for heuristics for NP-hard problems
- SATé-boosting for alignment methods
- SuperFine-boosting for supertree methods
- DACTAL-boosting for all methods

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

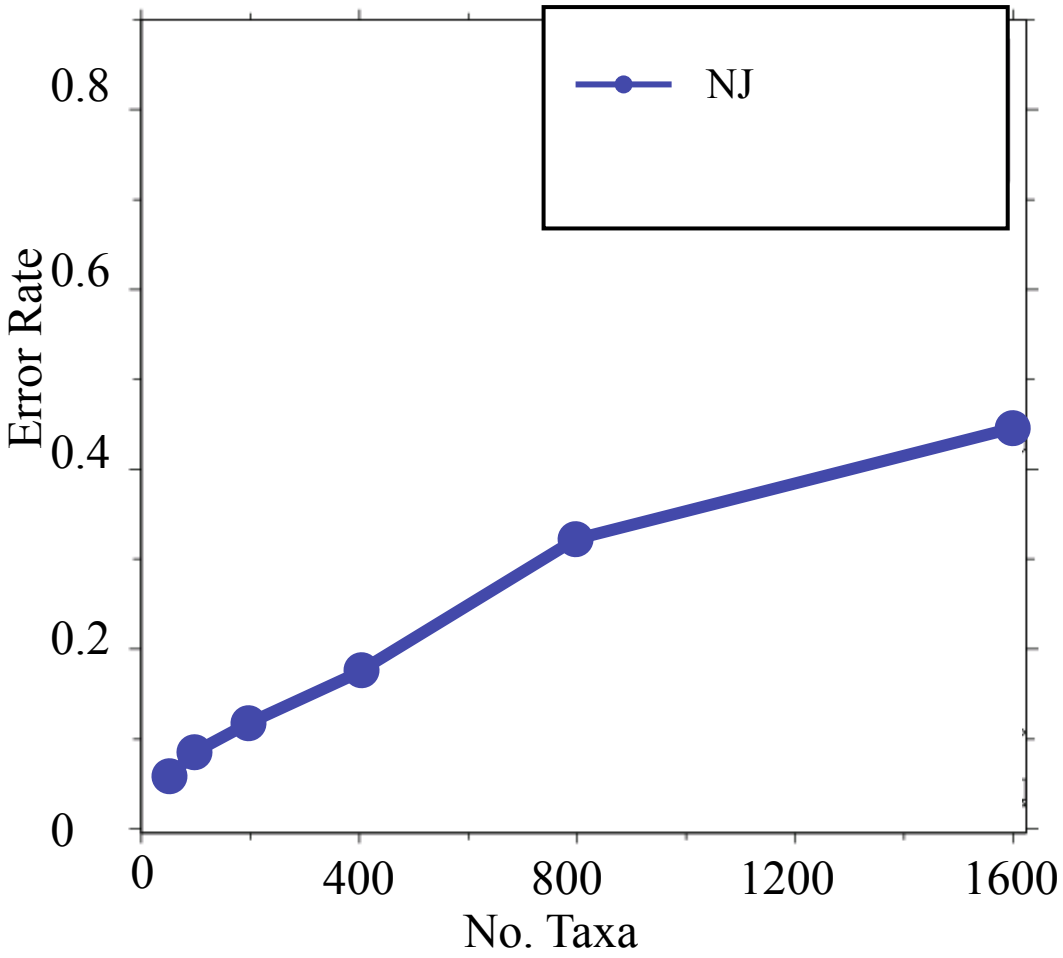


- DCMs “boost” the performance of phylogeny reconstruction methods.



Neighbor joining has poor accuracy on large diameter model trees

[Nakhleh et al. ISMB 2001]

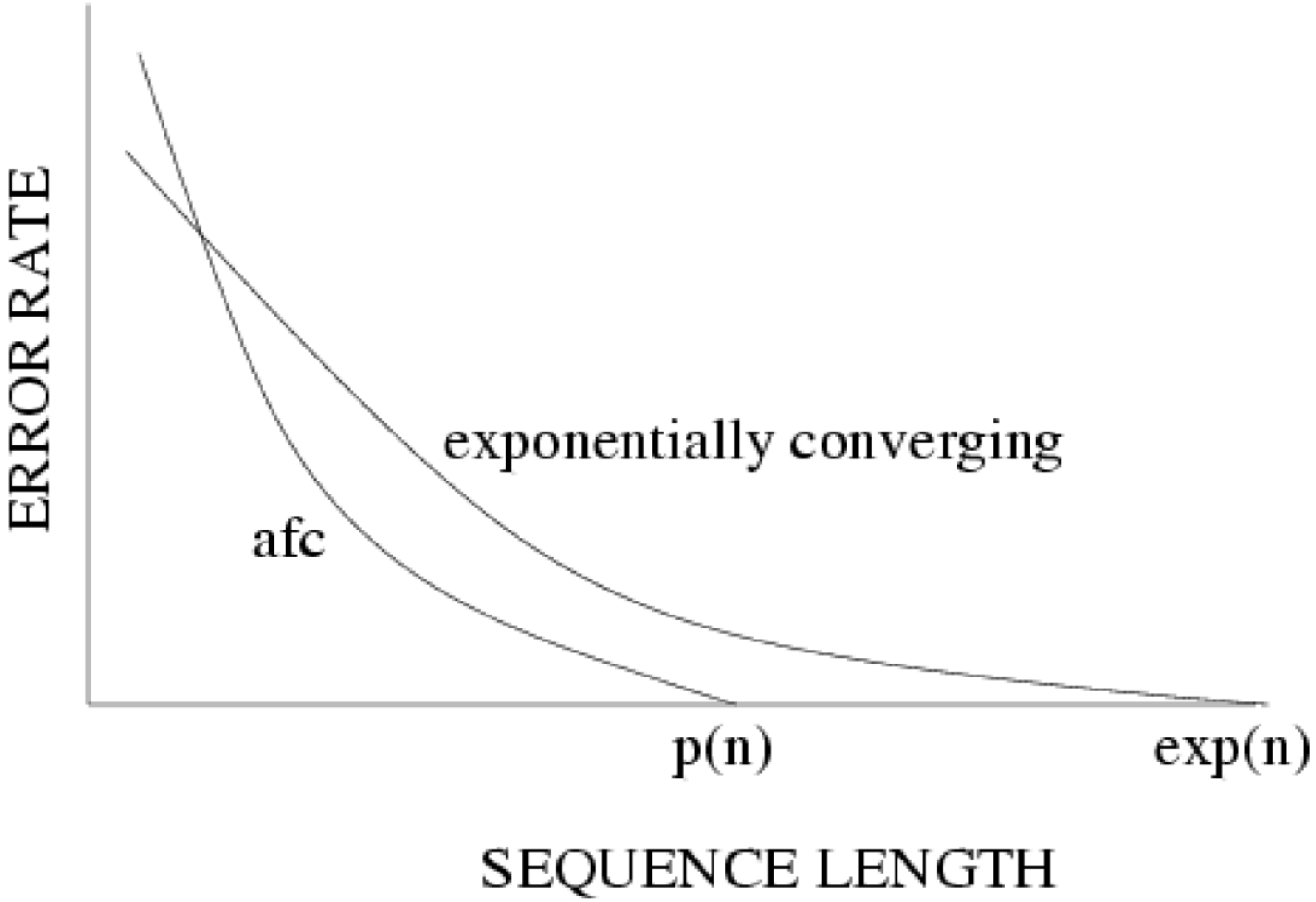


Simulation study

based upon fixed edge lengths, K2P model of evolution, sequence lengths fixed to 1000 nucleotides.

Error rates reflect proportion of incorrect edges in inferred trees.

Statistical consistency, convergence rates, and absolute fast convergence

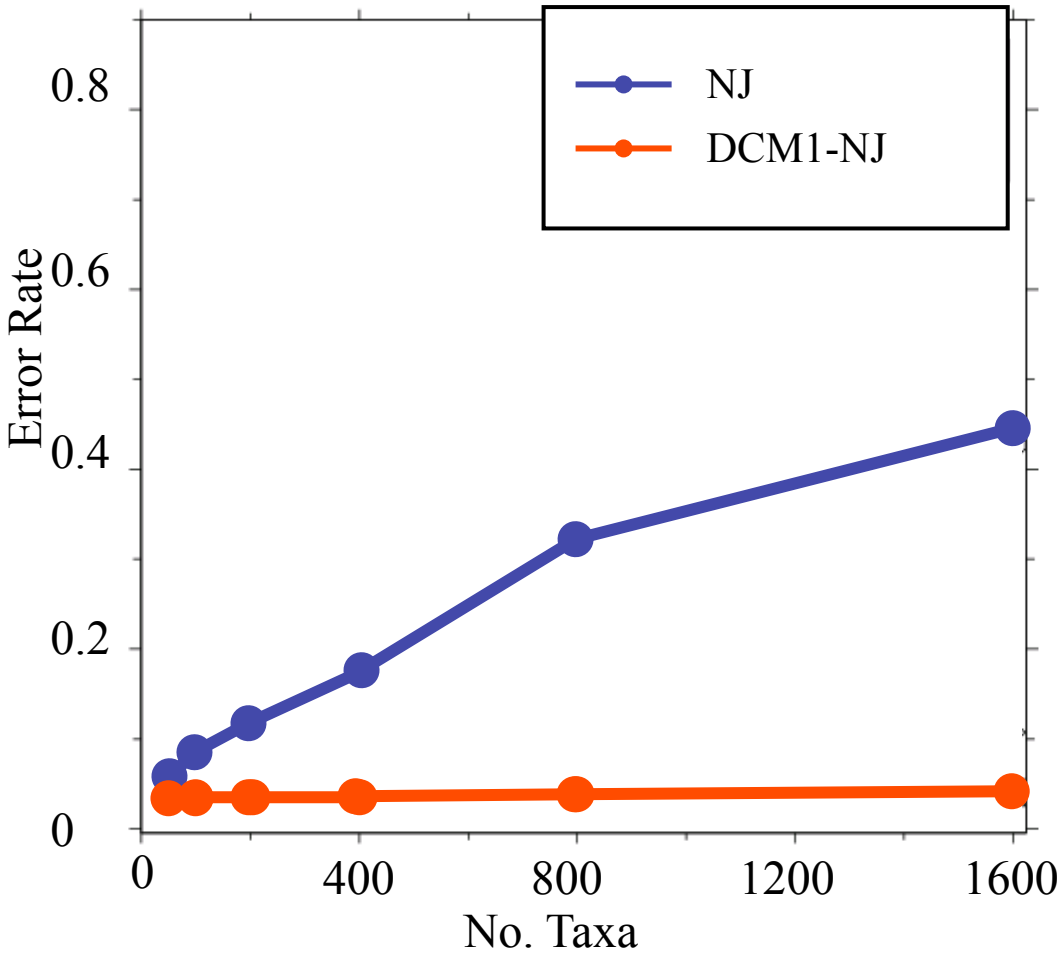


Neighbor Joining is consistent, but its sequence length requirement is exponential

- Atteson: Let T be a General Markov model tree defining additive matrix D. Then Neighbor Joining will reconstruct the true tree with high probability from sequences that are of length at least $O(\lg n \, e^{\max\{Dij\}})$.
- Note: $\max\{D_{ij}\} = O(g \, \mathrm{diam}(T))$, where g is the maximum length of any edge.

DCM1-boosting distance-based methods

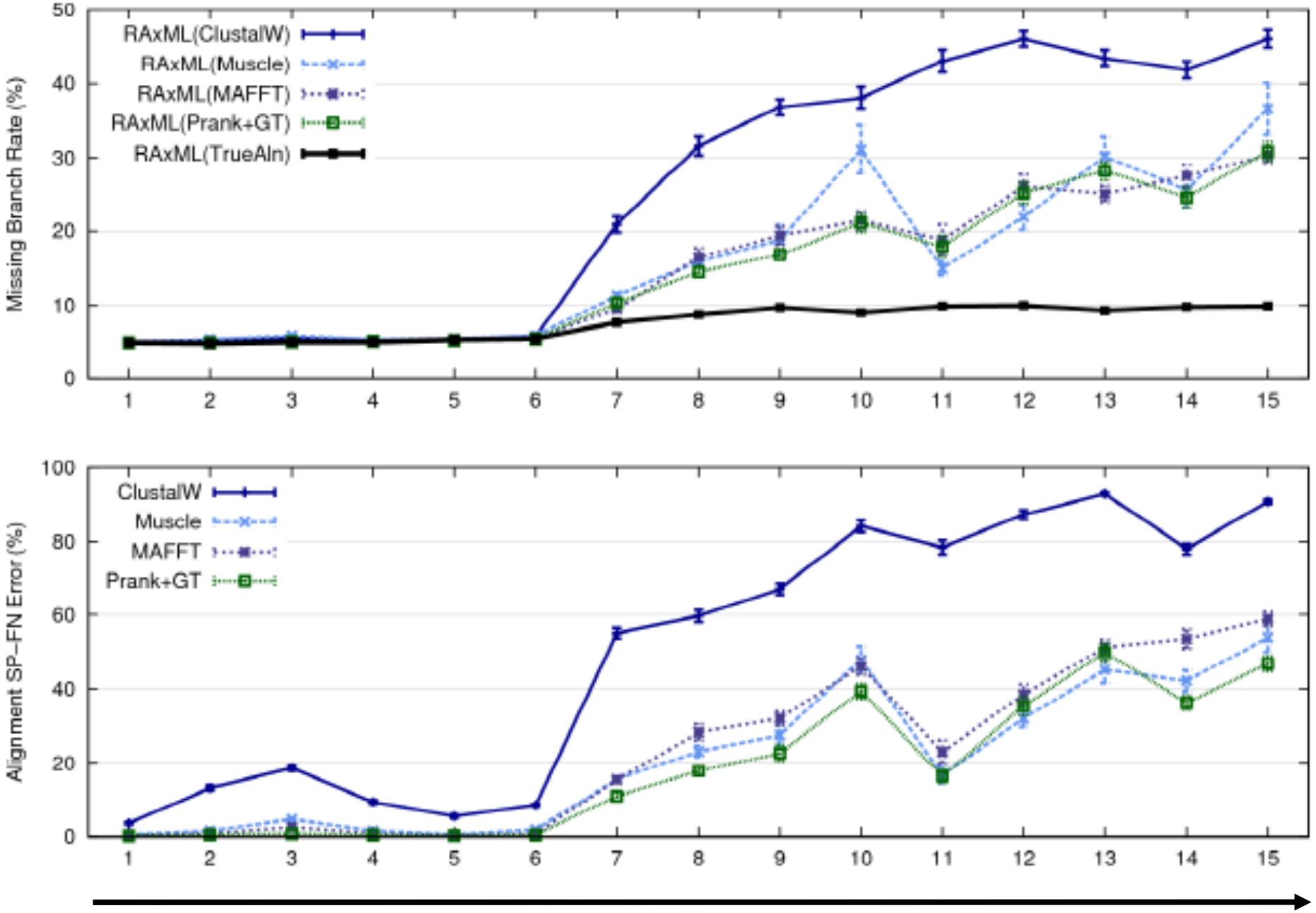
[Nakhleh et al. ISMB 2001 and Warnow et al. SODA 2001]



Theorem: DCM1-NJ converges to the true tree from polynomial length sequences

Alignment Estimation

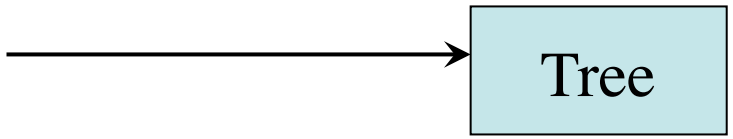
- Most phylogeny estimation methods assume that the true alignment is known.
- In fact, estimating the alignment is difficult, especially on large datasets.



1000 taxon models, ordered by difficulty (Liu et al., 2009)

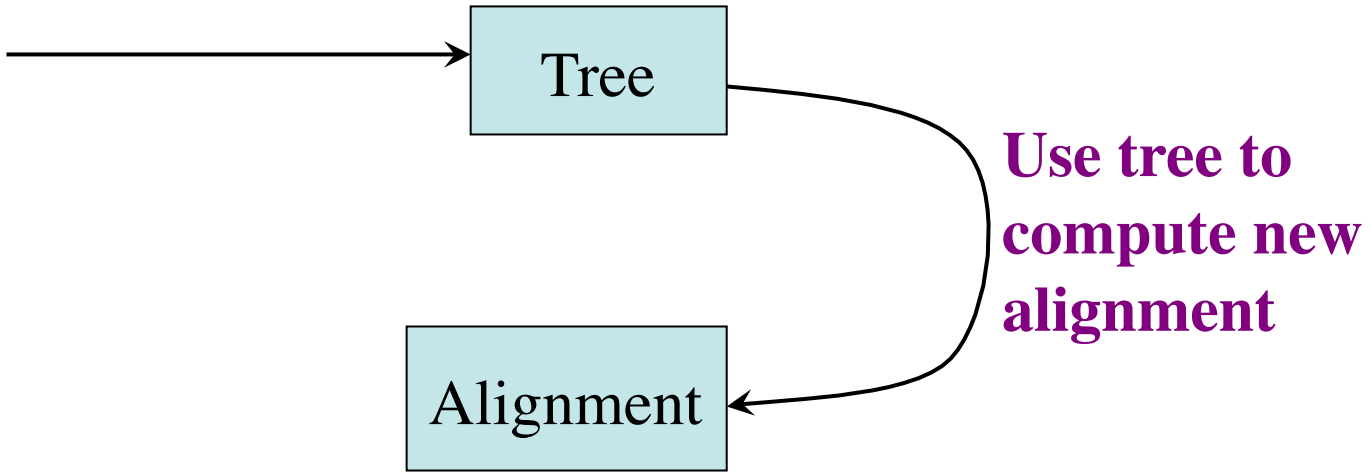
SATé Algorithm

Obtain initial alignment
and estimated ML tree

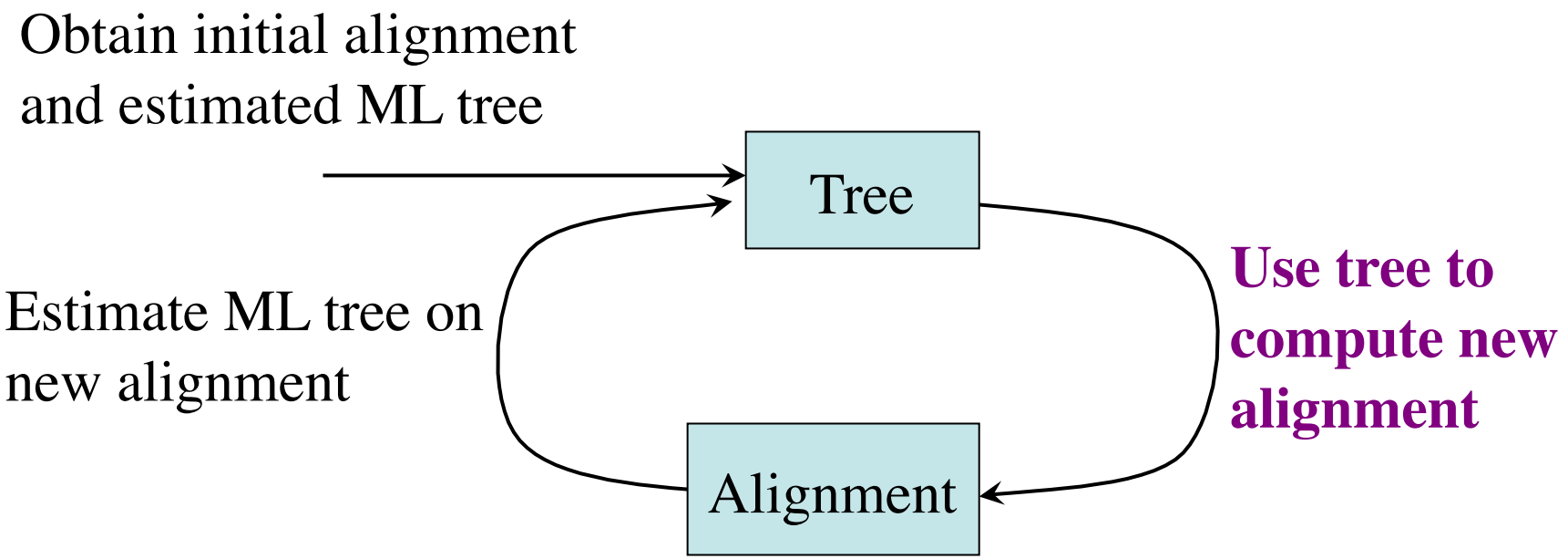


SATé Algorithm

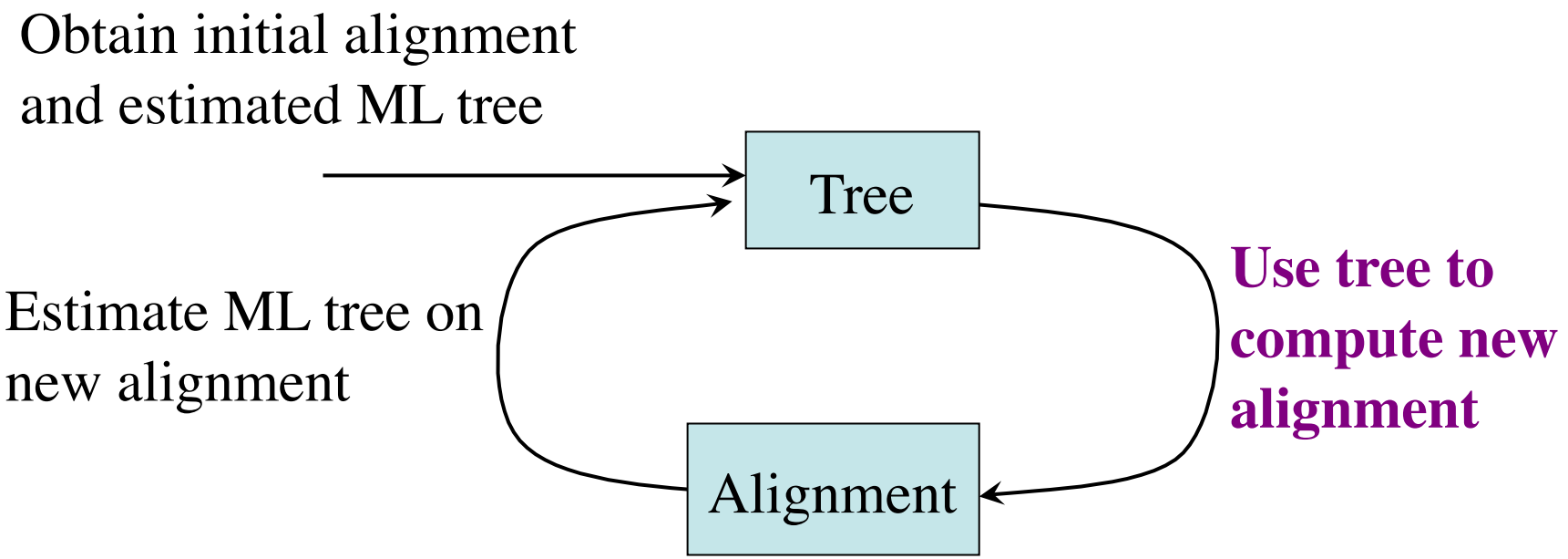
Obtain initial alignment
and estimated ML tree



SATé Algorithm



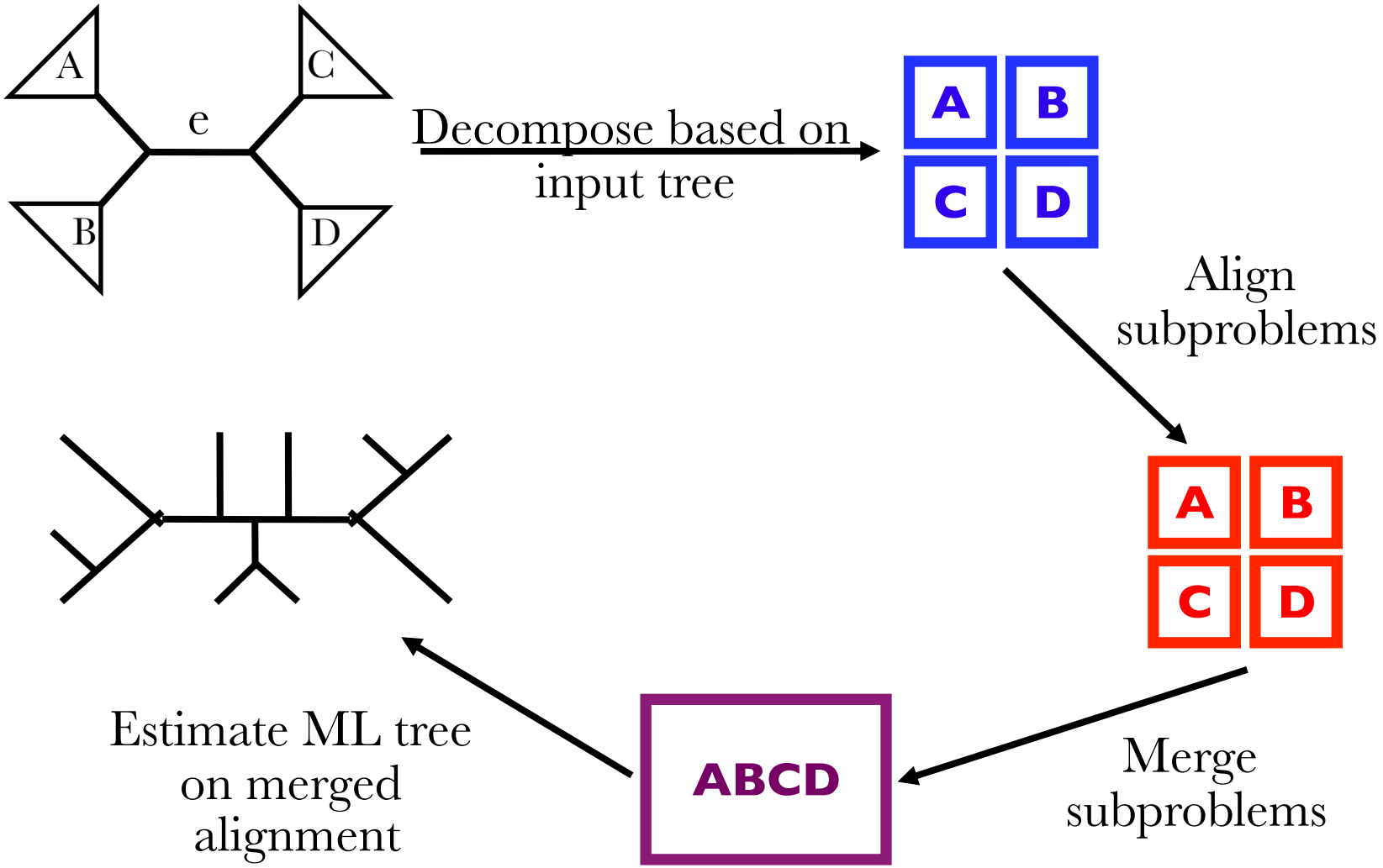
SATé Algorithm

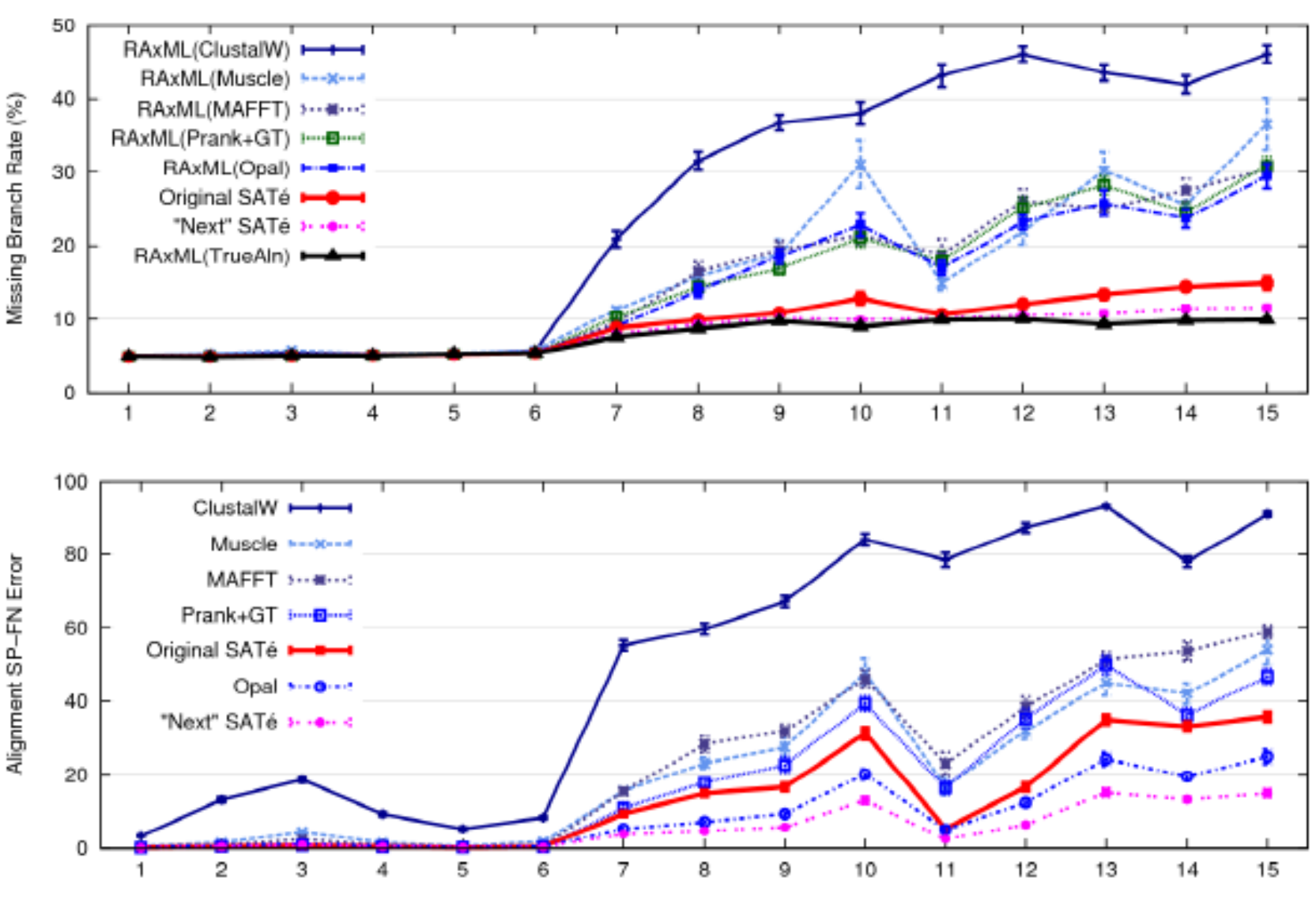


If new alignment/tree pair has worse ML score, realign using a different decomposition

Repeat until termination condition (typically, 24 hours)

One SATé iteration (cartoon)





SATé compared to two-phase methods on 1000 taxon datasets
Original SAT**é:** Liu et al. Science 2009
“Next” SAT**é**: Liu et al., Systematic Biology (in press)

SATé-boosting alignment methods

- Current alignment methods fail to return reasonable alignments on large datasets with high rates of indels and substitutions.
- SATé is an iterative divide-and-conquer technique that improves the accuracy of alignment methods.
- Trees computed on SATé-boosted alignments are much more accurate.
- SATé is very fast, finishing on 1000 taxon datasets on desktop computers in a few hours.

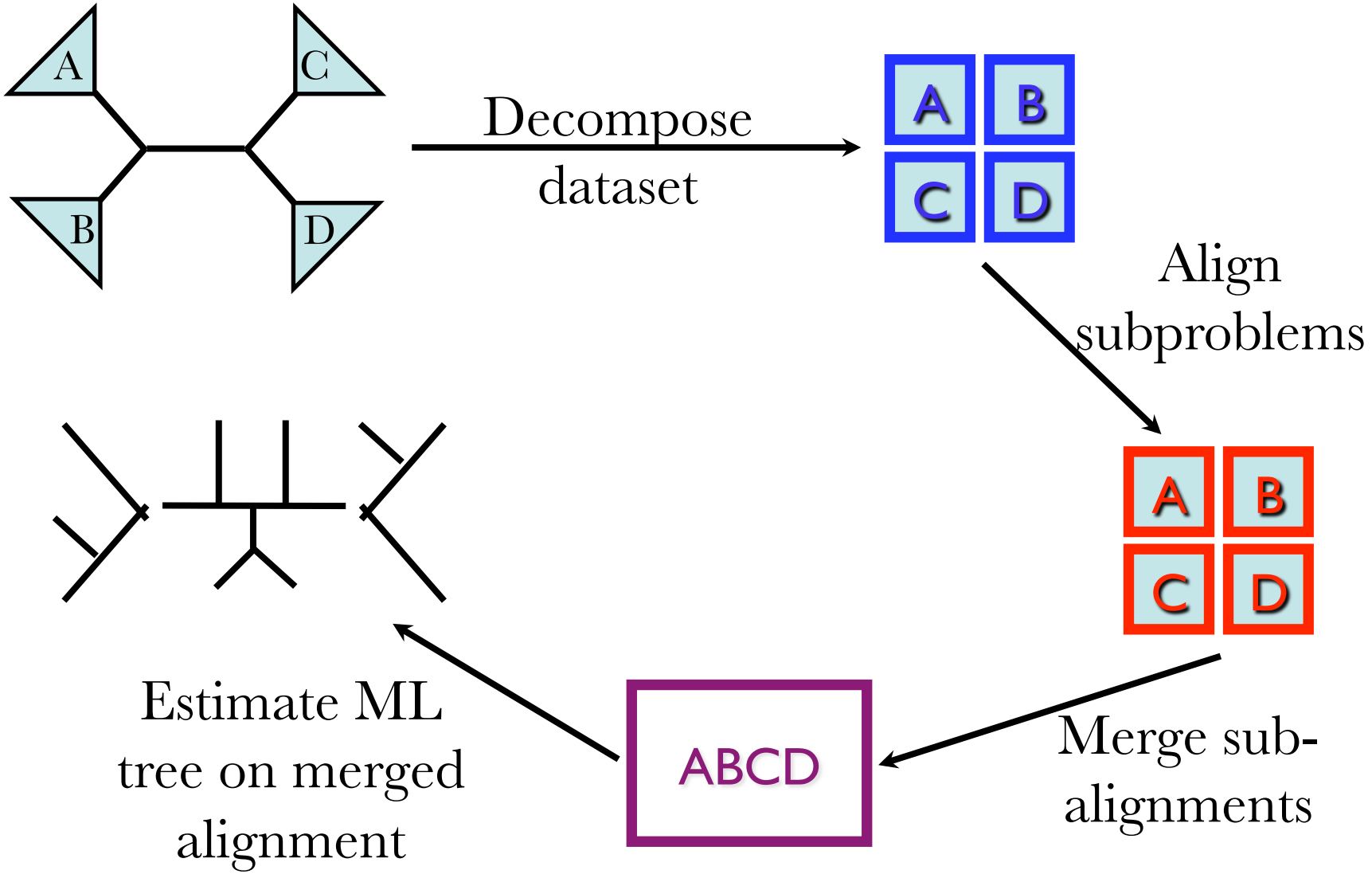
In use by many biologists world-wide.

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

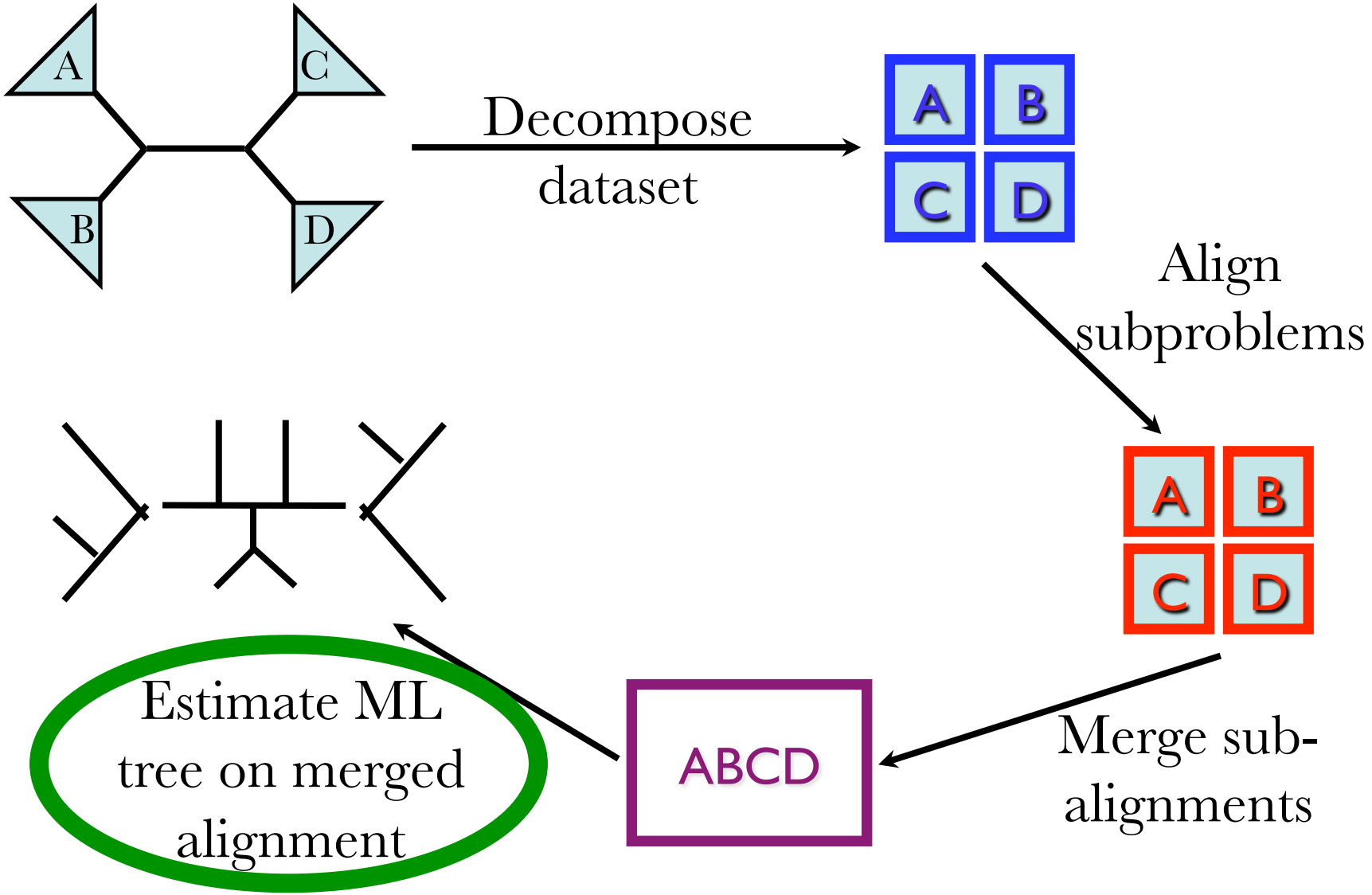
Why does SATé work well?

- Not because it finds the best alignment to optimize ML, treating gaps as missing data!
- Theorem: under Jukes-Cantor, if we allow the alignment to change arbitrarily, then the best alignment is “mono-typic”, and all trees are optimal.

Each SATé iteration



Limitations of SATé-I and -II



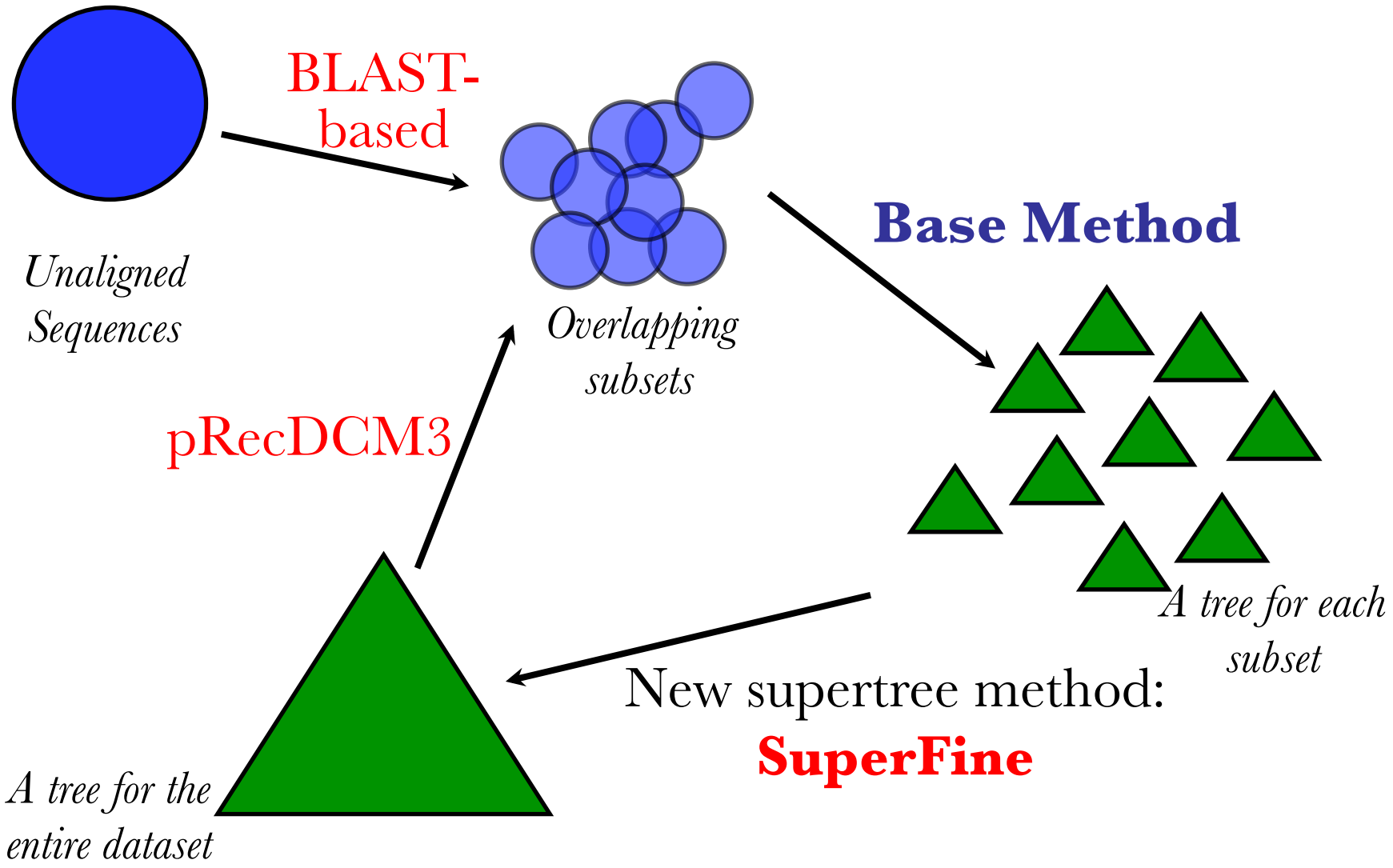
DACTAL-boosting

(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, in preparation

DACTAL-boosting



Superfine: Swenson et al., Systematic Biology, in press.

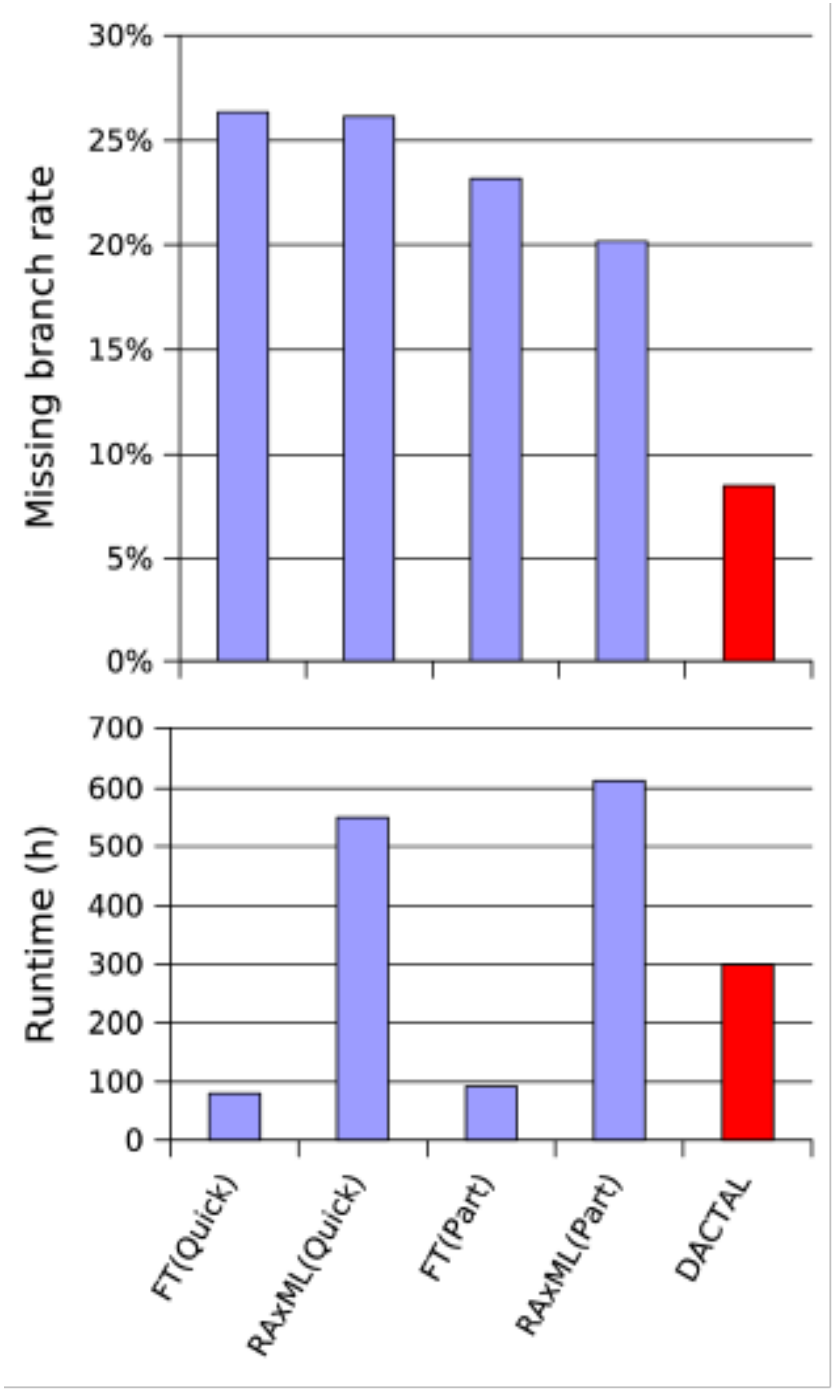
DACTAL-boosting
Base method: ML(MAFFT)

Benchmark: 3 very large biological datasets (with 6K to 28K sequences) from CRW website of curated rRNA sequence alignments

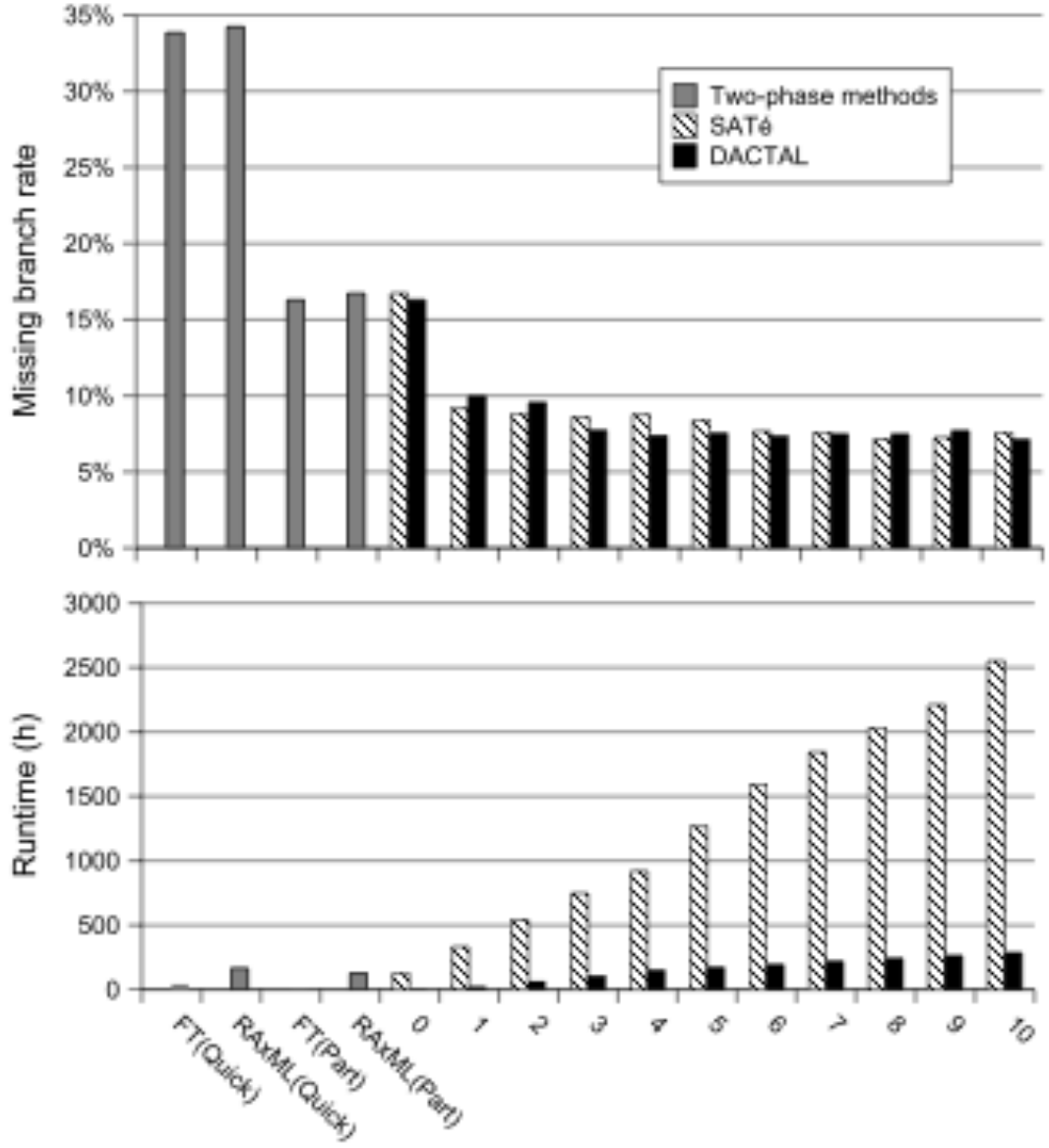
DACTAL runs for 5 iterations starting from FT(Part), and computes RAxML trees on MAFFT alignments of subsets of **250** sequences

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

FastTree (FT) and RAxML are ML methods



DACTAL vs SATé on the 16S.T dataset (~7000 sequences)



Summary for DACTAL and SATé

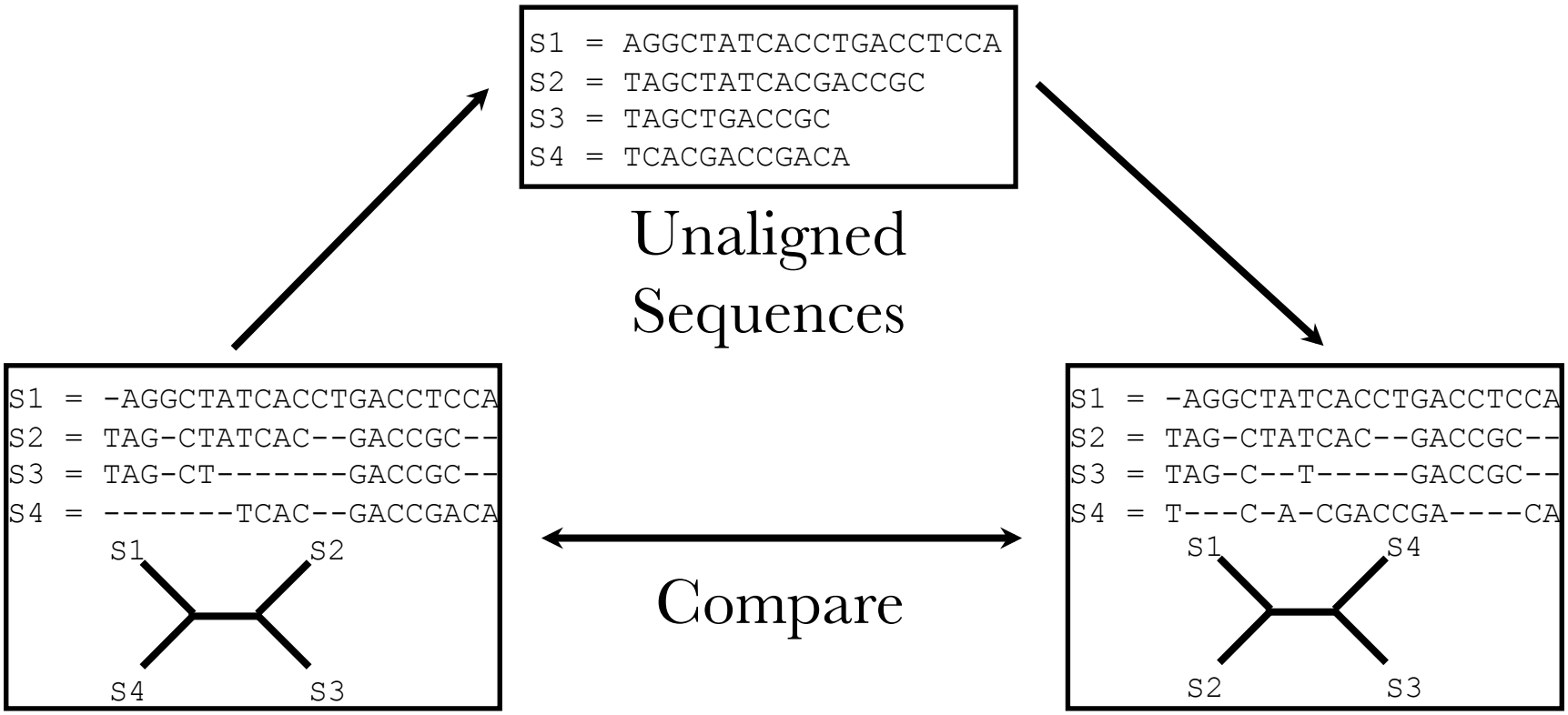
- Both meta-methods are highly robust to starting trees. DACTAL matches the accuracy of SATé on a per-iteration basis, but DACTAL iterations are faster and so it can analyze very large datasets very quickly.
- Following DACTAL with a SATé re-alignment yields very accurate alignments (and faster than just running SATé).
- Future work:
  - DACTAL- and SATé-boosting for statistically-based methods like Bali-Phy
  - DACTAL-boosting for estimating species trees from gene trees
  - Developing mathematical theory explaining why these meta-methods improve estimations

Acknowledgments

- Funding: Packard Foundation, NSF, Guggenheim Foundation, and Microsoft Research New England

- Collaborators: Daniel Huson, Randy Linder, Kevin Liu, Bernard Moret, Luay Nakhleh, Serita Nelesen, Sindhu Raghavan, Usman Roshan, Jerry Sun, Katherine St. John, Mike Steel, and Shel Swenson

Simulation Studies



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

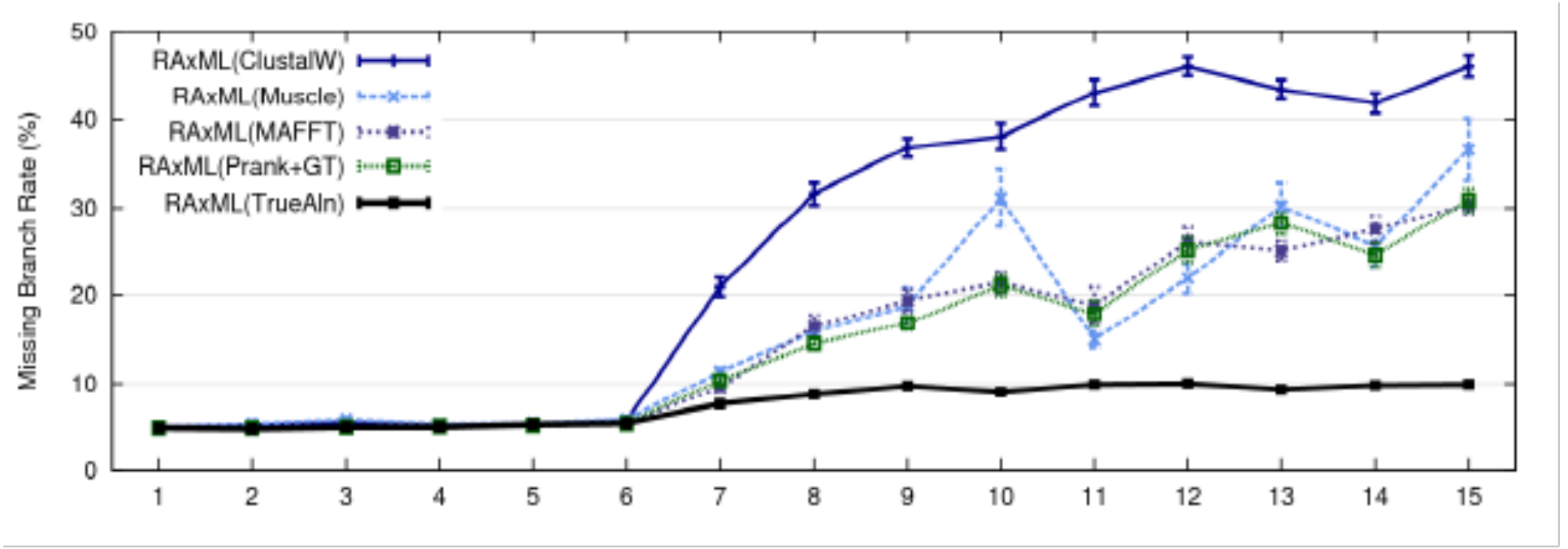
- Bayesian MCMC
- Maximum parsimony
- Maximum likelihood
- Neighbor joining
- FastME
- UPGMA
- Quartet puzzling
- Etc.

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

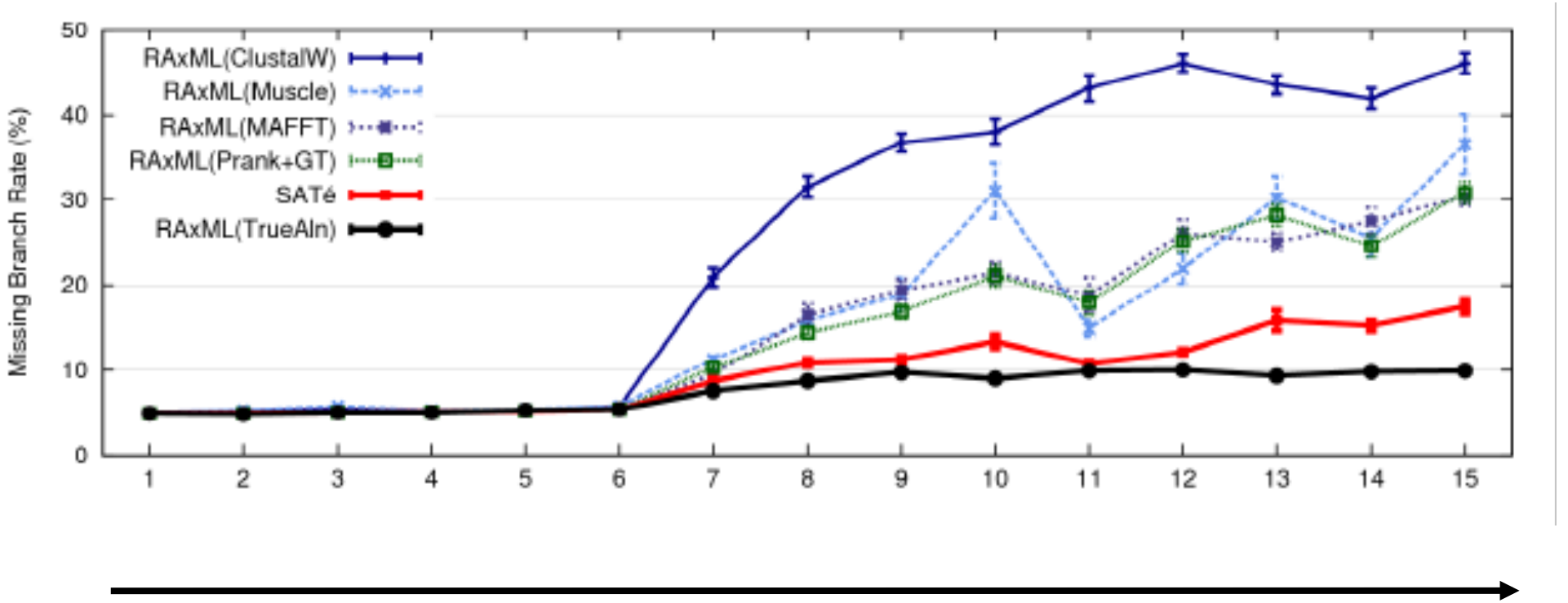
Problems with the two-phase approach

- 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.
- *Potentially useful markers are often discarded* if they are difficult to align.

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



1000 taxon models, ordered by difficulty



1000 taxon models, ordered by difficulty

24 hour SATé analysis, on desktop machines

(Similar improvements for biological datasets)

Research Projects

- Estimating multiple sequence alignments and phylogenies on large datasets
- Estimating species trees from gene trees
- Supertree methods
- Whole genome phylogeny using gene order and content
- Phylogenetic estimation under statistical models
- Datamining sets of trees and alignments
- Visualization of ultra-large trees
- Reticulate phylogeny detection and estimation