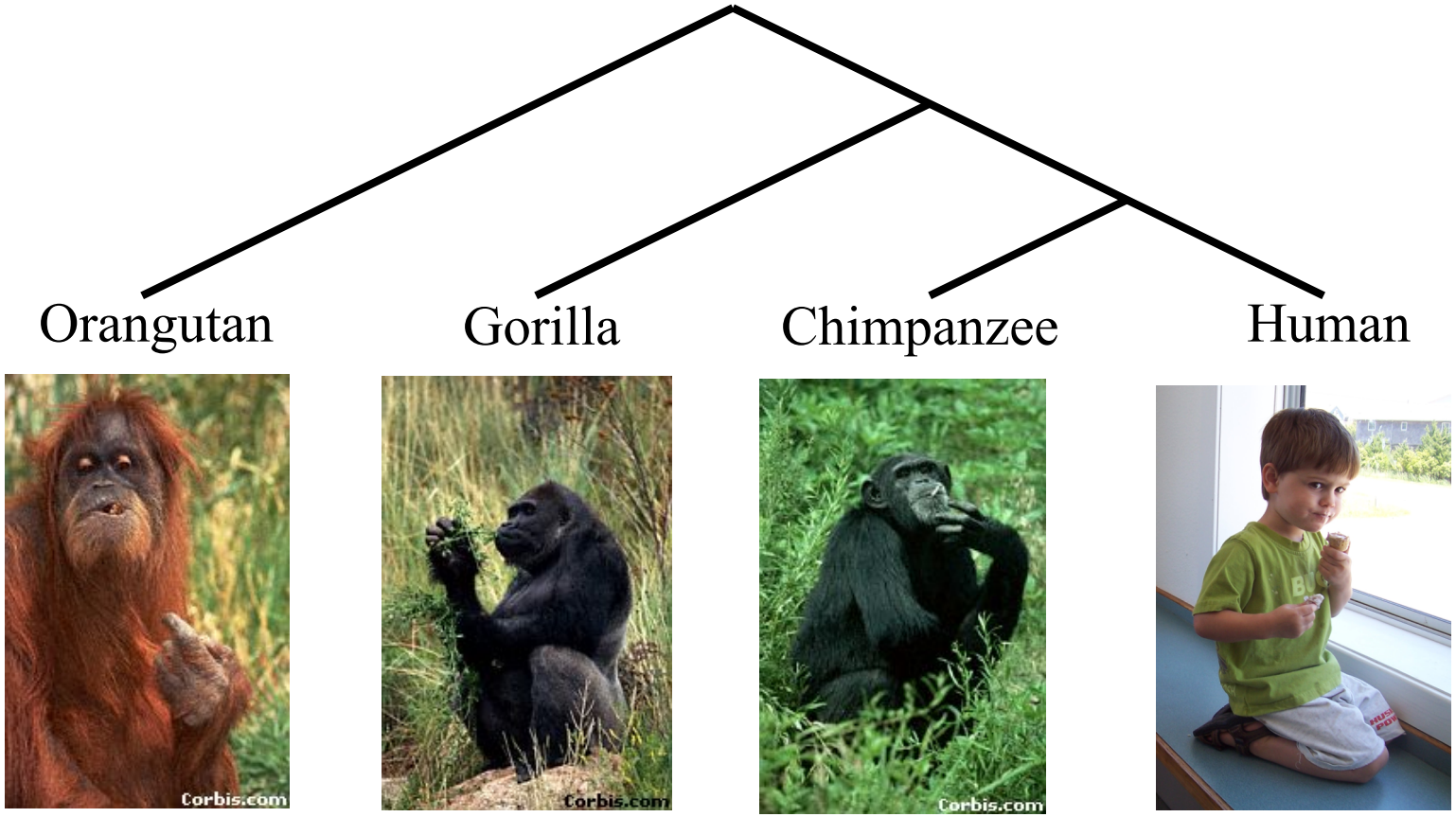
Ultra-Large Phylogeny Estimation Using SATé and DACTAL

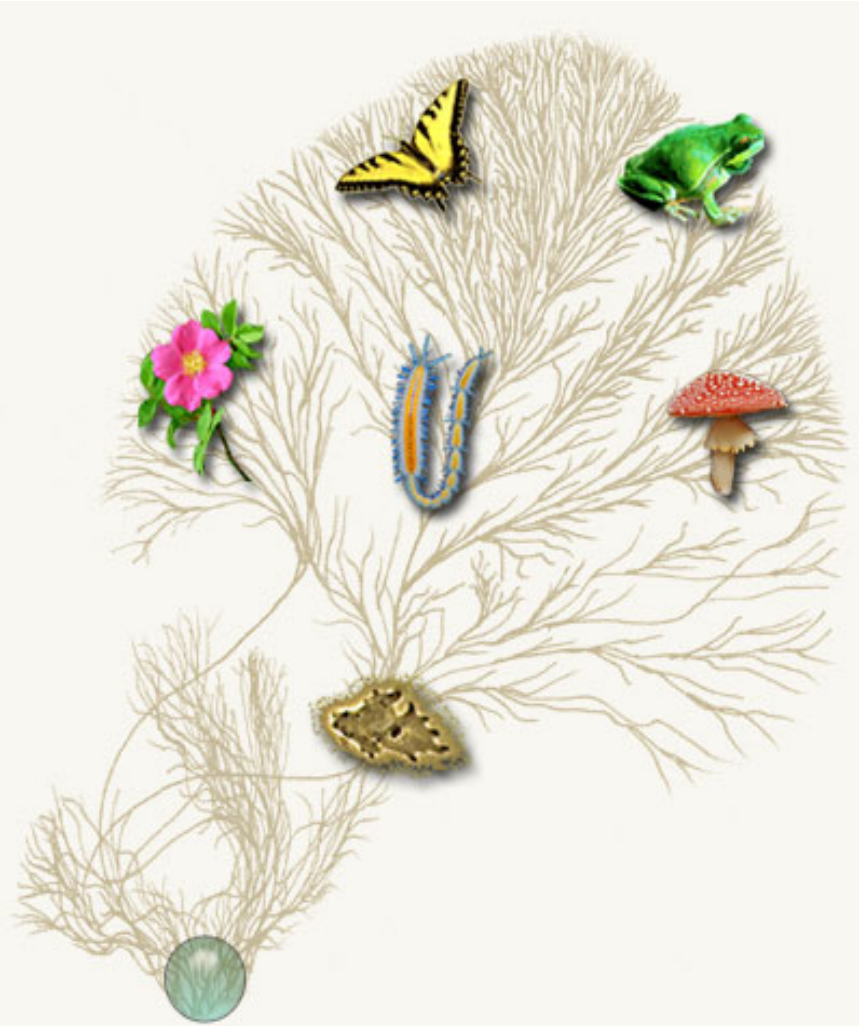
Tandy Warnow
Department of Computer Science
The University of Texas at Austin

Phylogeny (evolutionary tree)



From the Tree of the Life Website,
University of Arizona

How did life evolve on earth?

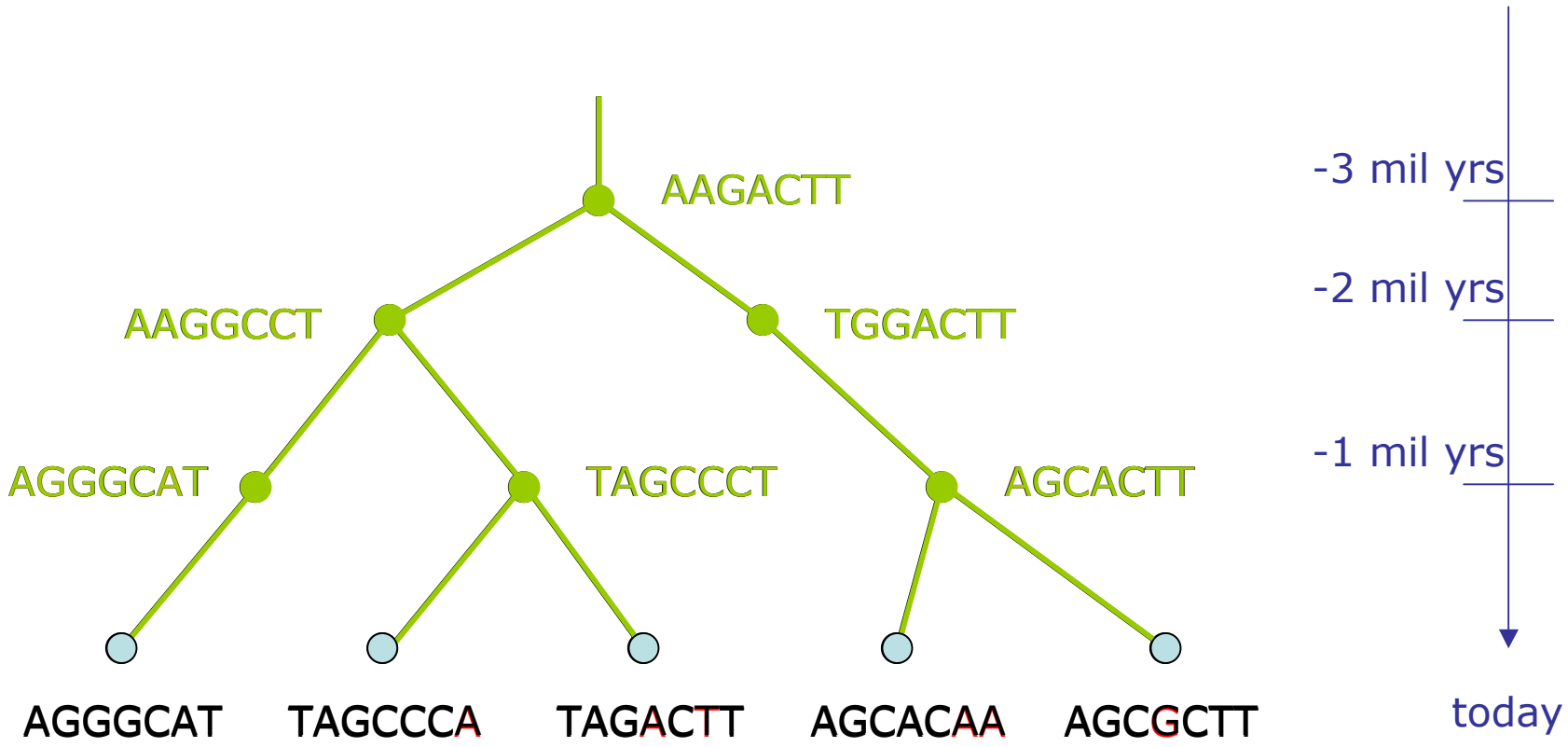


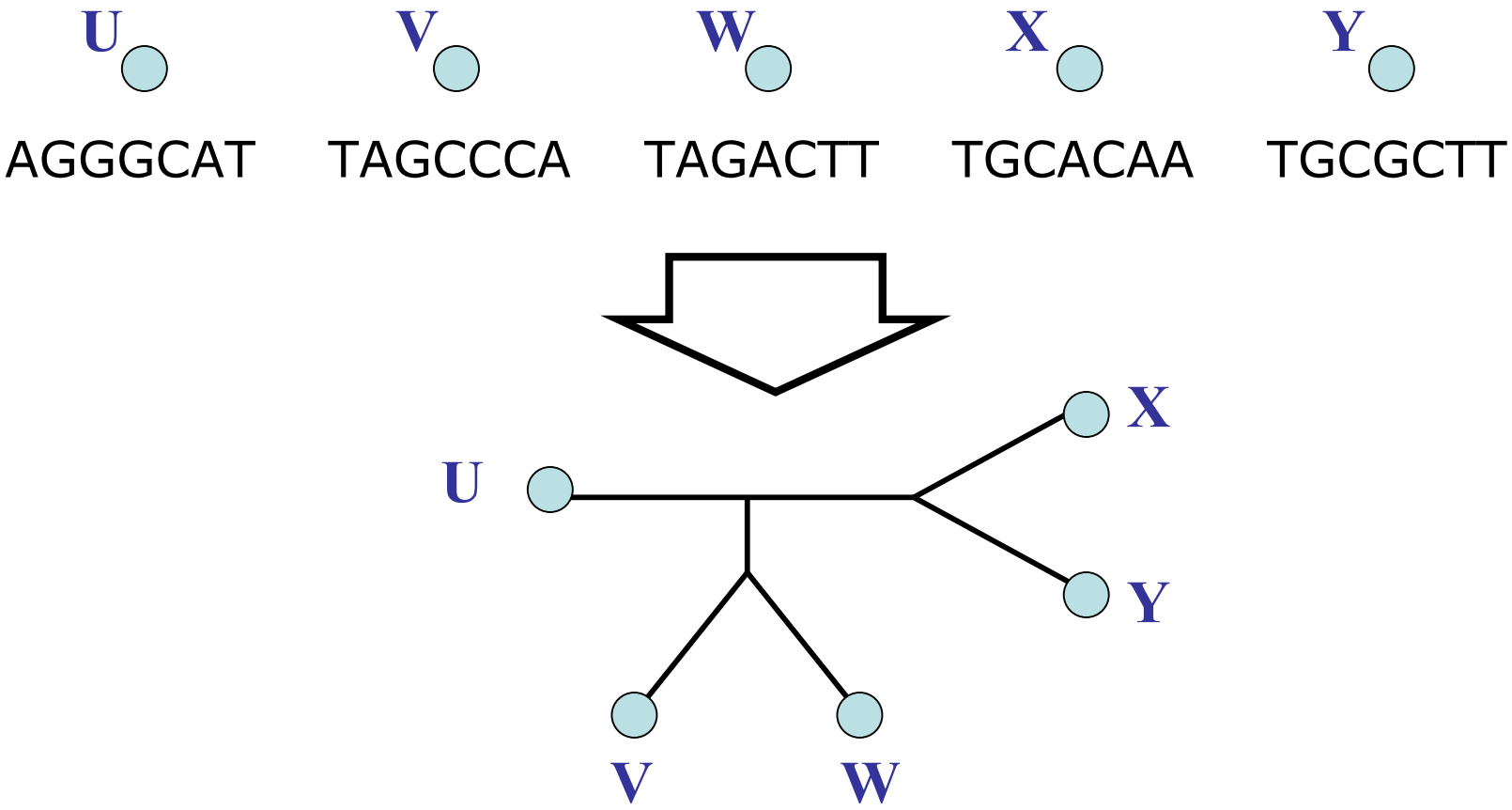
- **Courtesy of the Tree of Life project**

An international effort to understand how life evolved on earth

Biomedical applications: drug design, protein structure and function prediction, biodiversity.

DNA Sequence Evolution




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

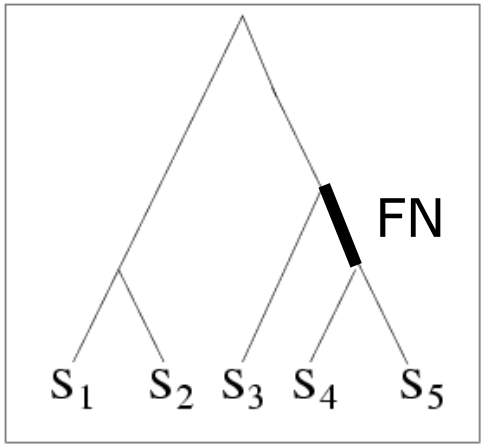
Markov Model of Site Evolution

Simplest (Jukes-Cantor):

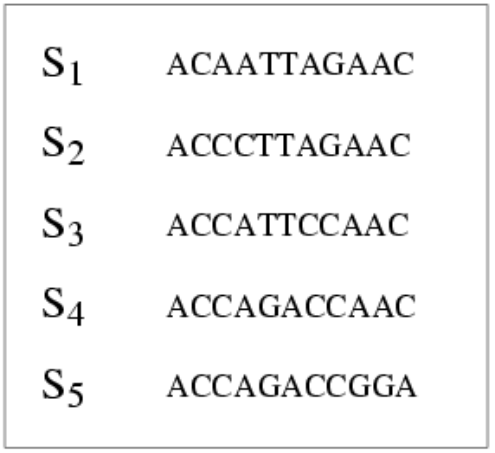
- The model tree T is binary and has substitution probabilities p(e) on each edge e.
- The state at the root is randomly drawn from {A,C,T,G} (nucleotides)
- If a site (position) changes on an edge, it changes with equal probability to each of the remaining states.
- The evolutionary process is Markovian.

More complex models (such as the General Markov model) are also considered, often with little change to the theory.

Quantifying Error



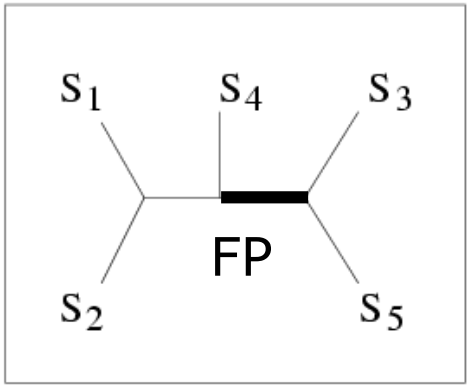
TRUE TREE



DNA SEQUENCES

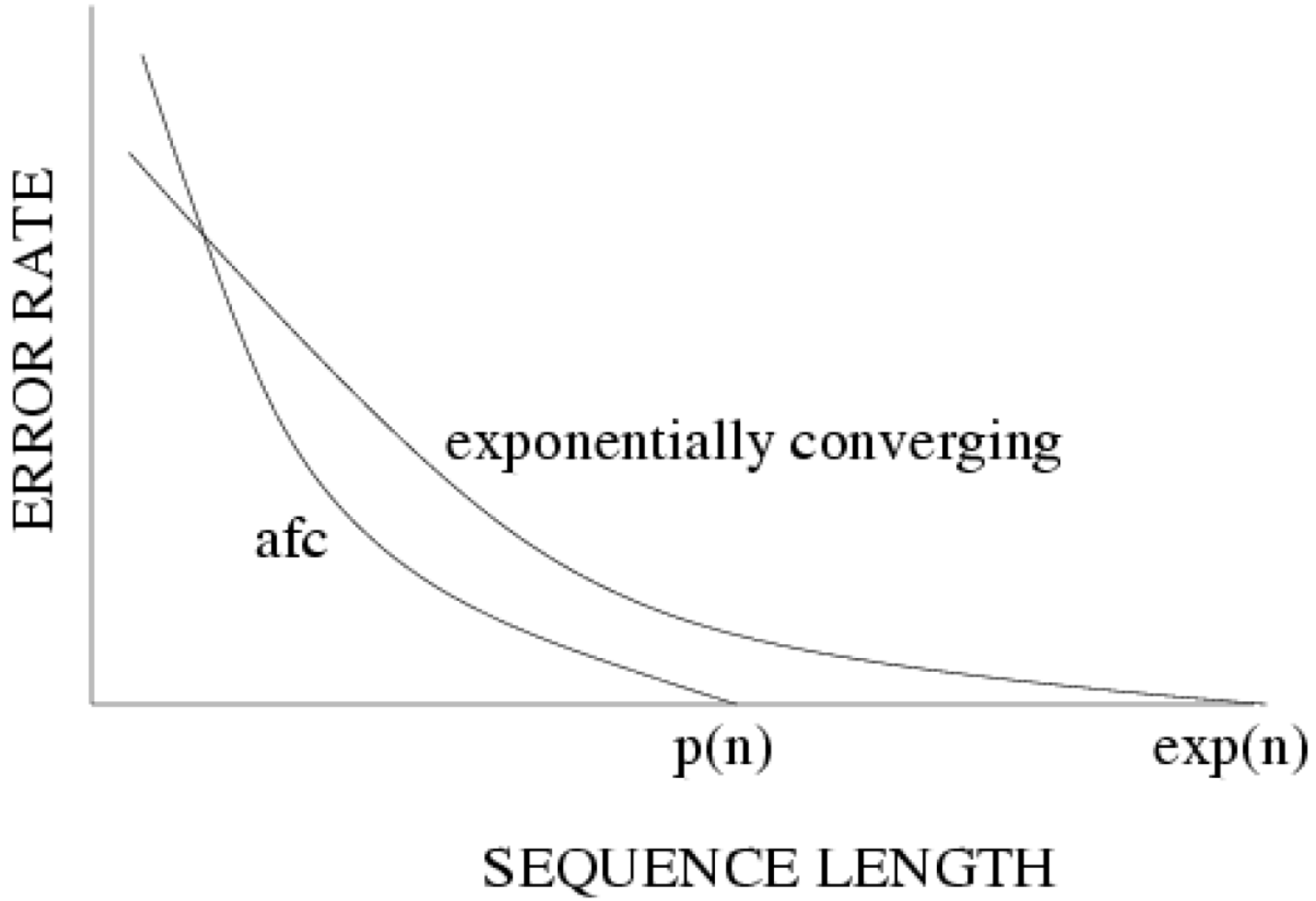
FN: false negative
(missing edge)
FP: false positive
(incorrect edge)

50% error rate

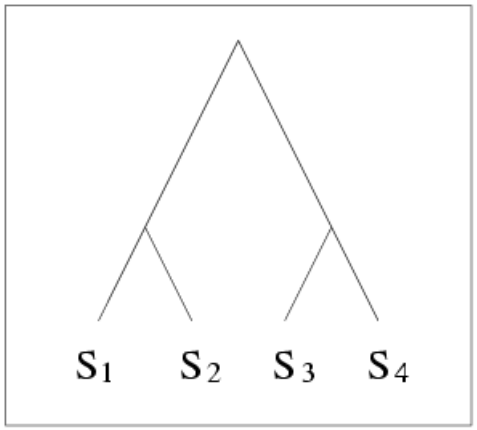


INFERRED TREE

Statistical consistency, exponential convergence, and absolute fast convergence (afc)



Distance-based estimation



TRUE TREE

S_1	ACAATTAGAAC
S_2	ACCCTTAGAAC
S_3	ACCATTCCAAC
S_4	ACCAGACCAAC

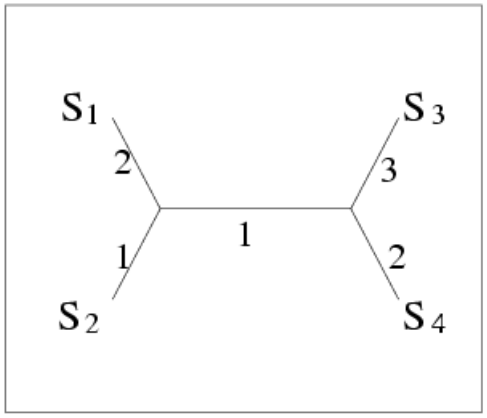
DNA SEQUENCES

STATISTICAL ESTIMATION OF PAIRWISE DISTANCES

	S_1	S_2	S_3	S_4
S_1	0	3	6	5
S_2		0	5	4
S_3			0	5
S_4				0

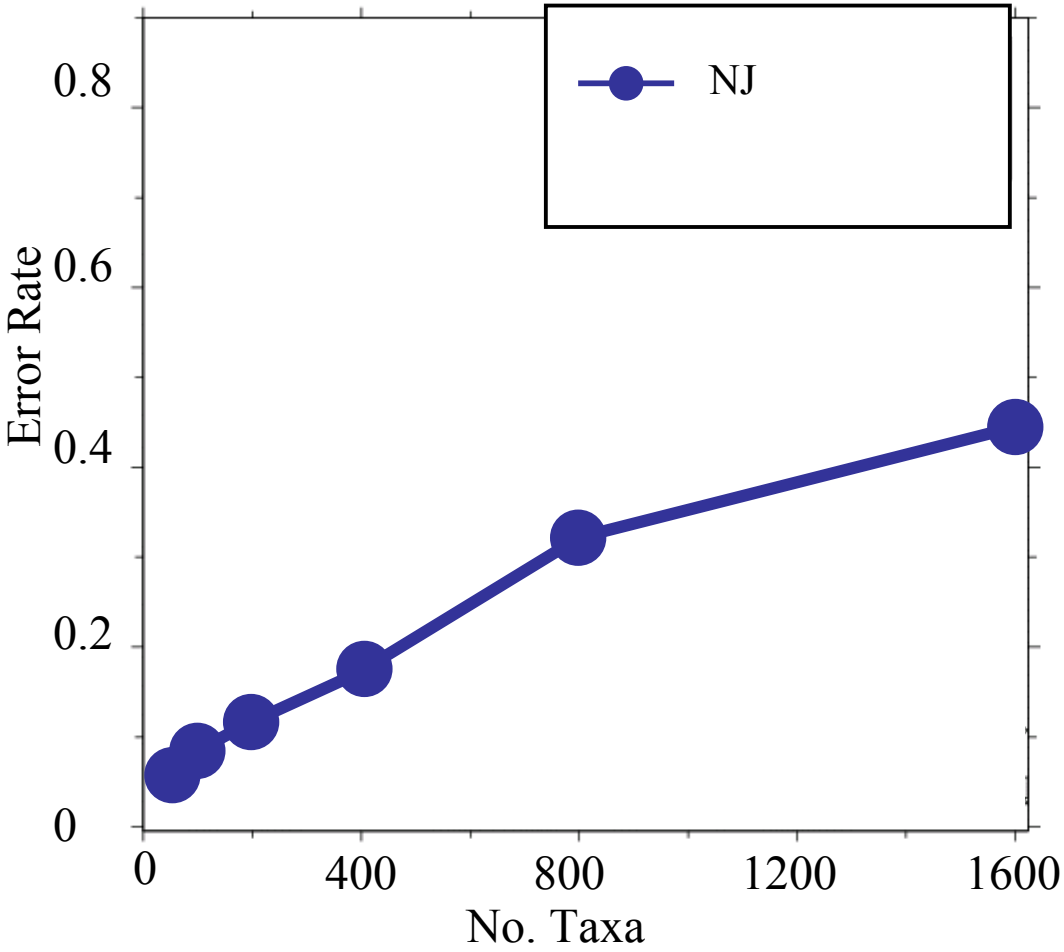
DISTANCE MATRIX

METHODS SUCH AS NEIGHBOR JOINING



INFERRED TREE

Performance on large diameter trees



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.

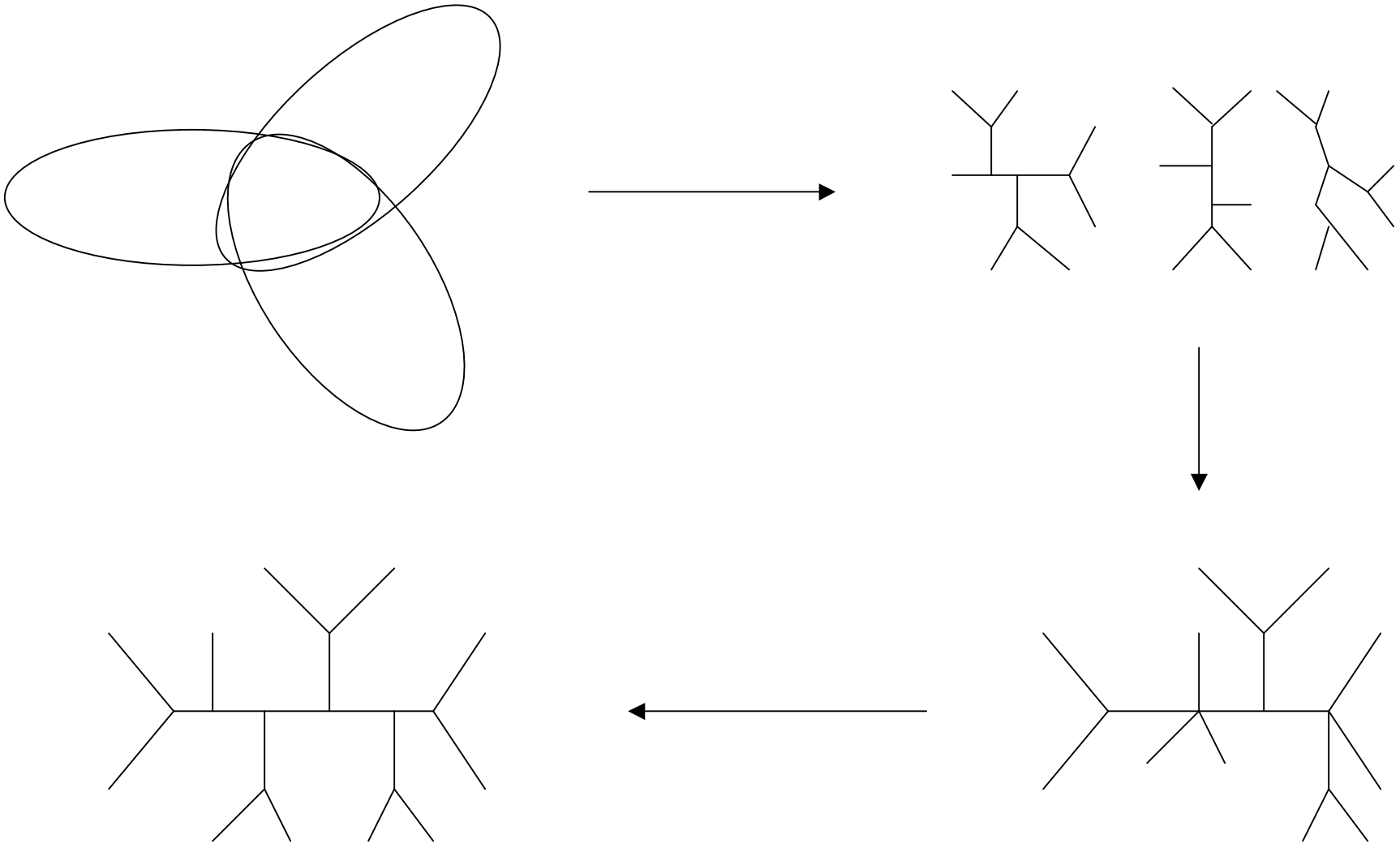
[Nakhleh et al. ISMB 2001]

Theorem (Erdos et al., Atteson): Neighbor joining (and some other methods) will return the true tree w.h.p. provided sequence lengths are exponential in the evolutionary diameter of the tree.

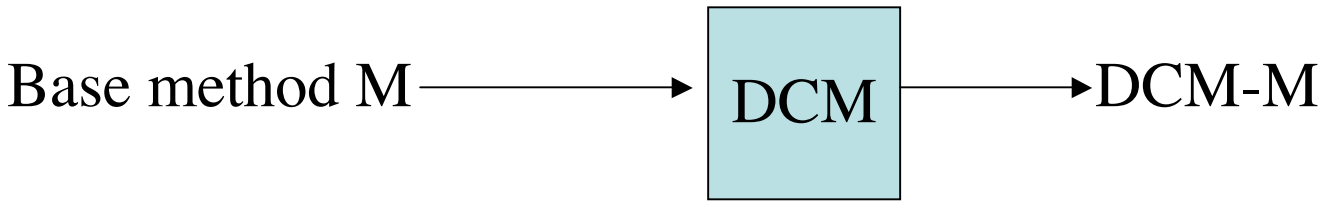
Sketch of proof:

- NJ (and other distance methods) guaranteed correct if *all* entries in the estimated distance matrix have low error
- Estimations of large distances require long sequences to have low error w.h.p.

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



- **DCMs "boost" the performance of phylogeny reconstruction methods.**



Phylogenetic "boosters" (meta-methods)

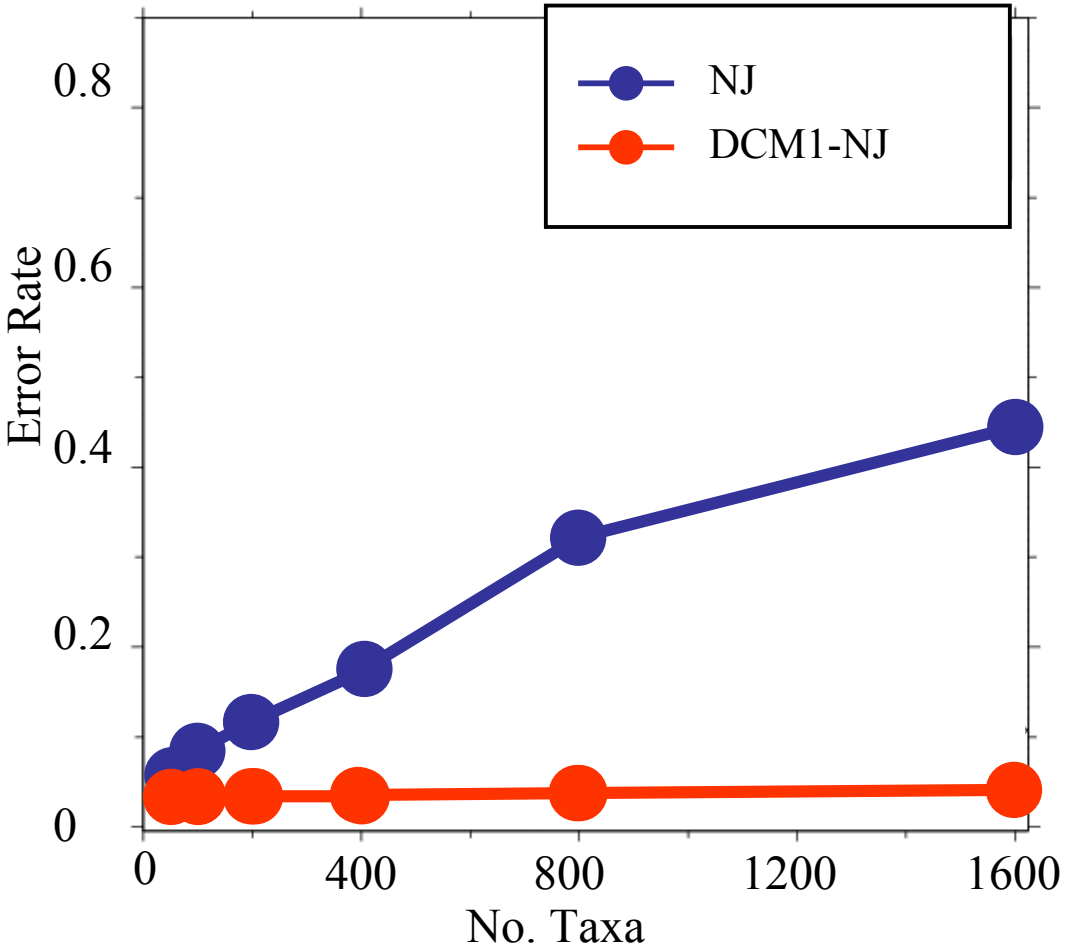
Goal: improve accuracy, speed, robustness, or theoretical guarantees of base methods

Examples:

- DCM-boosting for distance-based methods (1999)
- DCM-boosting for heuristics for NP-hard problems (1999)
- SATé-boosting for alignment methods (2009)
- SuperFine-boosting for supertree methods (2011)
- DACTAL-boosting for all phylogeny estimation methods (2011)
- SEPP-boosting for metagenomic analyses (2011)

DCM1-boosting distance-based methods

[Nakhleh et al. ISMB 2001]



Theorem (Warnow et al., SODA 2001): DCM1-NJ converges to the true tree from polynomial length sequences

Summary

DCM-NJ has better accuracy than NJ, and DCM-boosting of other distance-based method also produces very big improvements in accuracy

First afc methods developed by Erdos et al; later ones with even better theoretical performance (see papers by Daskalakis, Mossel, Roch, Gronau, Moran, Snir, and others).

Roch and collaborators have established a threshold for branch lengths, below which *logarithmic* sequence lengths can suffice for accuracy

What about more complex models?

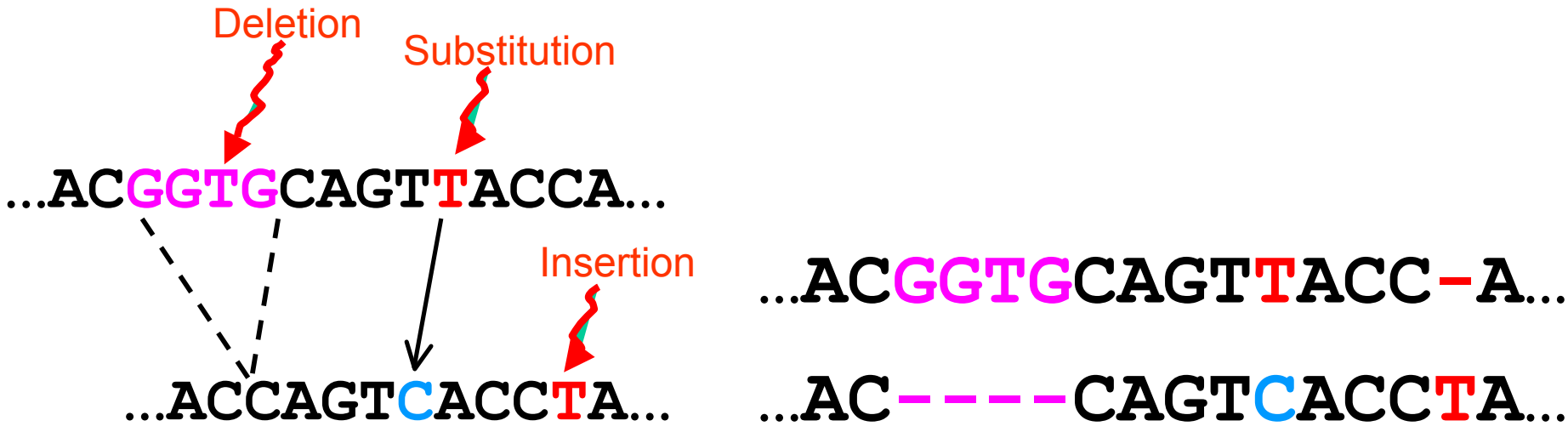
These results only apply when sequences evolve under these nice substitution-only models.

*What can we say about estimating trees when sequences evolve with insertions and deletions ("**indels**")?*

Today's talk:

some theory, some empirical performance

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



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

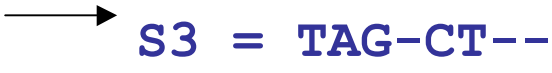
```
S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC
S3 = TAGCTGACCGC
S4 = TCACGACCGACA
```



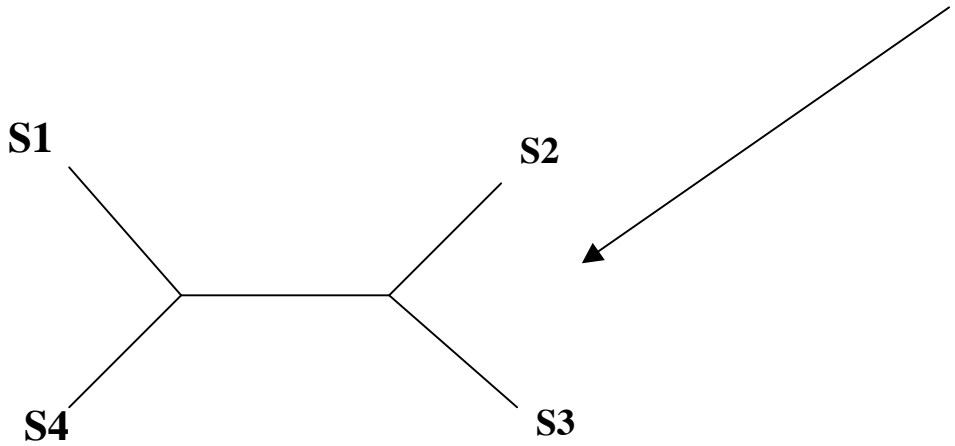
```
S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAC--GACCGC--
S3 = TAG-CT-------GACCGC--
S4 = -------TCAC--GACCGACA
```

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

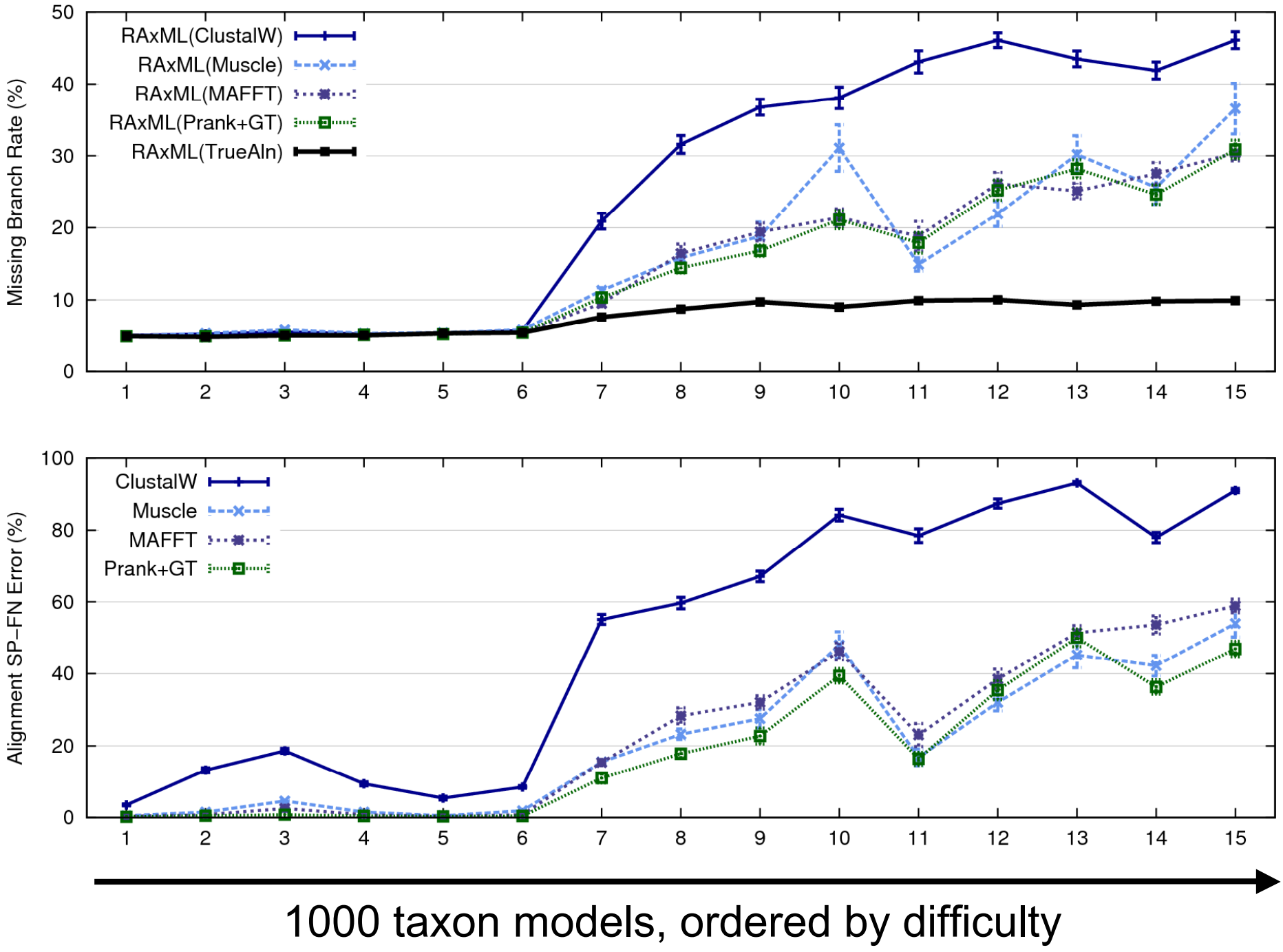
Alignment methods

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

Phylogeny methods

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

***RAxML**: best 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

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:

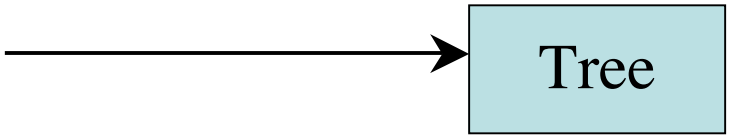
Simultaneous Alignment and Tree Estimation

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

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

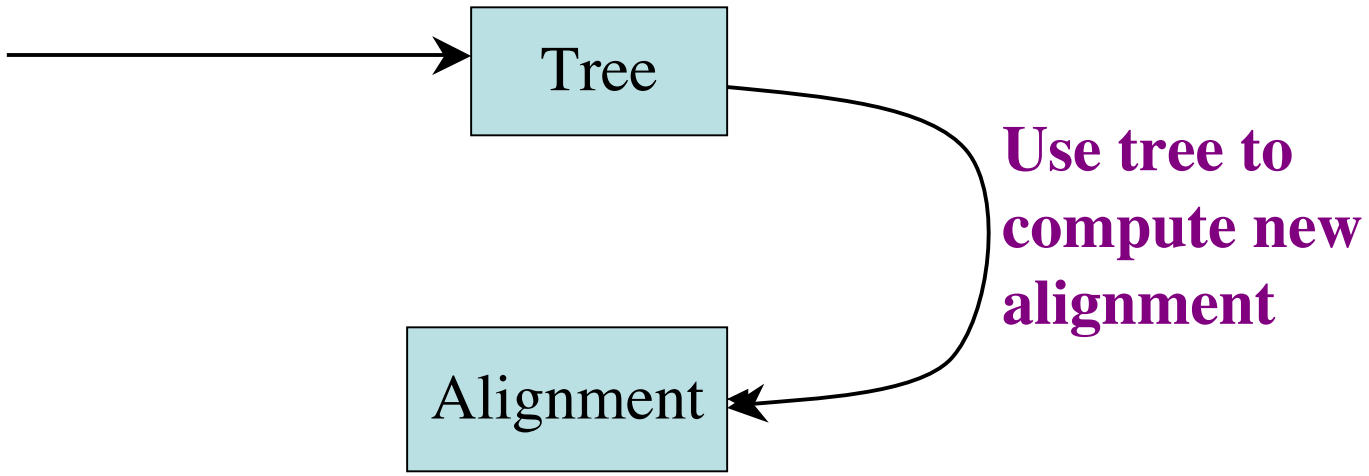
SATé Algorithm

Obtain initial alignment
and estimated ML tree

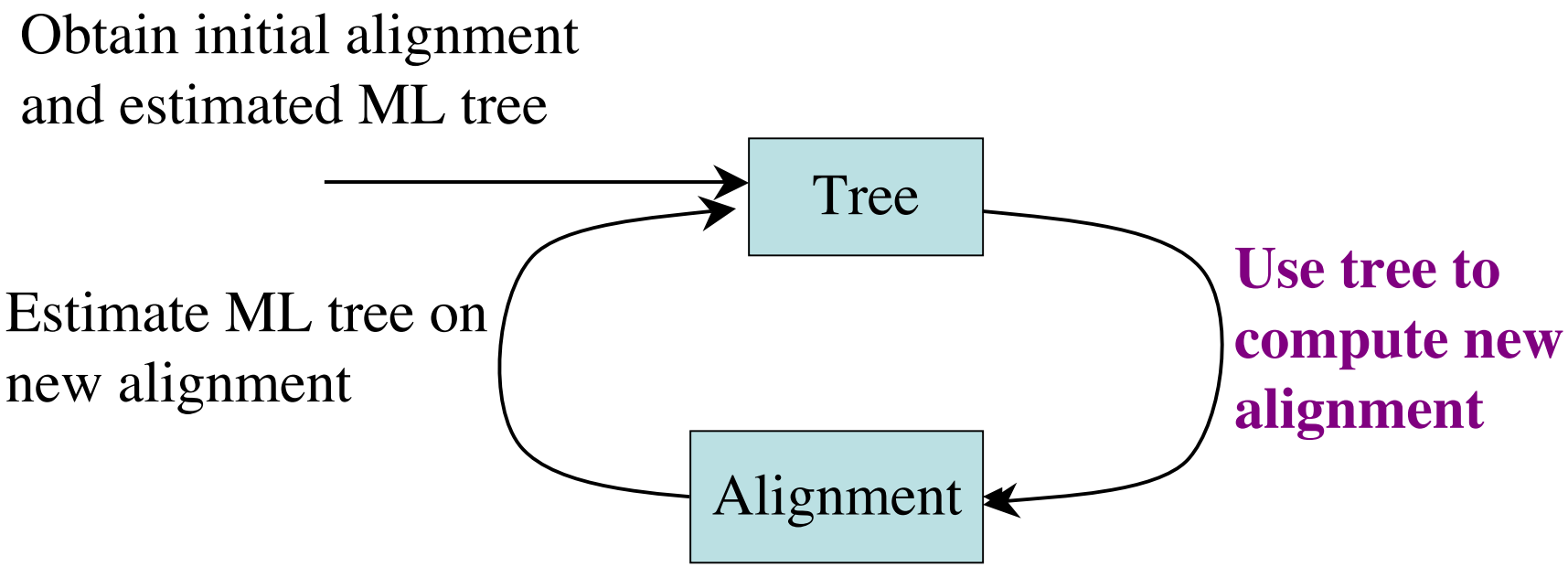


SATé Algorithm

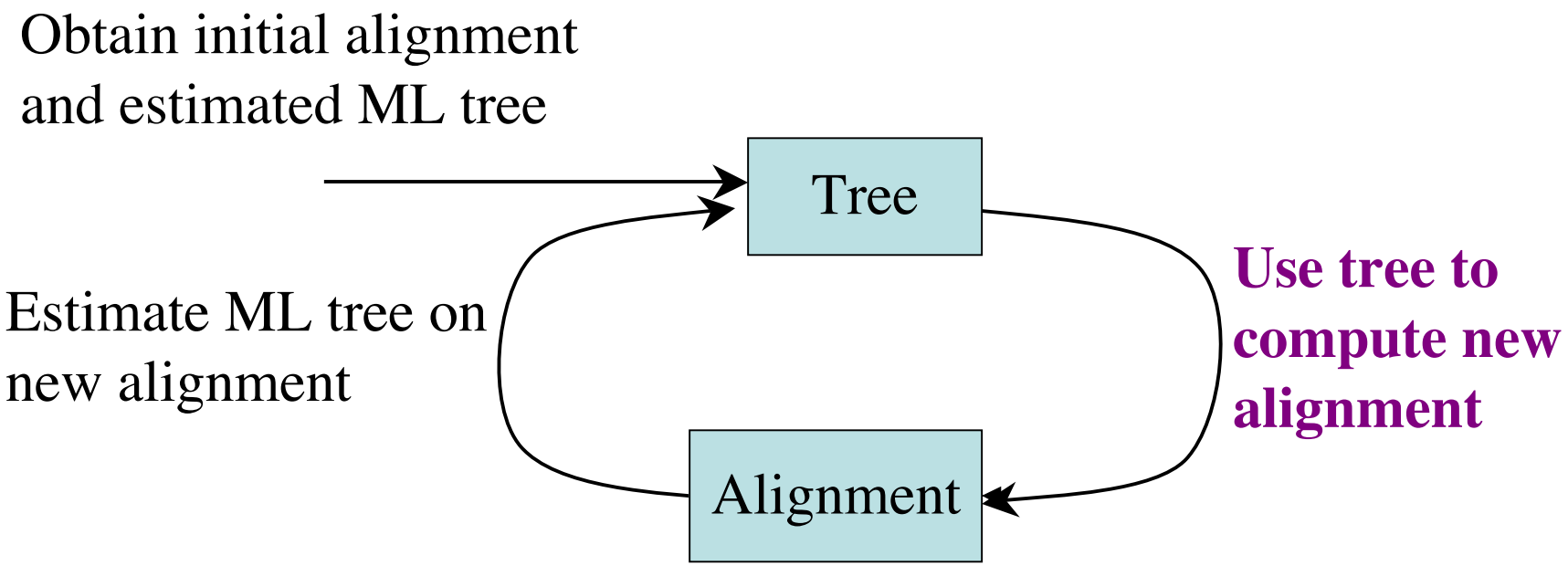
Obtain initial alignment
and estimated ML tree



SATé Algorithm



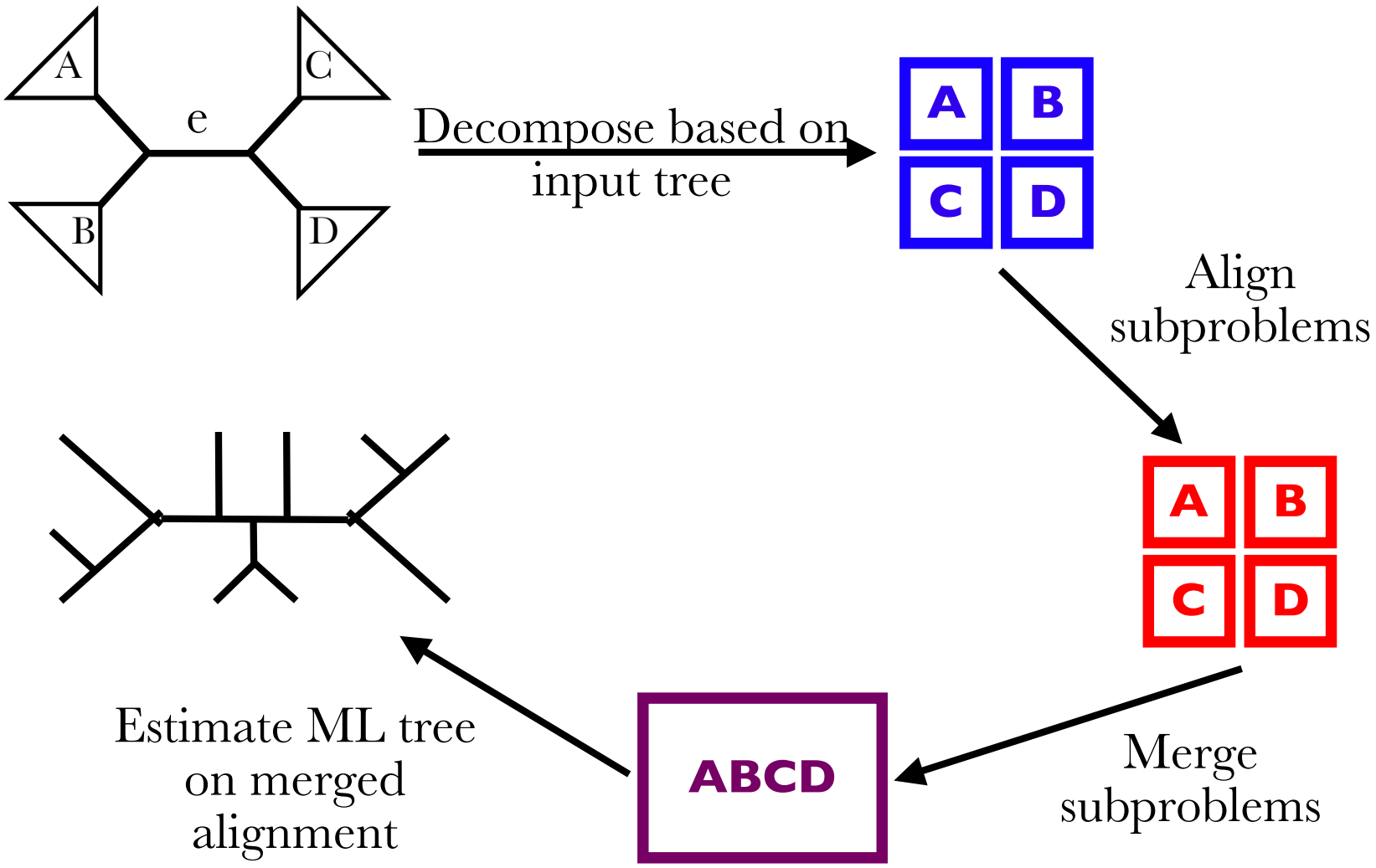
SATé Algorithm

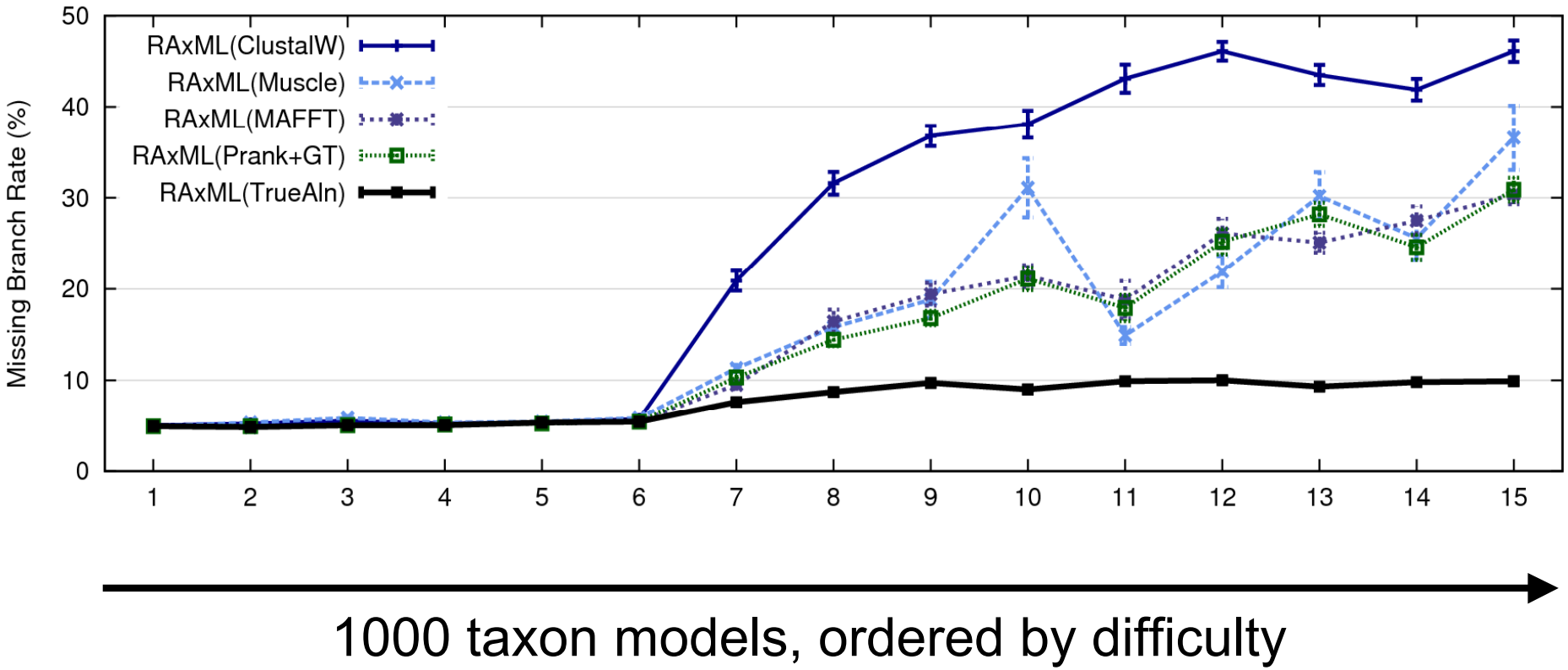


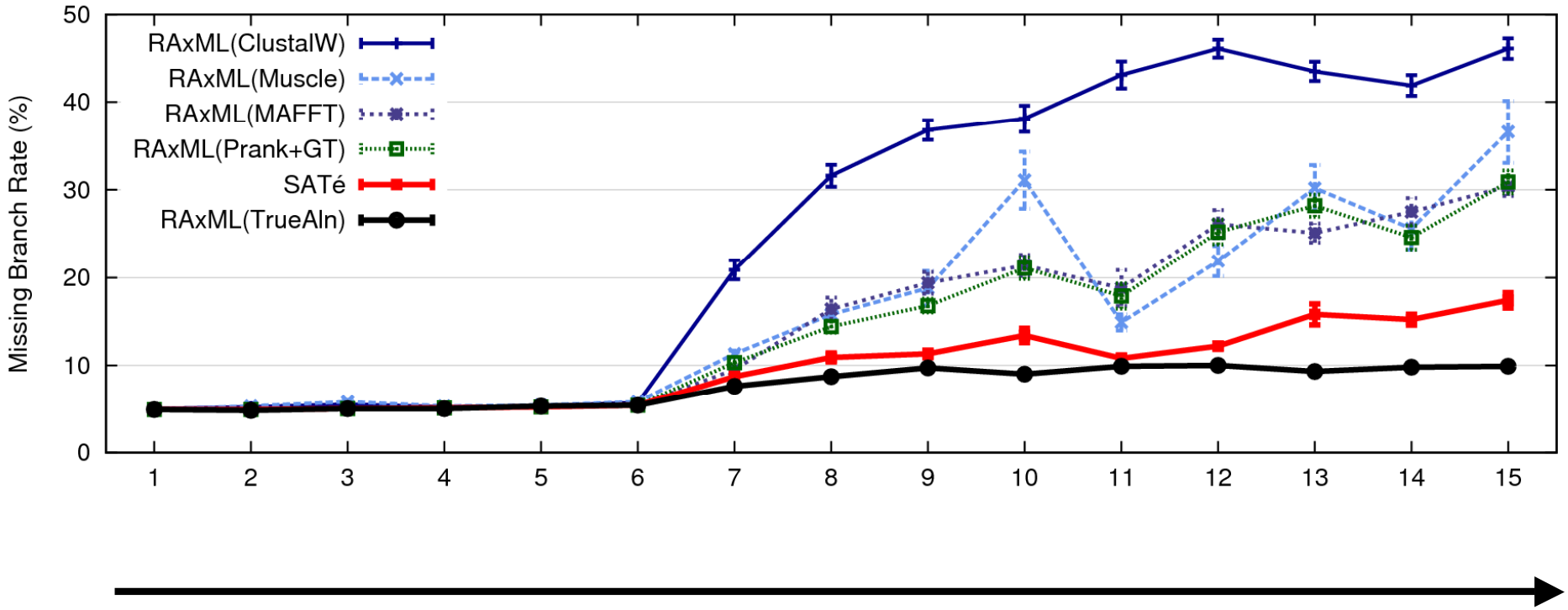
If new alignment/tree pair has worse ML score, realign using a different decomposition

Repeat until termination condition (typically, 24 hours)

One SATé iteration (really 32 subsets)



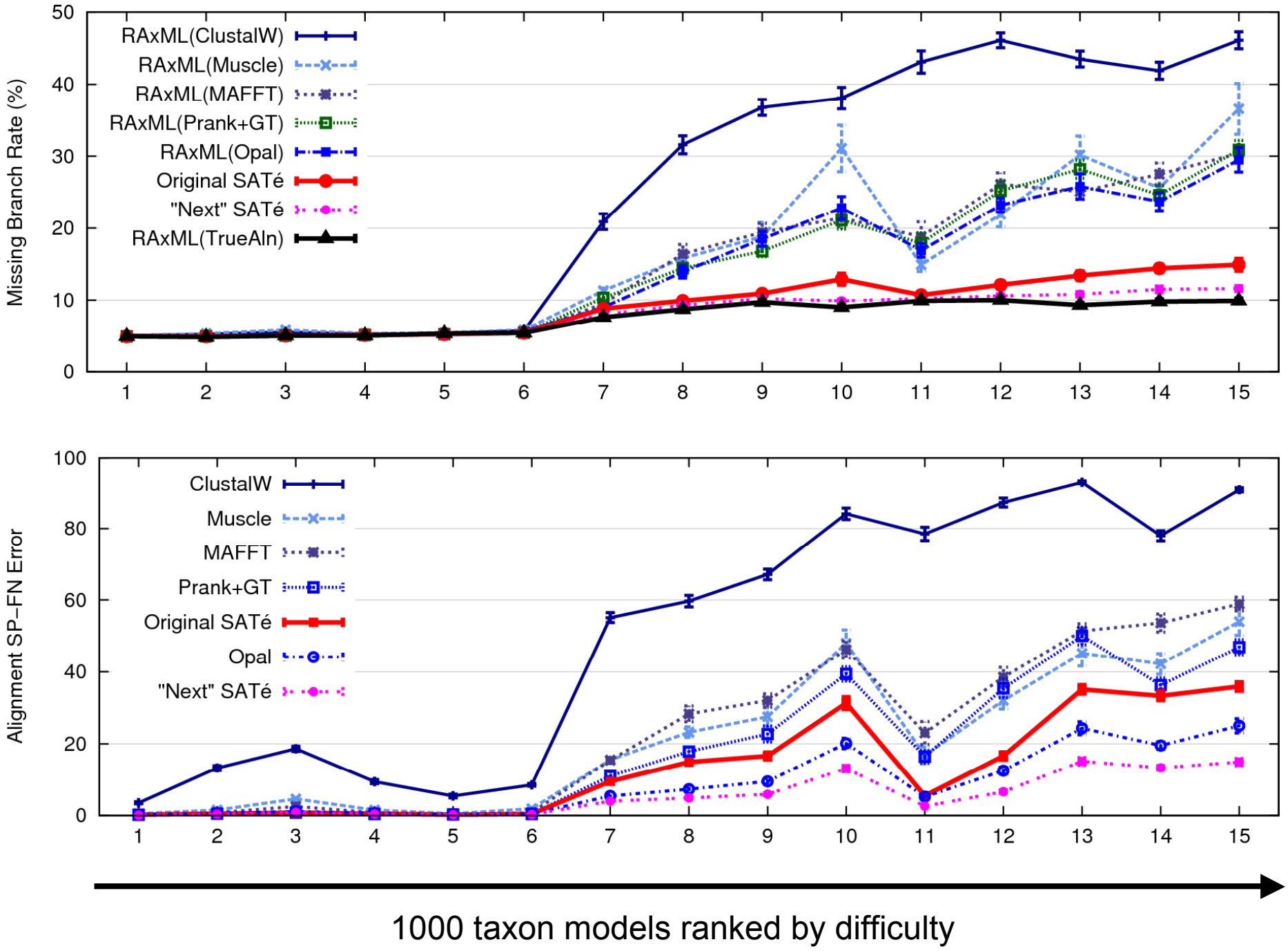

RAxML(ClustalW)
RAxML(Muscle)
RAxML(MAFFT)
RAxML(Prank+GT)
RAxML(TrueAln)
Missing Branch Rate (%)
0
10
20
30
40
50
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
1000 taxon models, ordered by difficulty



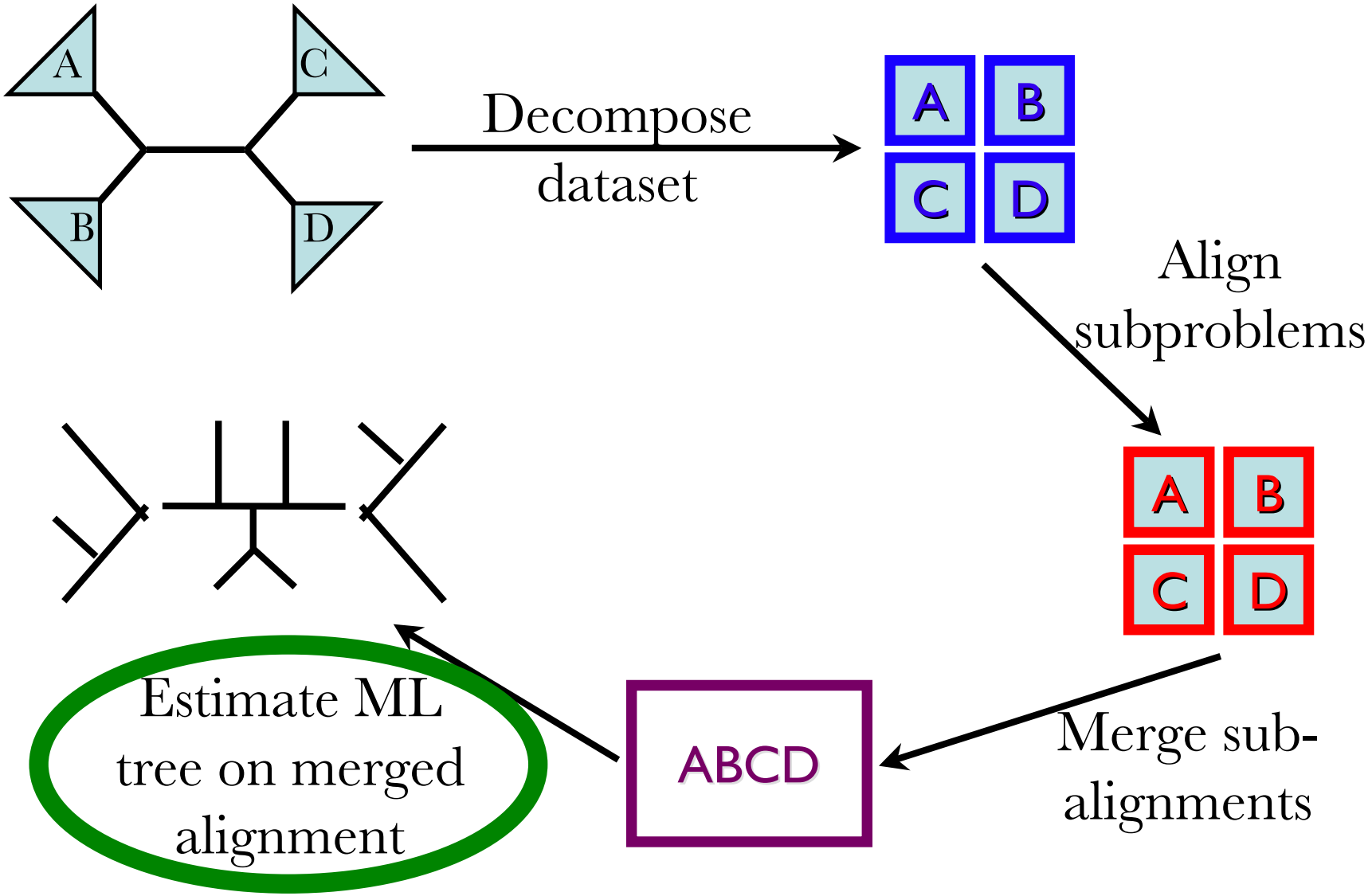
1000 taxon models, ordered by difficulty

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

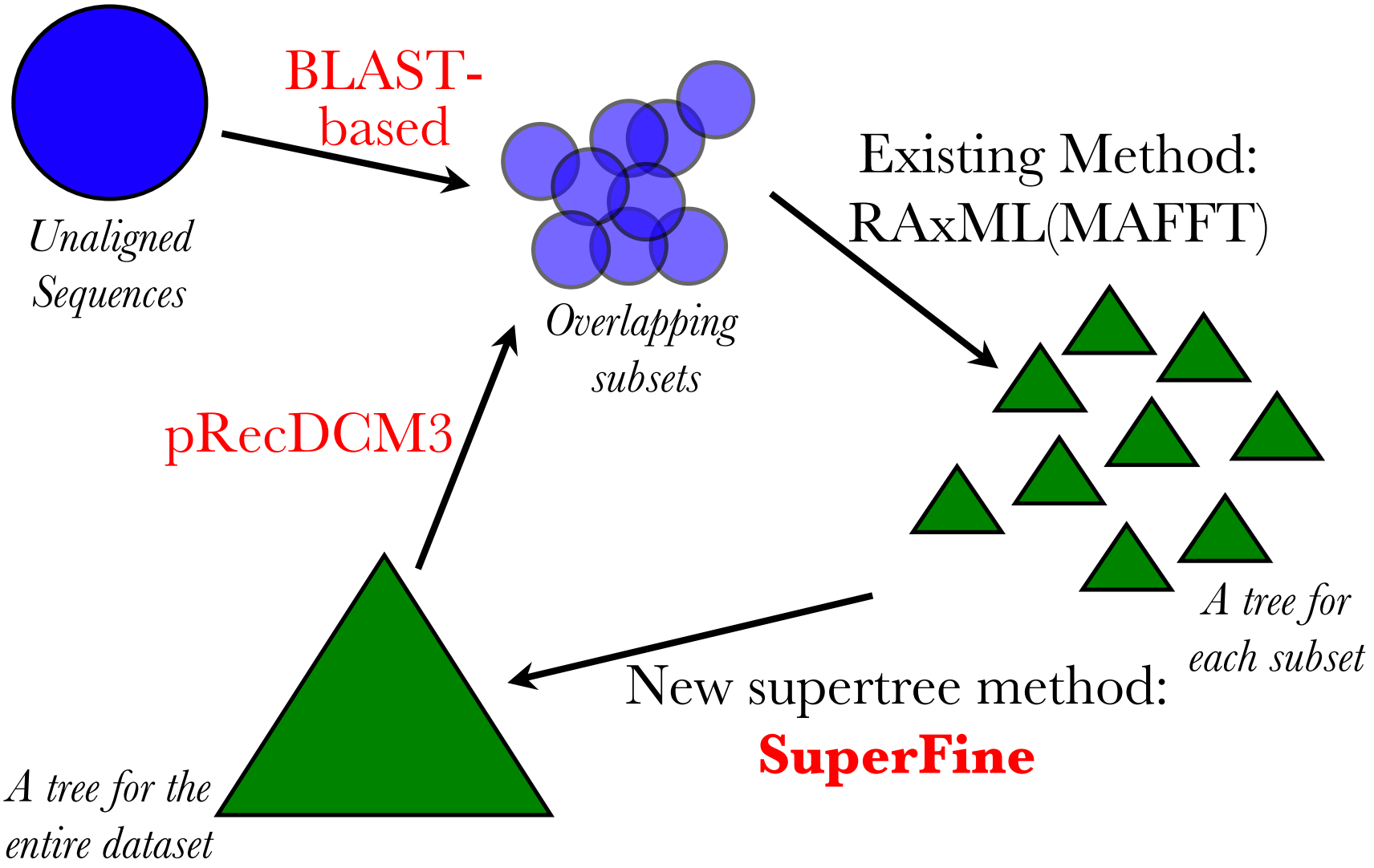
Limitations of SATé-I and -II



DACTAL

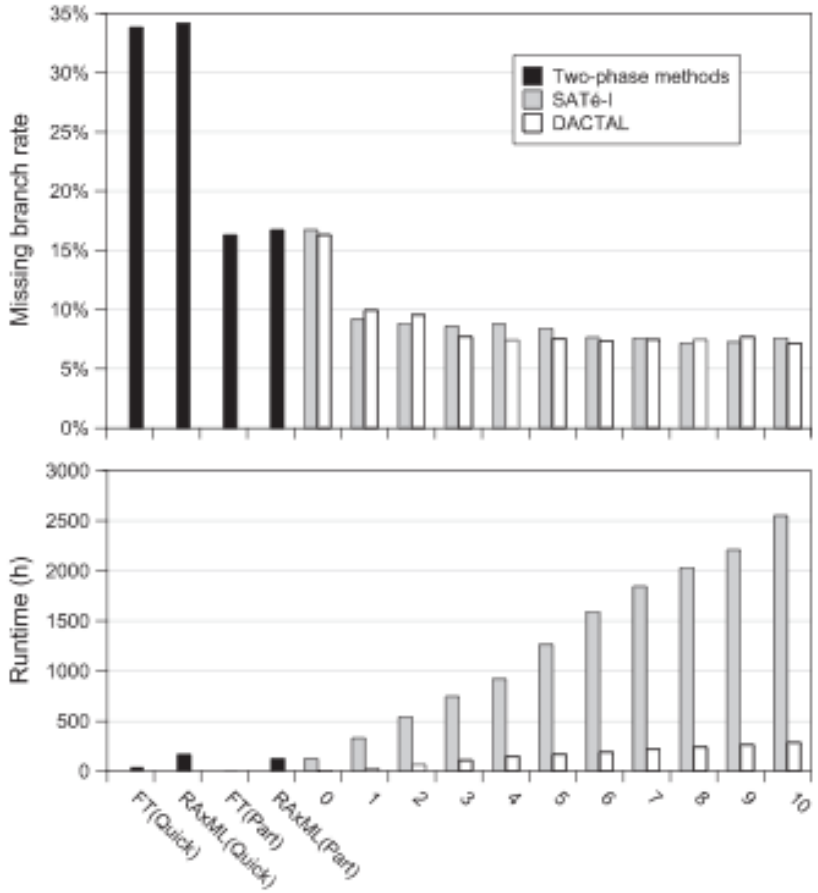
- DACTAL: Divide-and-Conquer Trees (almost) without alignments (Nelesen et al., submitted)

DACTAL


BLAST-based
Unaligned Sequences
Overlapping subsets
Existing Method: RAxML(MAFFT)
A tree for each subset
New supertree method:
SuperFine
pRecDCM3
A tree for the entire dataset

DACTAL vs. SATé

- **16S.T, 7350 rRNA sequences, curated structural alignment, ML bootstrap tree**



DACTAL vs 2-phase methods

CRW: Comparative RNA database,

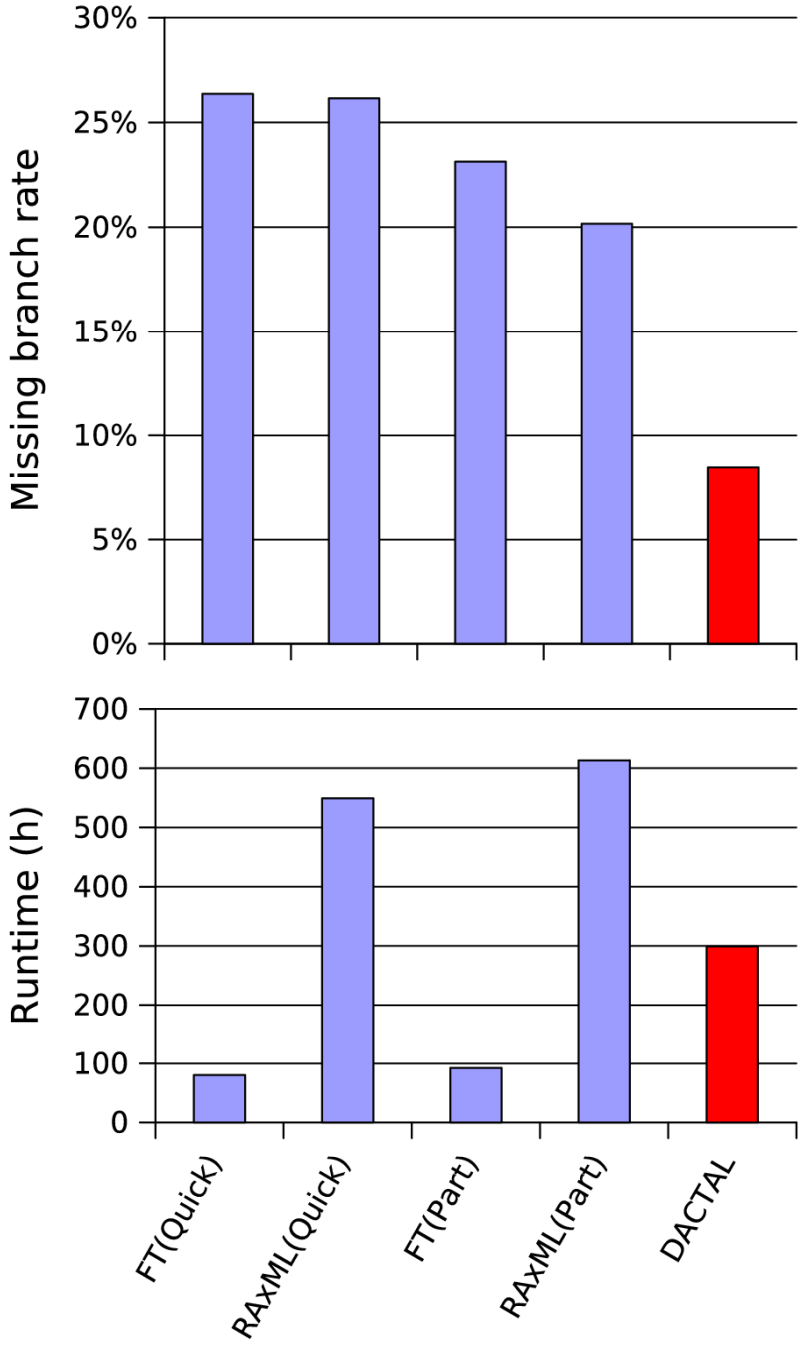
Three 16S datasets with 6,323 to 27,643 sequences

Reference alignments based on secondary structure

Reference trees are 75% RAxML bootstrap trees

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

FastTree (FT) and RAxML are ML methods



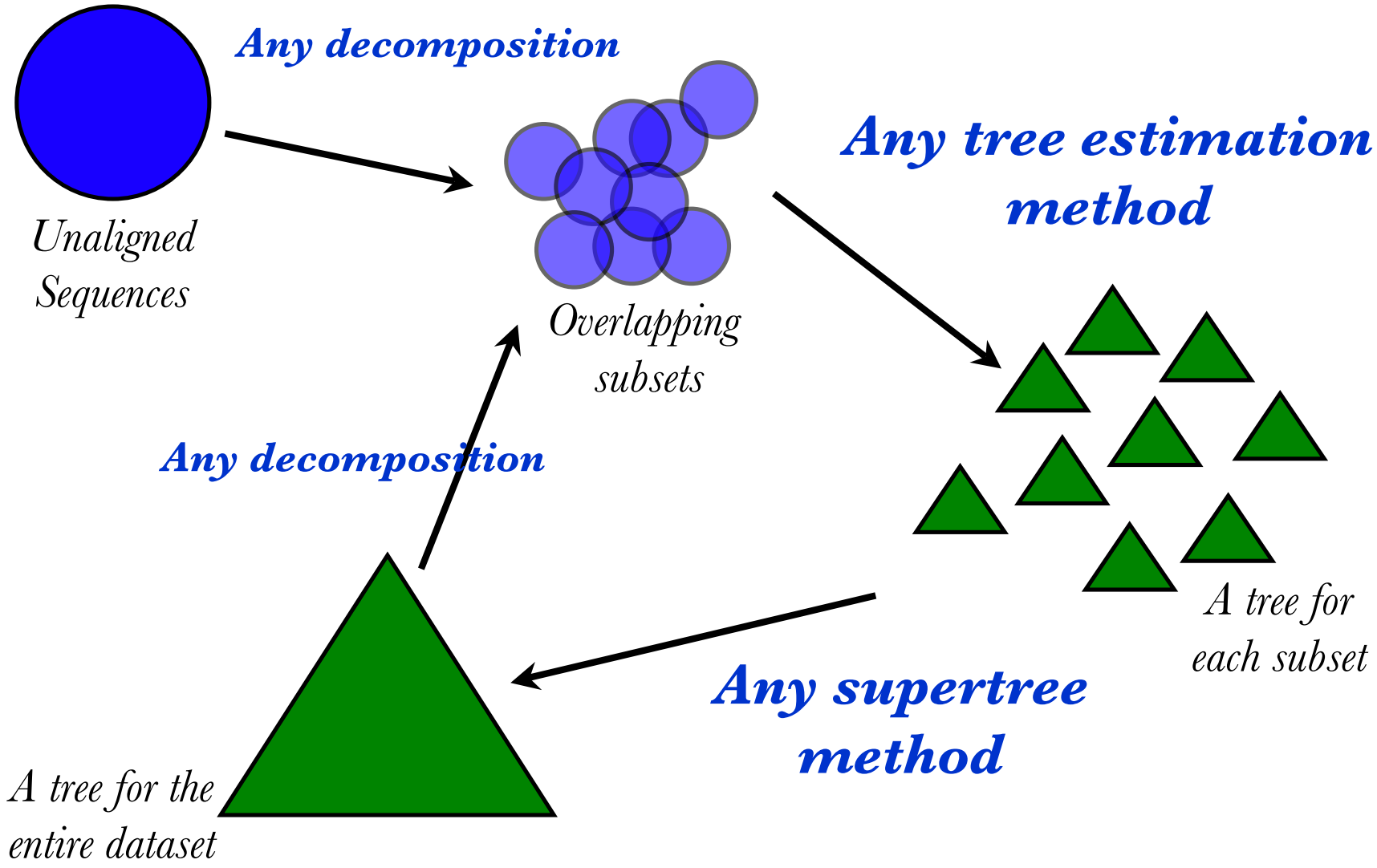
Observations

- SATé and DACTAL outperform two-phase methods with respect to topological accuracy on large, hard-to-align datasets.
- DACTAL outperforms SATé on the largest datasets.
- We do not have any theoretical explanation for why these methods perform well.

Some open questions

- What is the sequence length requirement for maximum likelihood?
- Are trees identifiable under models including long gaps?
- Why do SATé and DACTAL perform well?
- Under standard implementations of ML, gaps are treated as missing data: what are the consequences?

DACTAL


Any decomposition
Unaligned Sequences
Overlapping subsets
Any tree estimation method
A tree for each subset
Any supertree method
Any decomposition
A tree for the entire dataset

Phylogenetic "boosters" (meta-methods)

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

Examples:

- DCM-boosting for distance-based methods (1999)
- DCM-boosting for heuristics for NP-hard problems (1999)
- SATé-boosting for alignment methods (2009)
- SuperFine-boosting for supertree methods (2011)
- DACTAL-boosting for all phylogeny estimation methods (2011)
- SEPP-boosting for metagenomic analyses (2011)

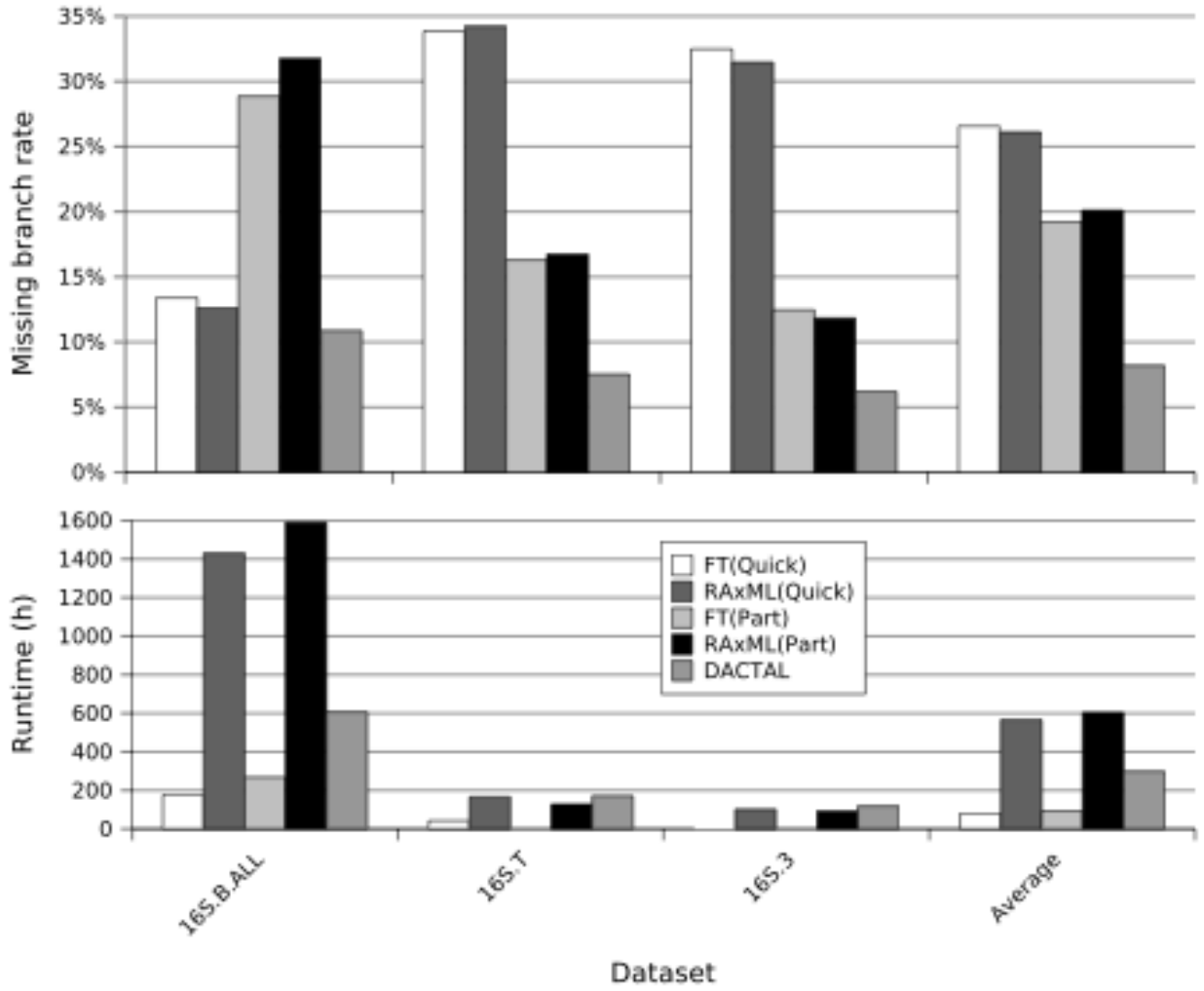
Implications

- Divide-and-conquer methods can greatly improve the accuracy and speed of phylogeny and alignment estimation.
- Theoretical performance doesn’t predict empirical performance.
- Many open questions result from considering phylogeny estimation with indels.

Acknowledgments

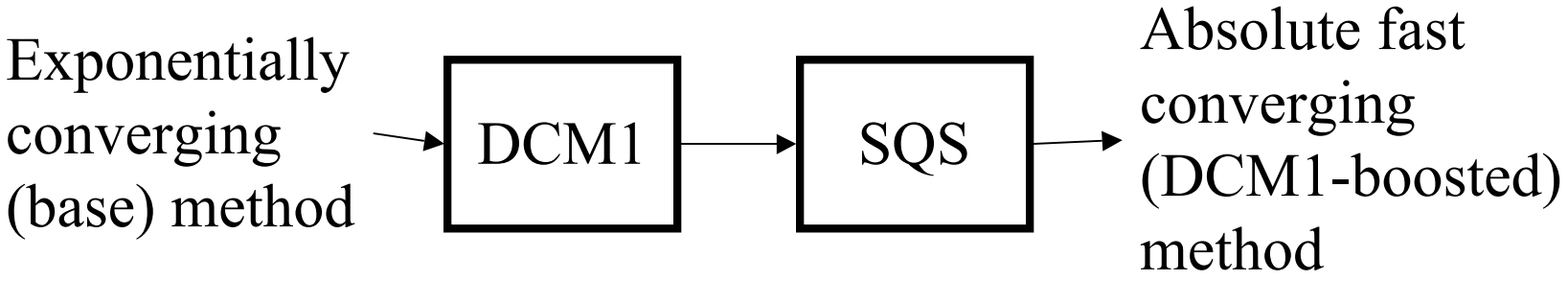
- Microsoft Research New England
- National Science Foundation: Assembling the Tree of Life (ATOL), ITR, and IGERT grants (0733029, 0331453, 0114387)
- The John P. Simon Guggenheim Foundation
- Collaborators: Randy Linder, Bernard Moret, Mark Holder, Jiaye Yu, Alexis Stamatakis, Mike Steel, Katherine St. John, Peter Erdos, Laszlo Szekely, Kevin Liu, Luay Nakhleh, Serita Nelesen, Sindhu Raghavan, Usman Roshan, Jerry Sun, Rahul Suri, Shel Swenson, and Li-San Wang.

DACTAL vs. 2-phase methods



DCM1-boosting:

Warnow, St. John, and Moret, SODA 2001



- The DCM1 phase produces a collection of trees (one for each threshold), and the SQS phase picks the “best” tree.
- For a given threshold, the base method is used to construct trees on small subsets (defined by the threshold) of the taxa. These small trees are then combined into a tree on the full set of taxa.