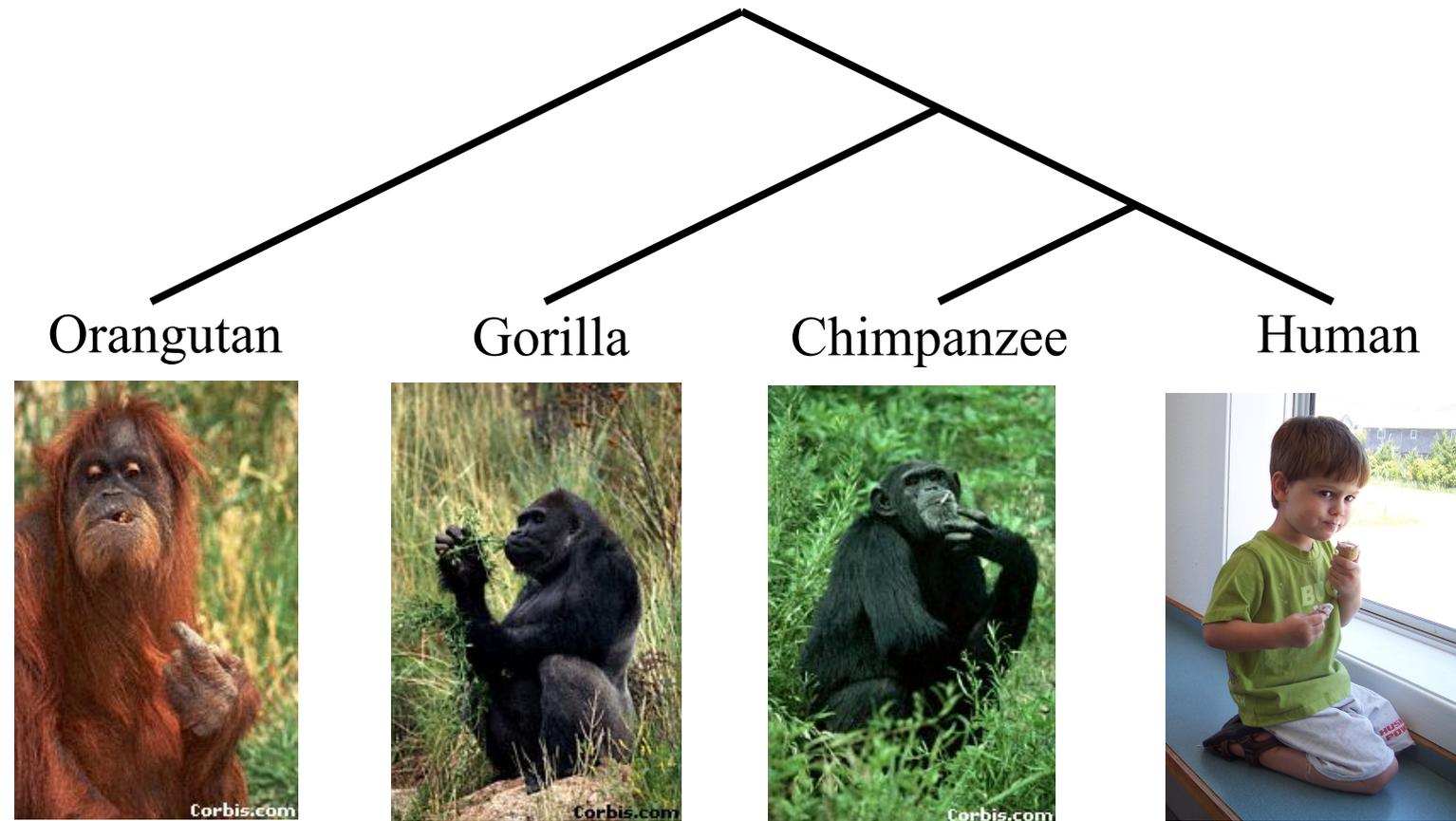


Ultra-large phylogeny estimation

Tandy Warnow

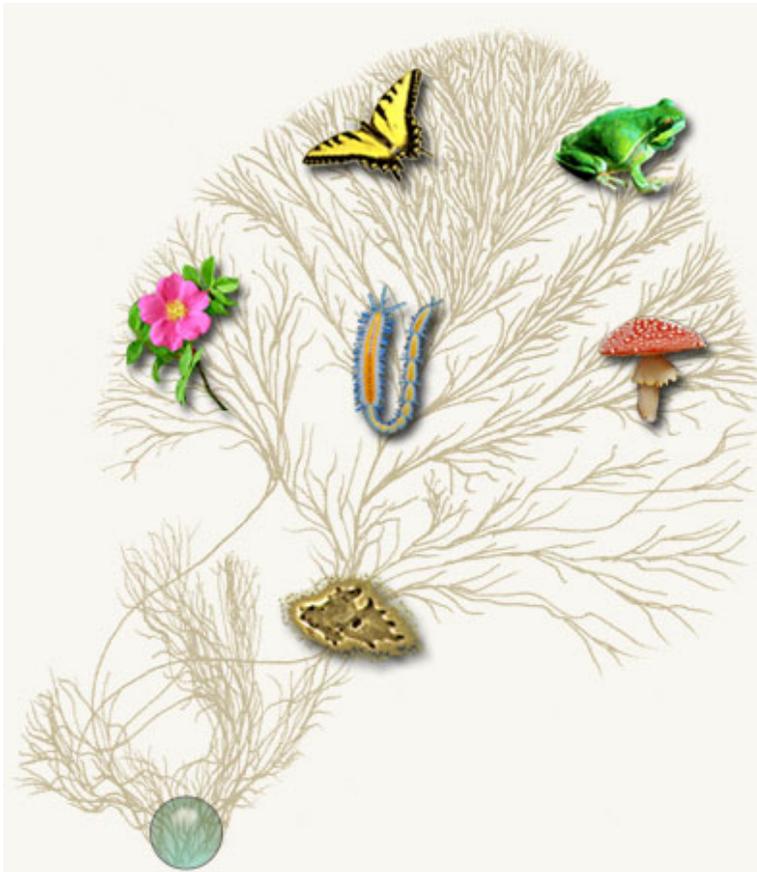
University of Texas at Austin

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

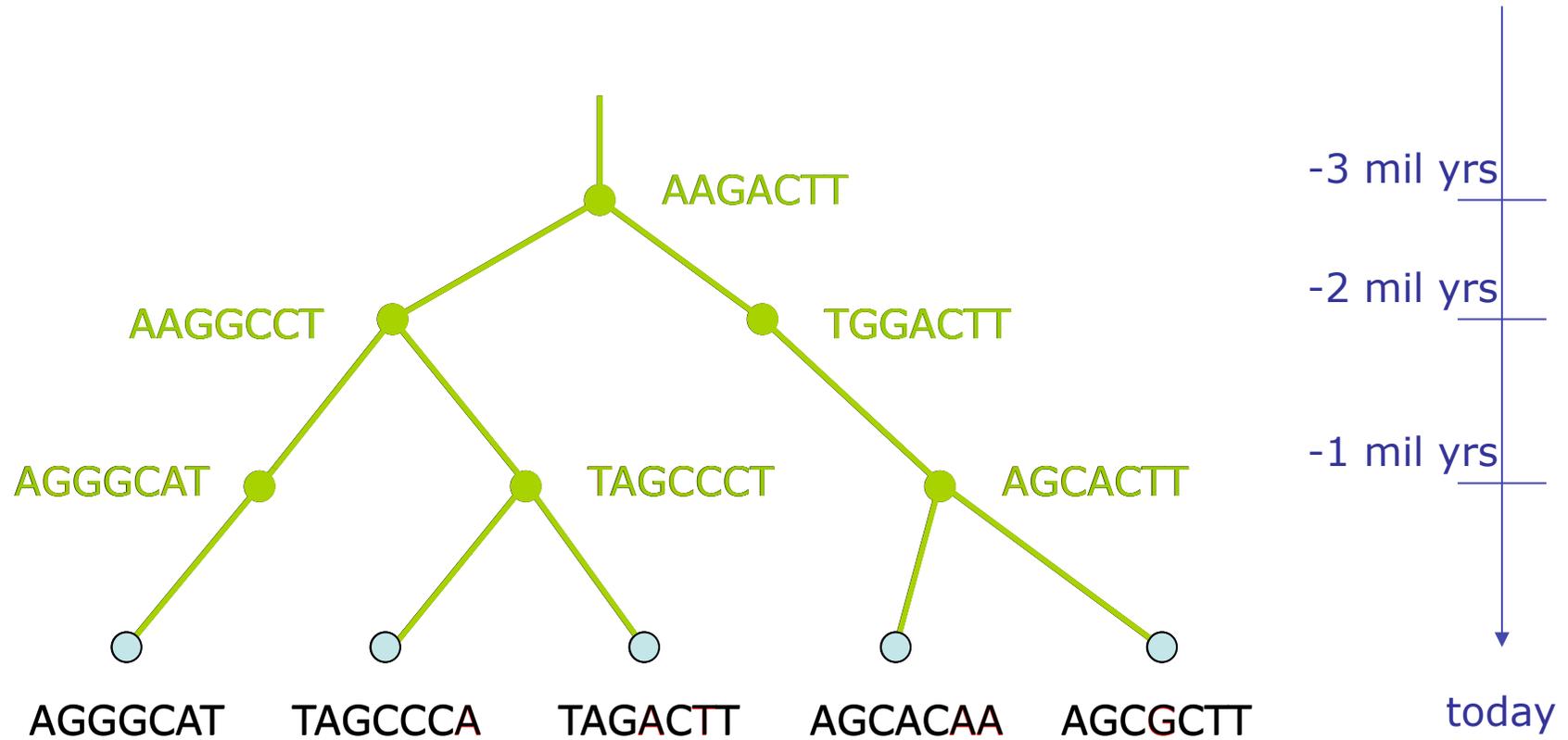
Statistical performance issues

Millions of taxa

Important applications

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

DNA Sequence Evolution



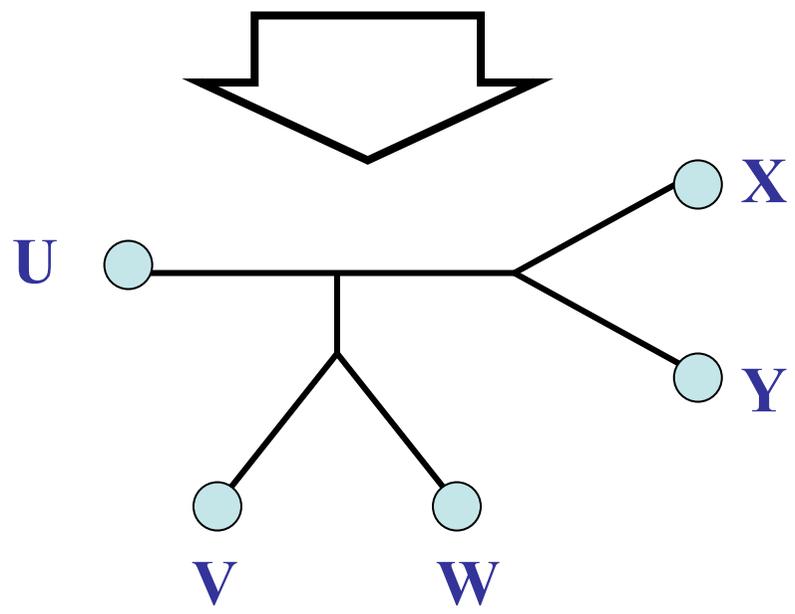
U
AGGGCAT

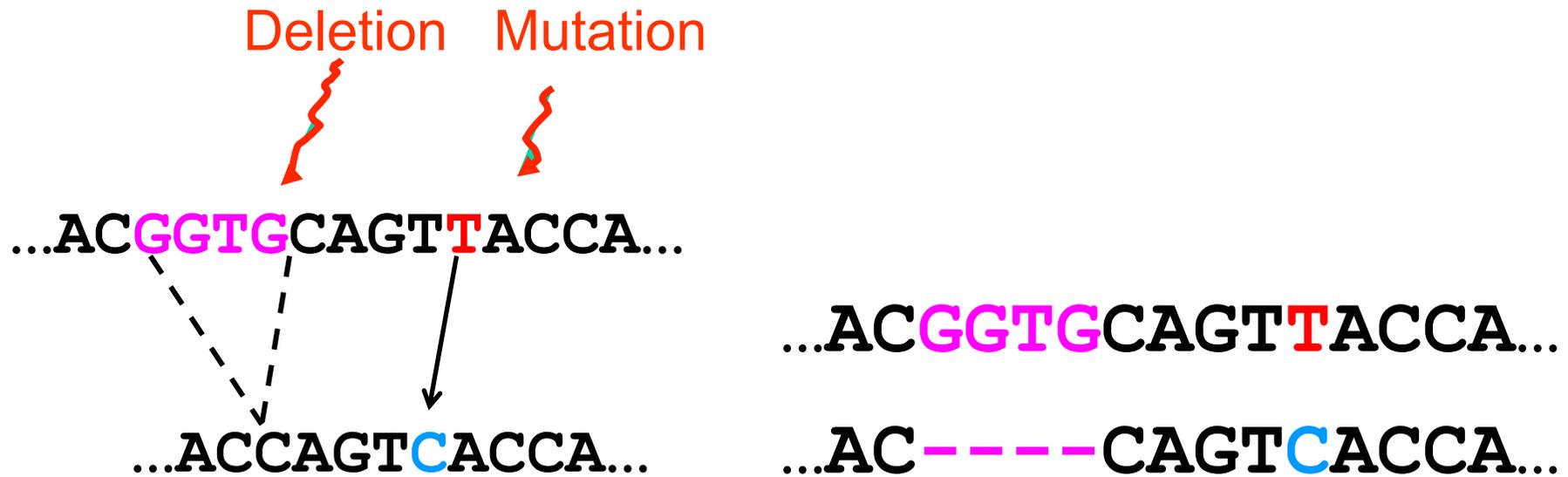
V
TAGCCCA

W
TAGACTT

X
AGCACAA

Y
AGCGCTT





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)

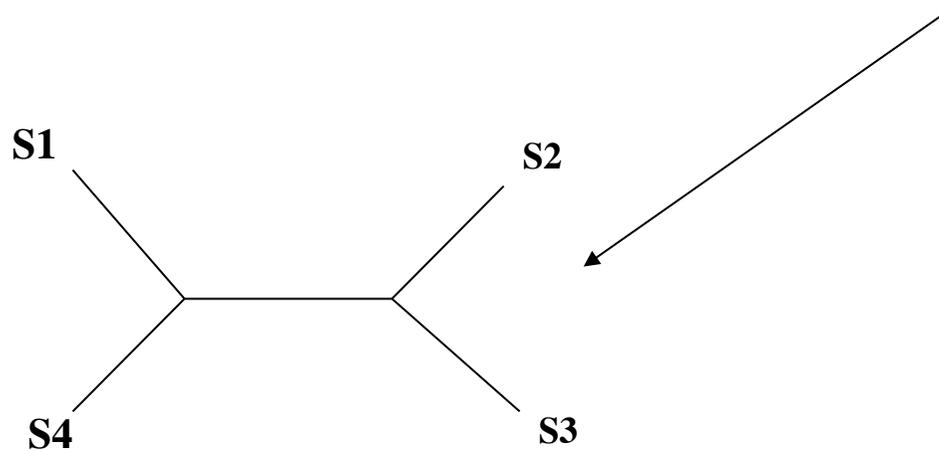
S1 = AGGCTATCACCTGACCTCCA	→	S1 = -AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC		S2 = TAG-CTATCAC--GACCGC--
S3 = TAGCTGACCGC		S3 = TAG-CT-----GACCGC--
S4 = TCACGACCGACA		S4 = -----TCAC--GACCGACA

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.

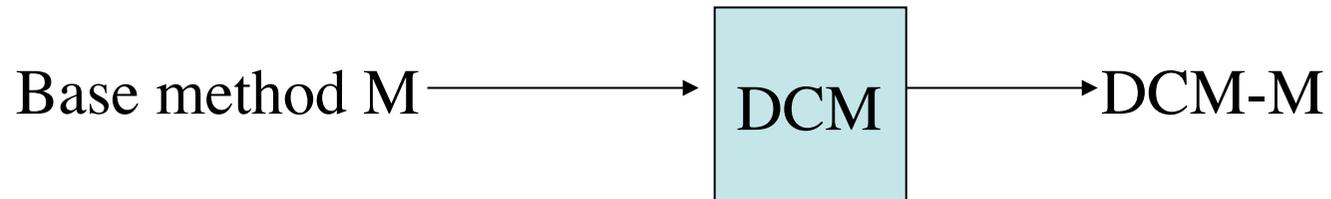
Phylogenetic “boosters” (meta-methods)

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

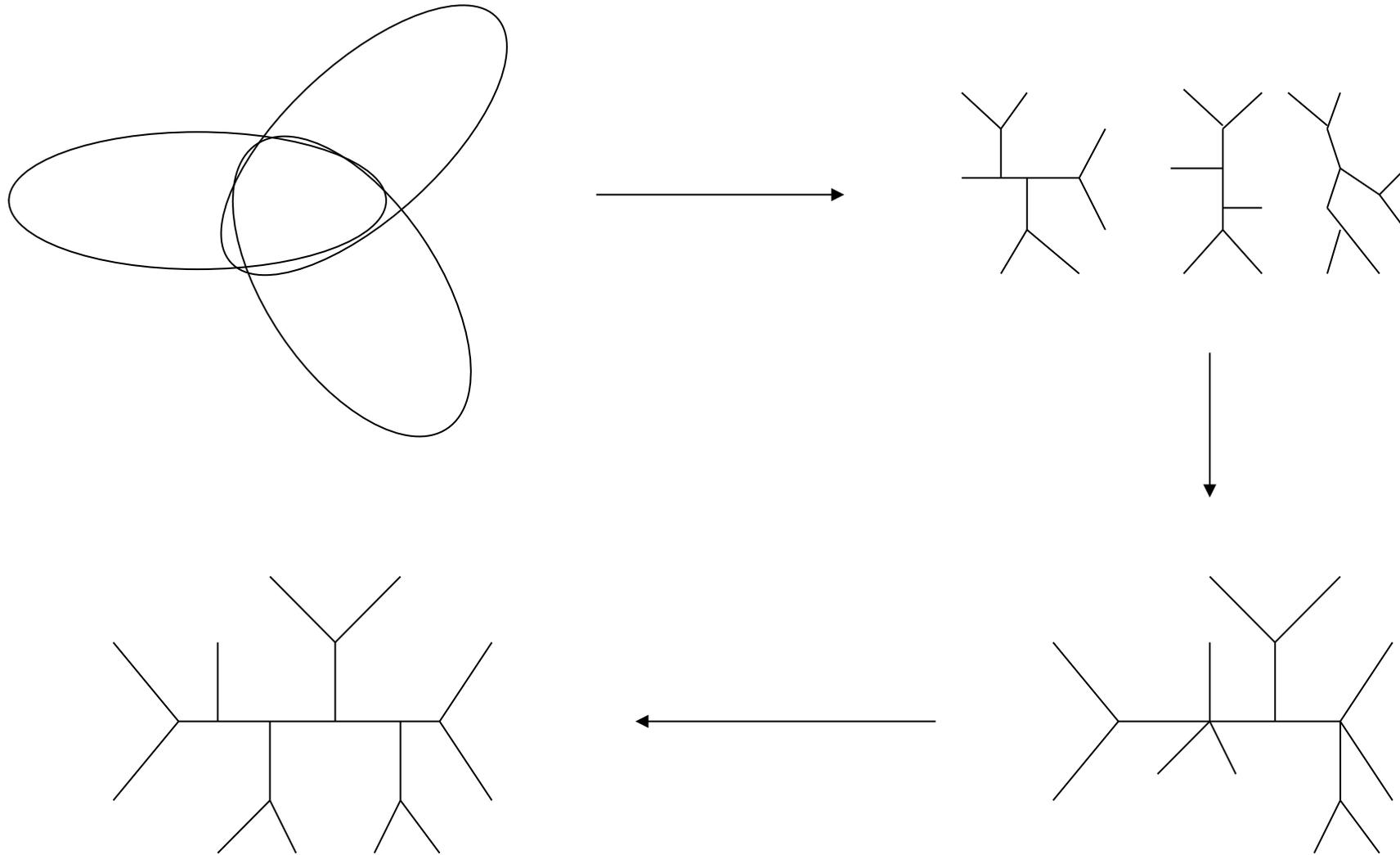
Examples:

- DCM-boosting for heuristics for NP-hard problems (e.g., gene order phylogeny)
- DCM-boosting for distance-based methods
- SATé-boosting for alignment methods
- SuperFine-boosting for supertree methods
- DACTAL-boosting for all methods

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



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



Graph-theoretic divide-and-conquer (DCM's)

- Define a **triangulated** graph so that its vertices correspond to the input taxa
- Compute a decomposition of the graph into overlapping subgraphs, thus defining a decomposition of the taxa into overlapping subsets.
- Apply the “base method” to each subset of taxa, to construct a subtree
- Merge the subtrees into a single tree on the full set of taxa.

Phylogenetic “boosters” (meta-methods)

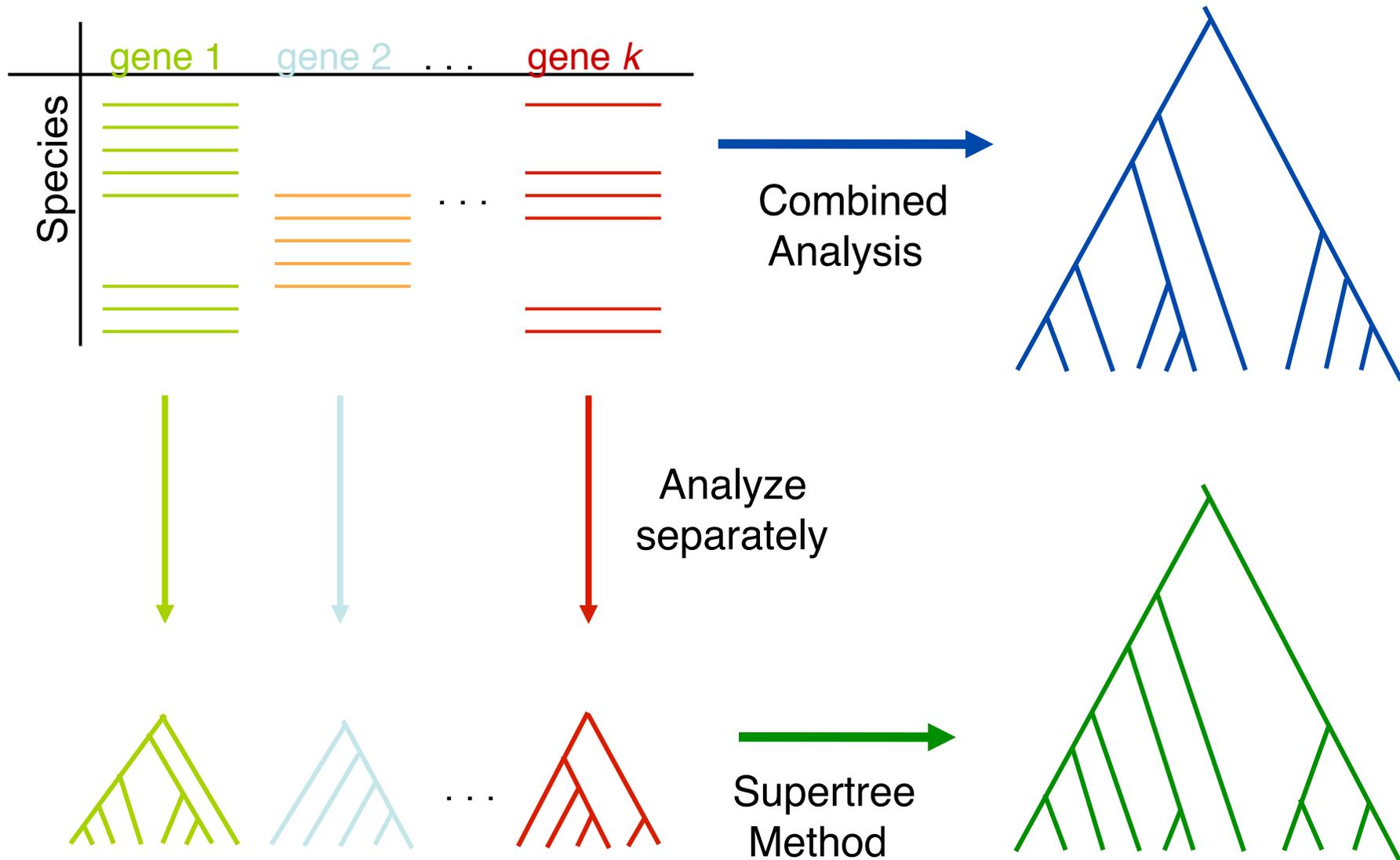
Goal: improve accuracy, speed, robustness, or theoretical guarantees of base methods

Examples:

- DCM-boosting for heuristics for NP-hard problems (e.g., gene order phylogeny)
- SuperFine-boosting for supertree methods
- DCM-boosting for distance-based methods
- SATé-boosting for alignment methods
- DACTAL-boosting for all methods

Part I: SuperFine

Multi-marker phylogenetics

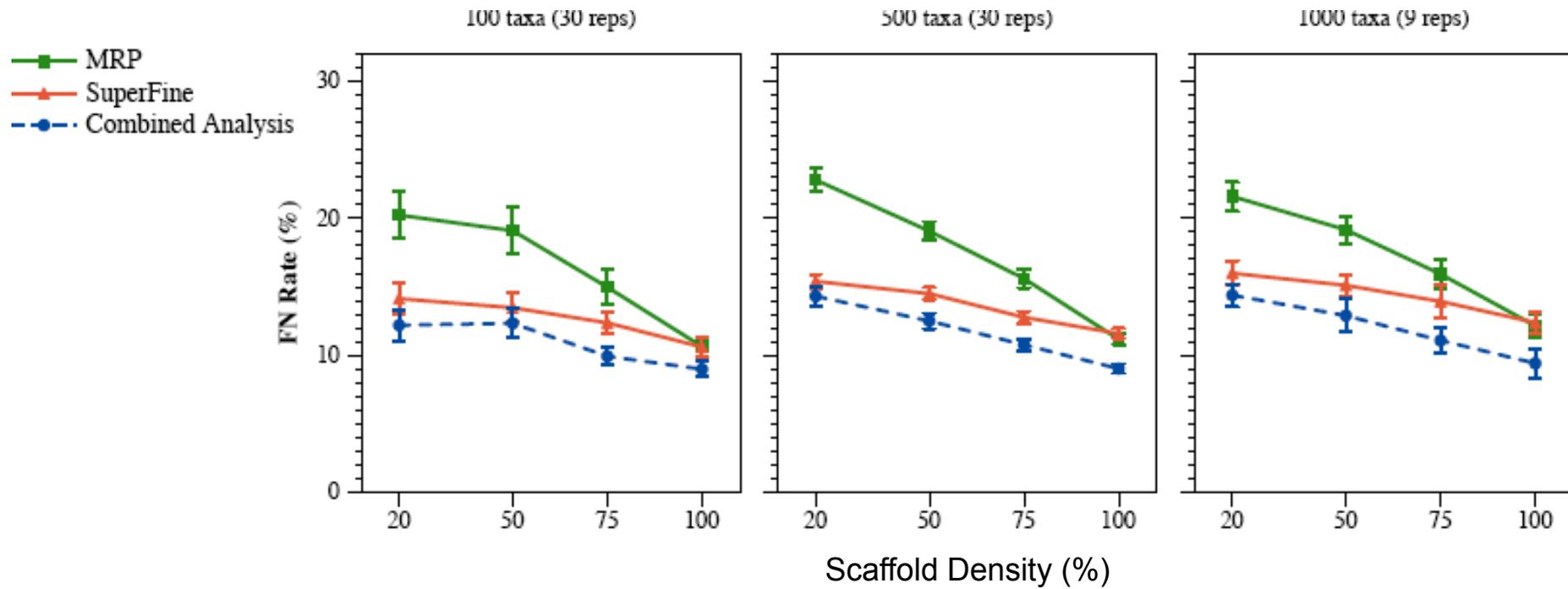


SuperFine: new supertree method

- Step 1: construct a partially resolved constraint tree, using The Strict Consensus Merger (SCM) (Huson et al. 1999)
- Step 2: use favorite supertree method to refine each polytomy (e.g., MRP)

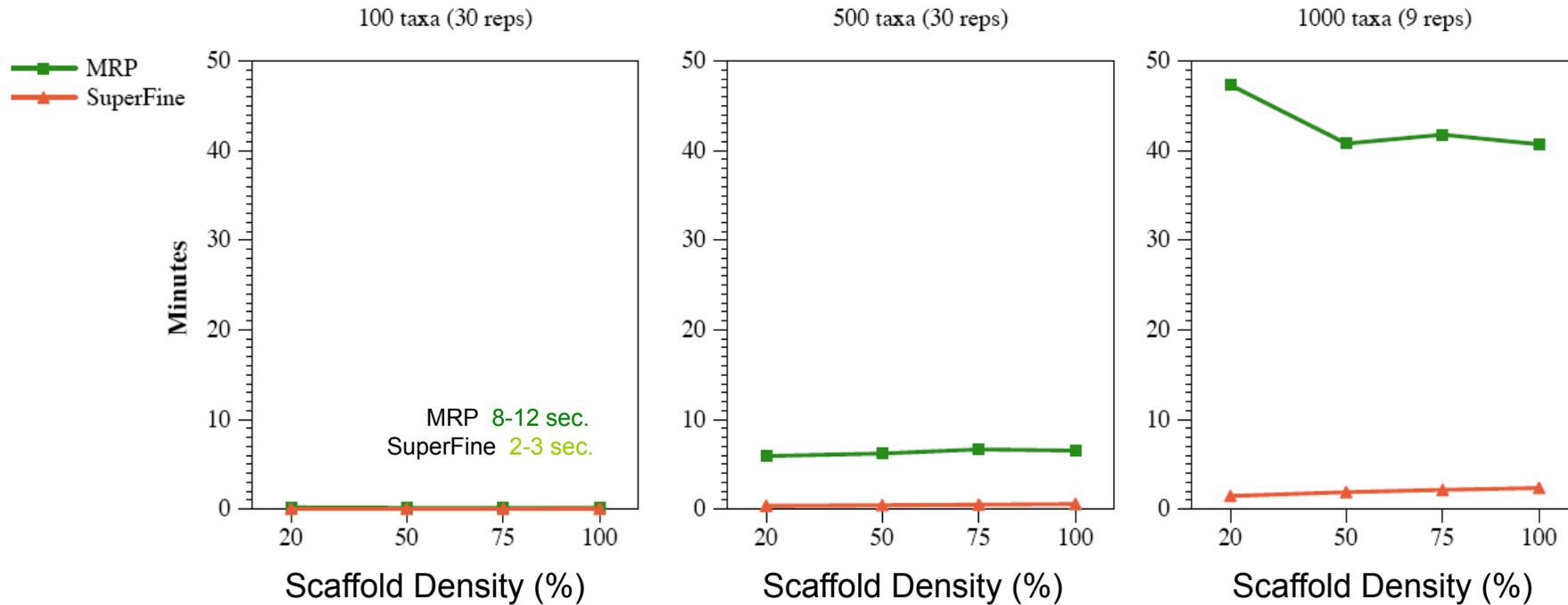
Swenson, Suri, Linder, and Warnow, in press SystBio

False Negative Rate



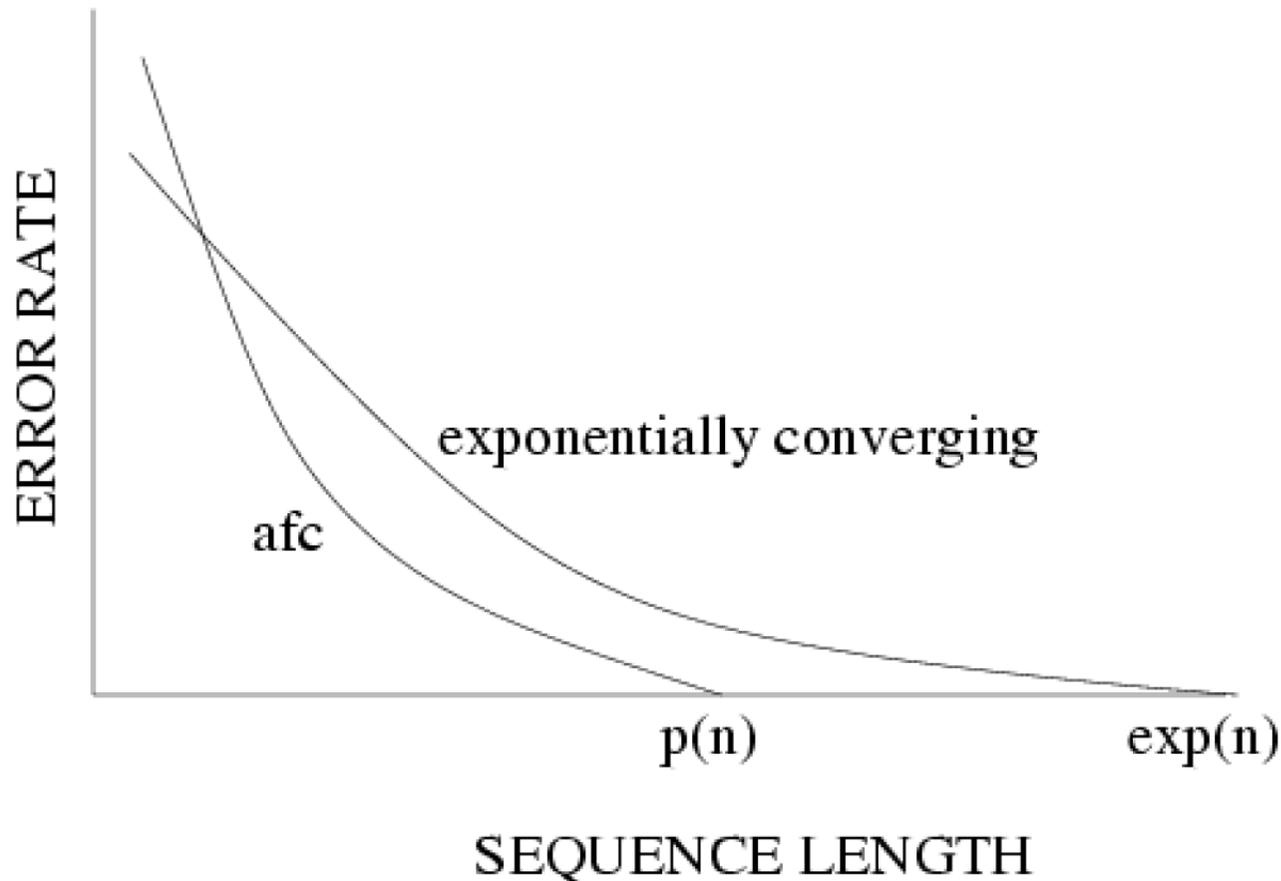
Running Time

SuperFine vs. MRP



Part II: DCM-boosting for distance-based methods

Statistical consistency, convergence rates, and absolute fast convergence



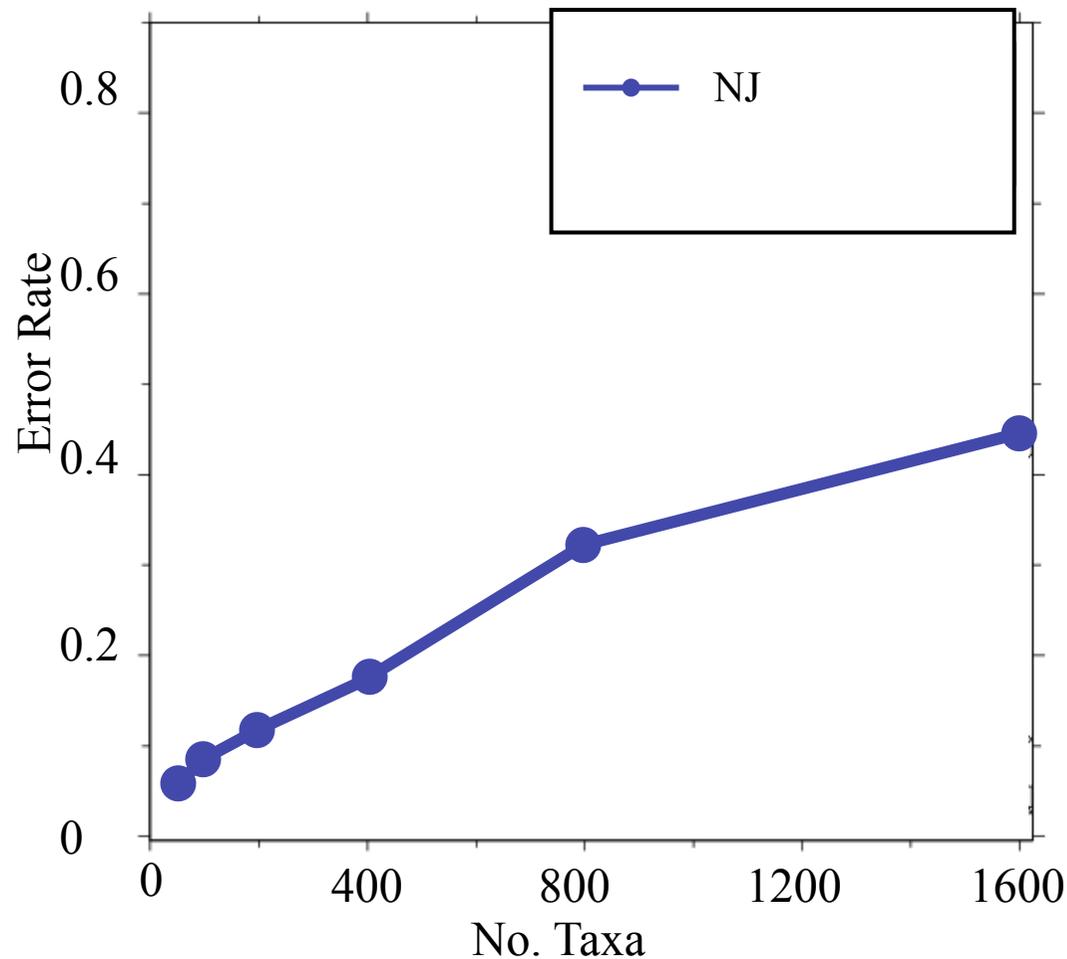
First afc methods in the Erdos, Steel, Szekely and Warnow papers (ICALP 1997, RSA 1999, TCS 1999)

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 $O(\lg n e^{\max\{D_{ij}\}})$.
- Note: $\max \{D_{ij}\} = O(g \text{ diam}(T))$, where g is the maximum length of any edge.

Neighbor joining has poor accuracy on large diameter model trees

[Nakhleh et al. ISMB 2001]

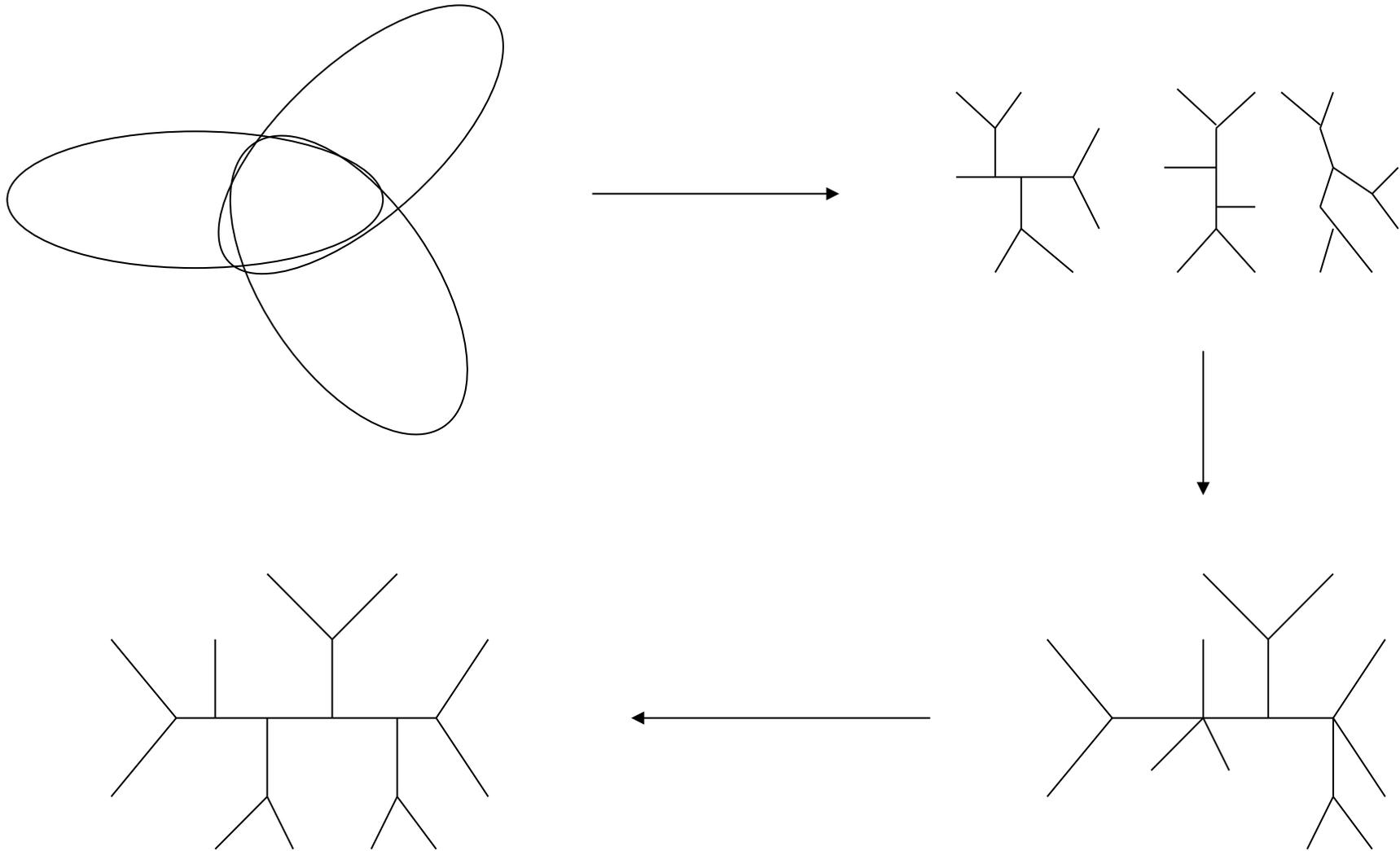


Simulation study

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.

Divide-and-conquer for phylogeny estimation



DCM1 Decompositions

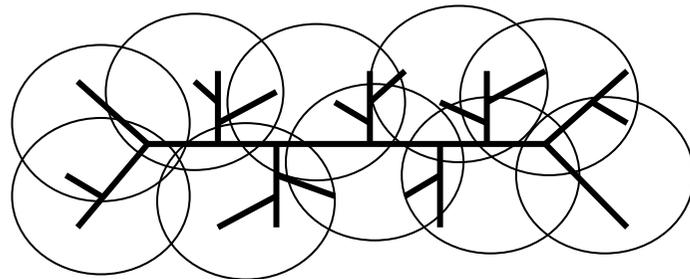
Input: Set S of sequences, distance matrix d , threshold value $q \in \{d_{ij}\}$

1. Compute threshold graph

$$G_q = (V, E), V = S, E = \{(i, j) : d(i, j) \leq q\}$$

2. Perform minimum weight triangulation (note: if d is an additive matrix, then the threshold graph is provably **triangulated**).

DCM1 decomposition : Compute maximal cliques



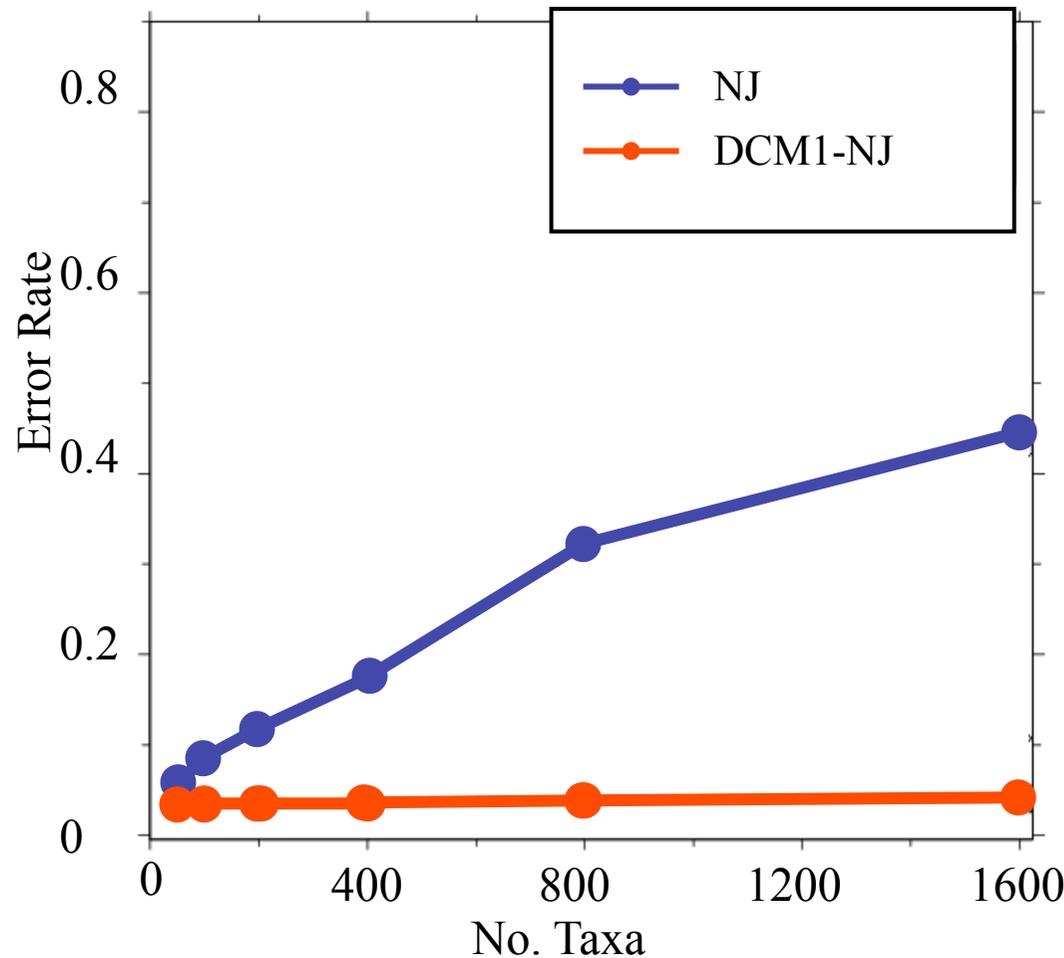
Huson, Nettles and Warnow, RECOMB 1999

Improving upon NJ

- Construct trees on a number of smaller diameter subproblems, and merge the subtrees into a tree on the full dataset.
- Our approach:
 - Phase I: produce $O(n^2)$ trees (one for each diameter)
 - Phase II: pick the “best” tree from the set.

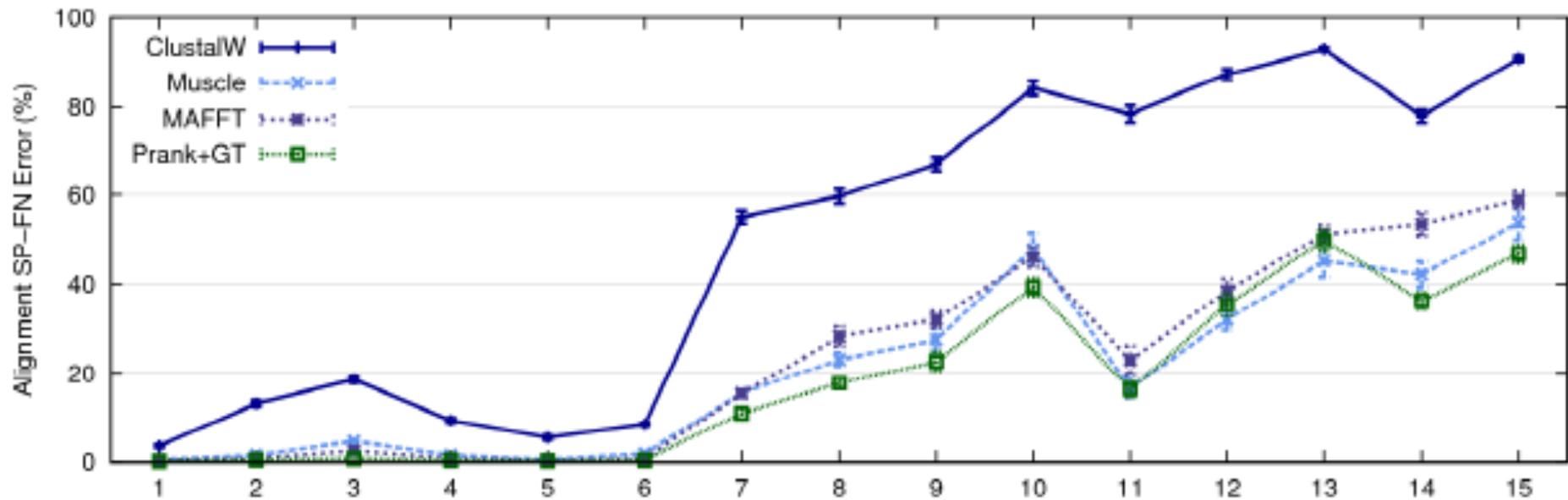
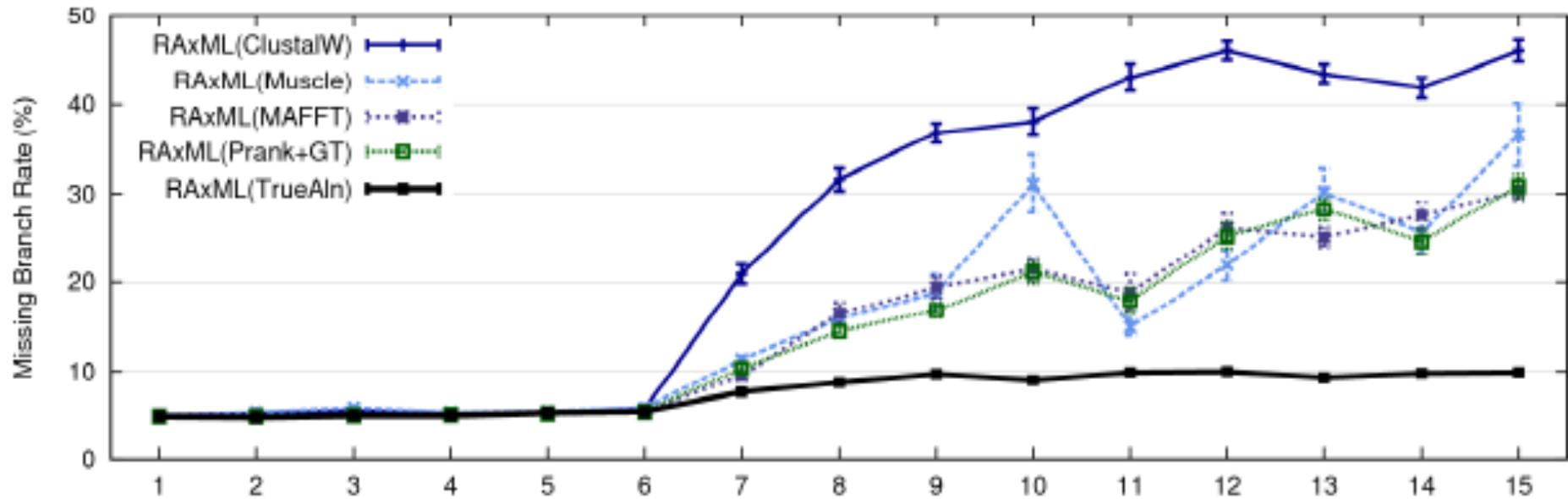
DCM1-boosting distance-based methods

*[Nakhleh et al. ISMB 2001 and
Warnow, Moret and St. John, SODA 2001]*



Theorem:
DCM1-NJ is afc
(converges to
the true tree
from **polynomial**
length
sequences)

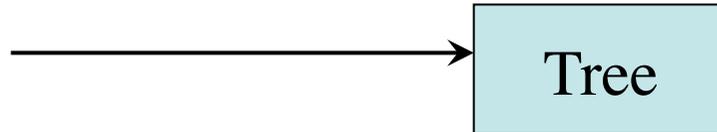
Part III: SATé-boosting for alignment estimation



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

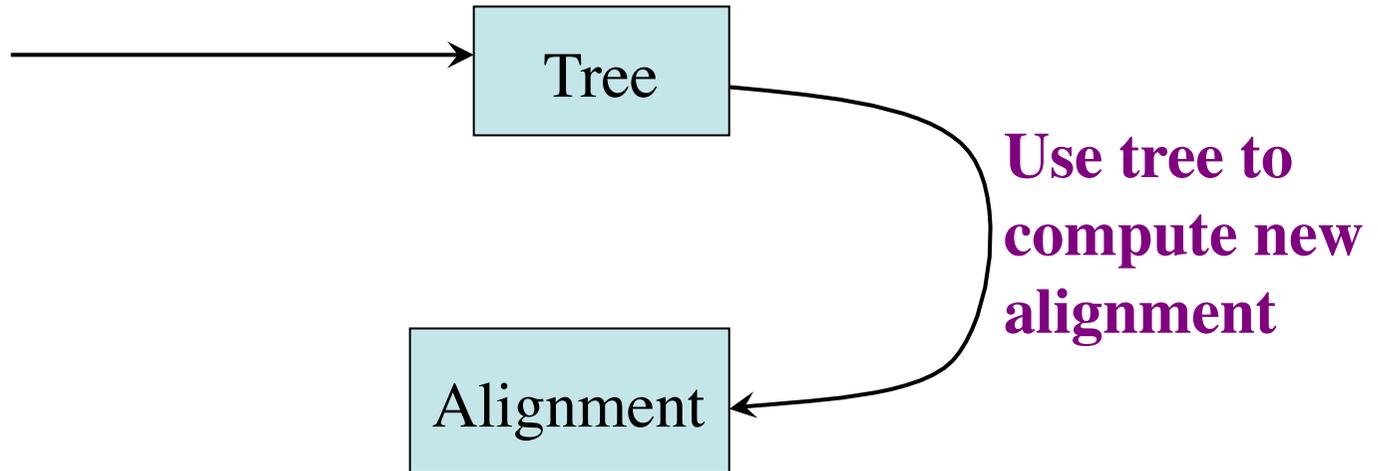
SATé Algorithm

Obtain initial alignment
and estimated ML tree



SATé Algorithm

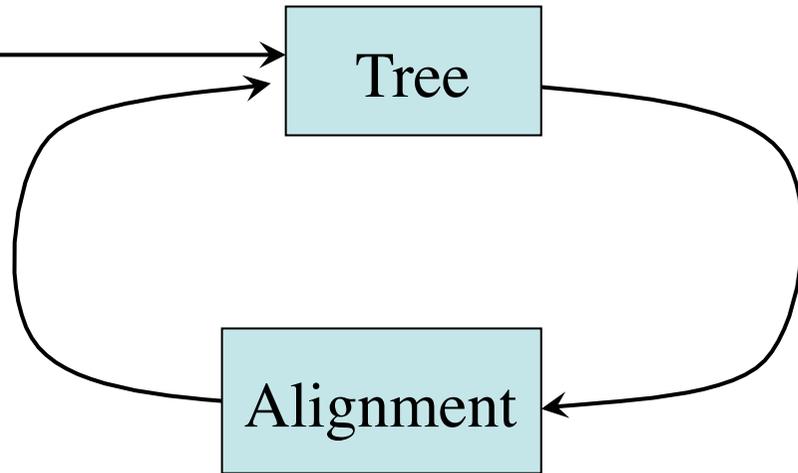
Obtain initial alignment
and estimated ML tree



SATé Algorithm

Obtain initial alignment
and estimated ML tree

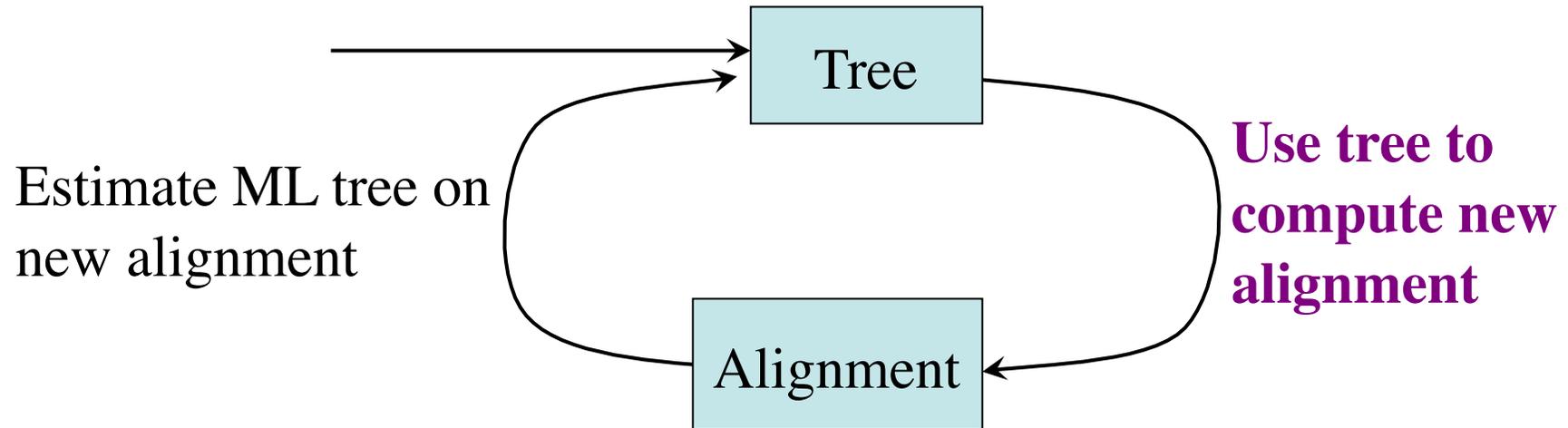
Estimate ML tree on
new alignment



**Use tree to
compute new
alignment**

SATé Algorithm

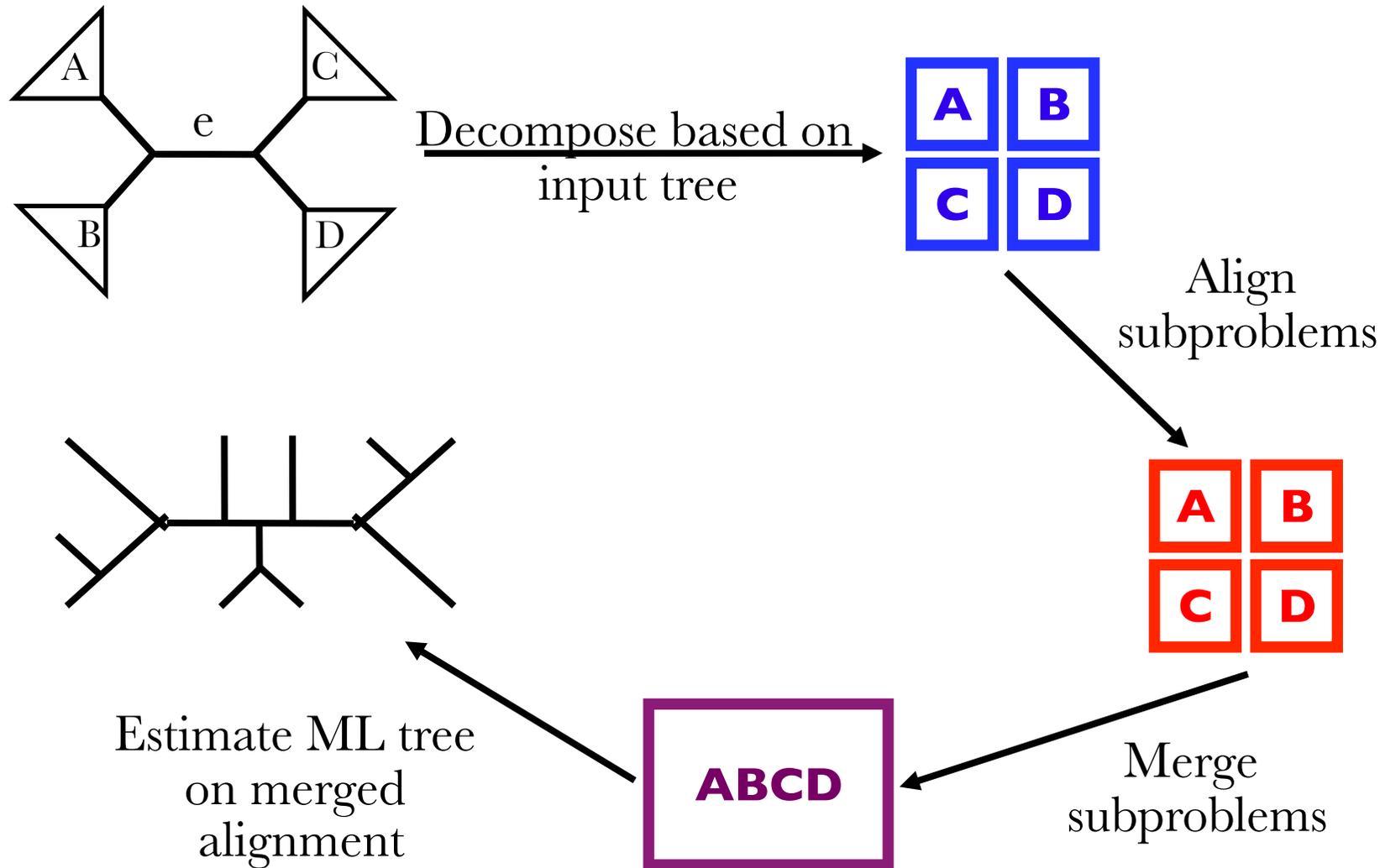
Obtain initial alignment
and estimated ML tree

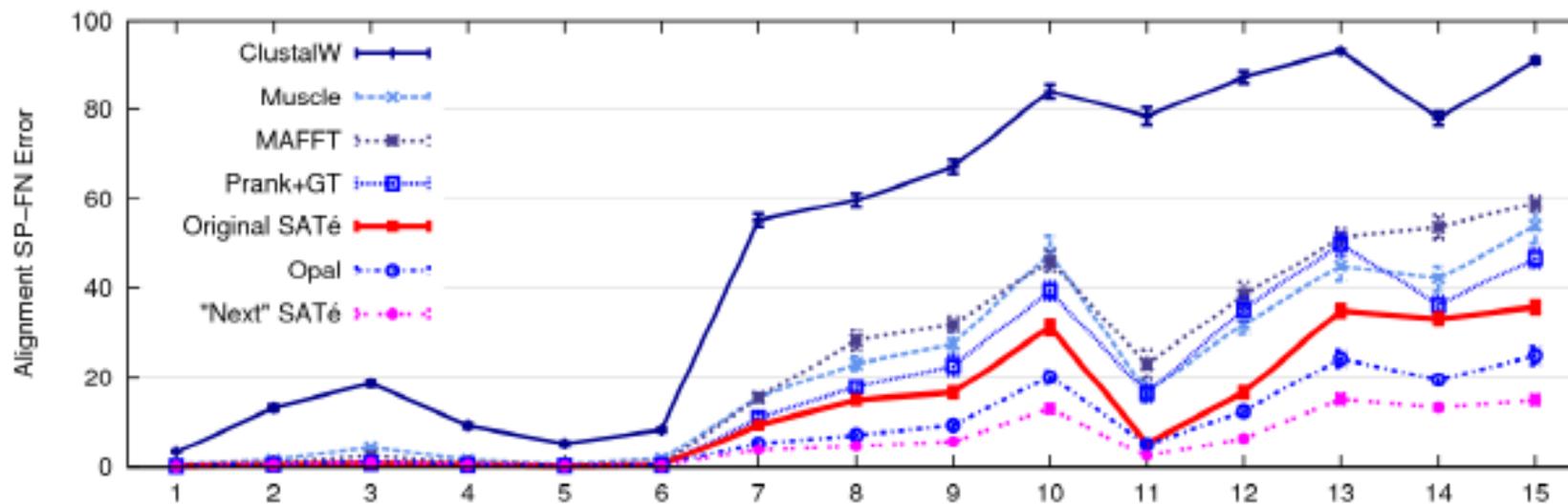
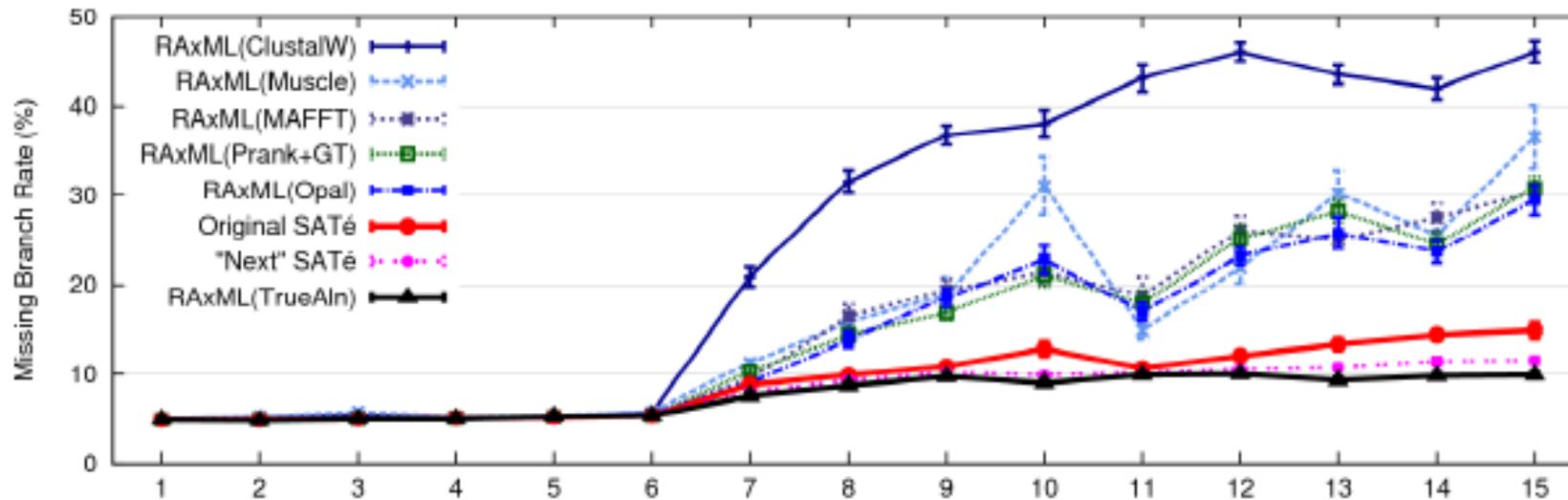


If new alignment/tree pair has worse ML score, realign using a different decomposition

Repeat until termination condition (typically, 24 hours)

One SATé iteration (cartoon)



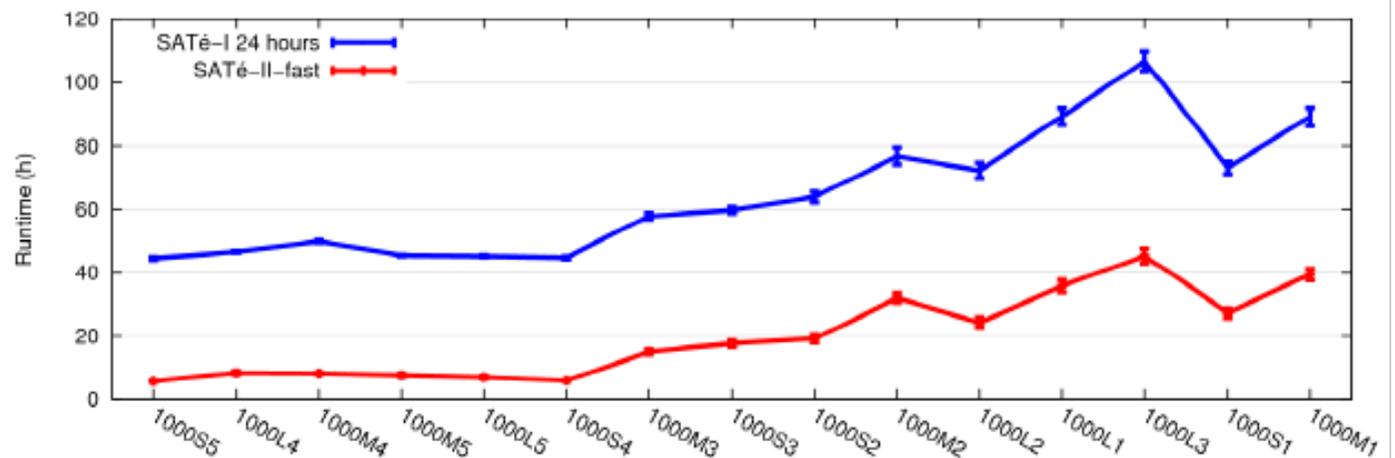
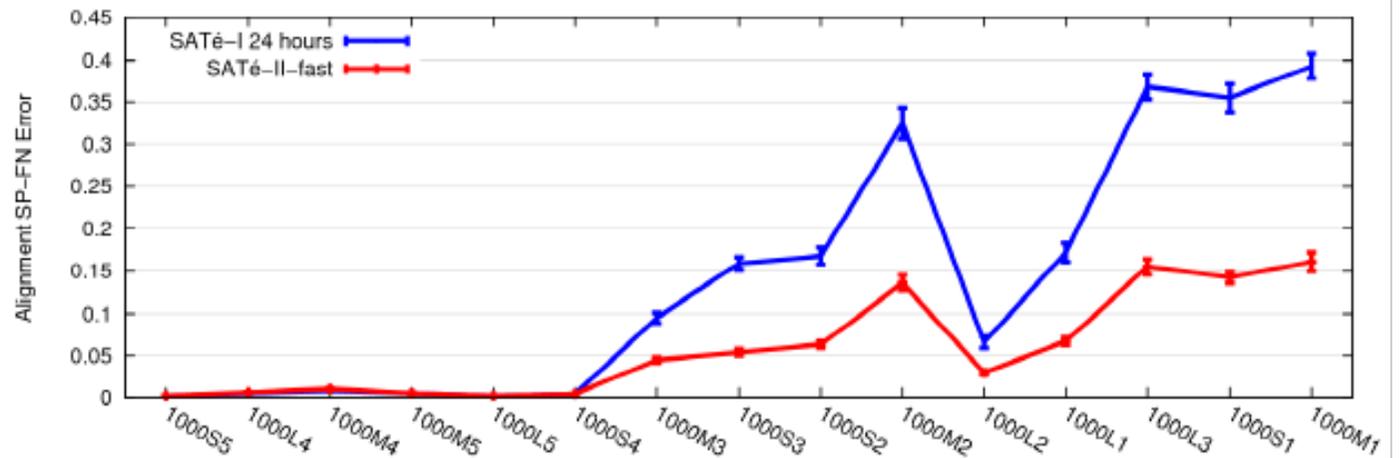
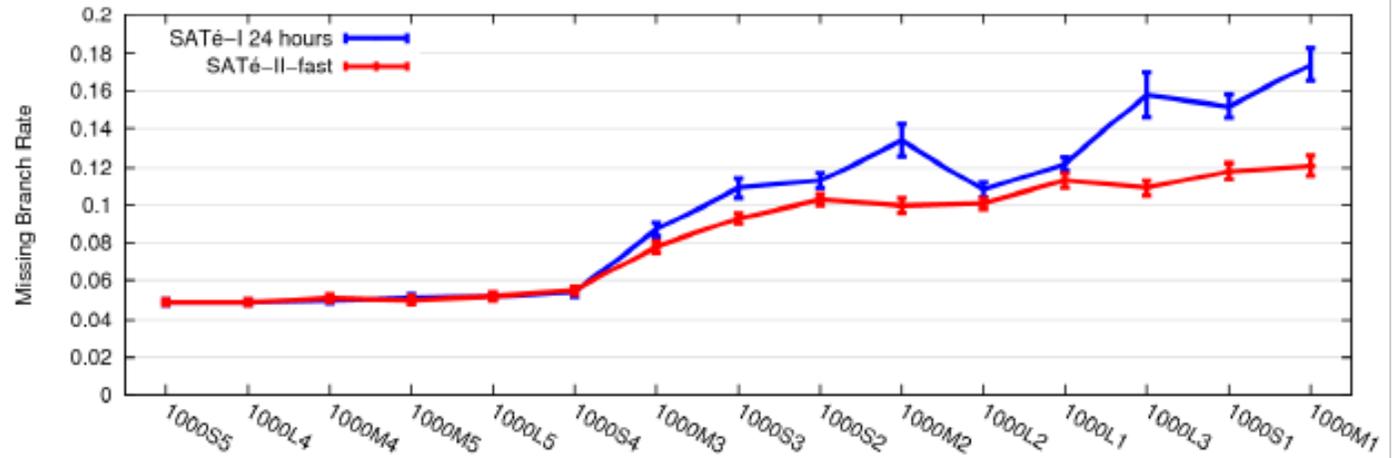


SATé compared to two-phase methods on 1000 taxon datasets
 Original SATé: Liu et al. Science 2009
 "Next" SATé: Liu et al., Systematic Biology (in press)

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

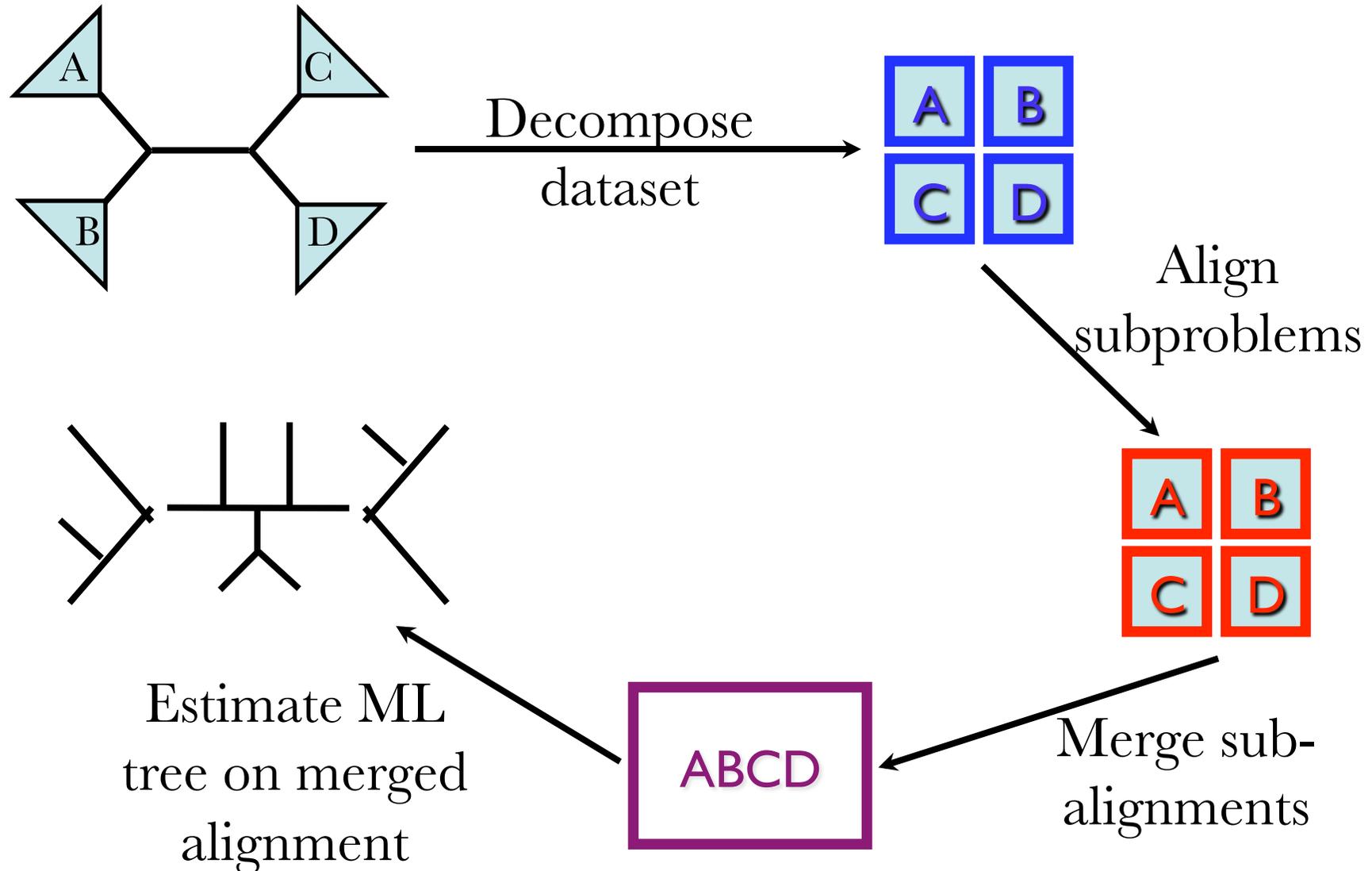


Why does SATé work well?

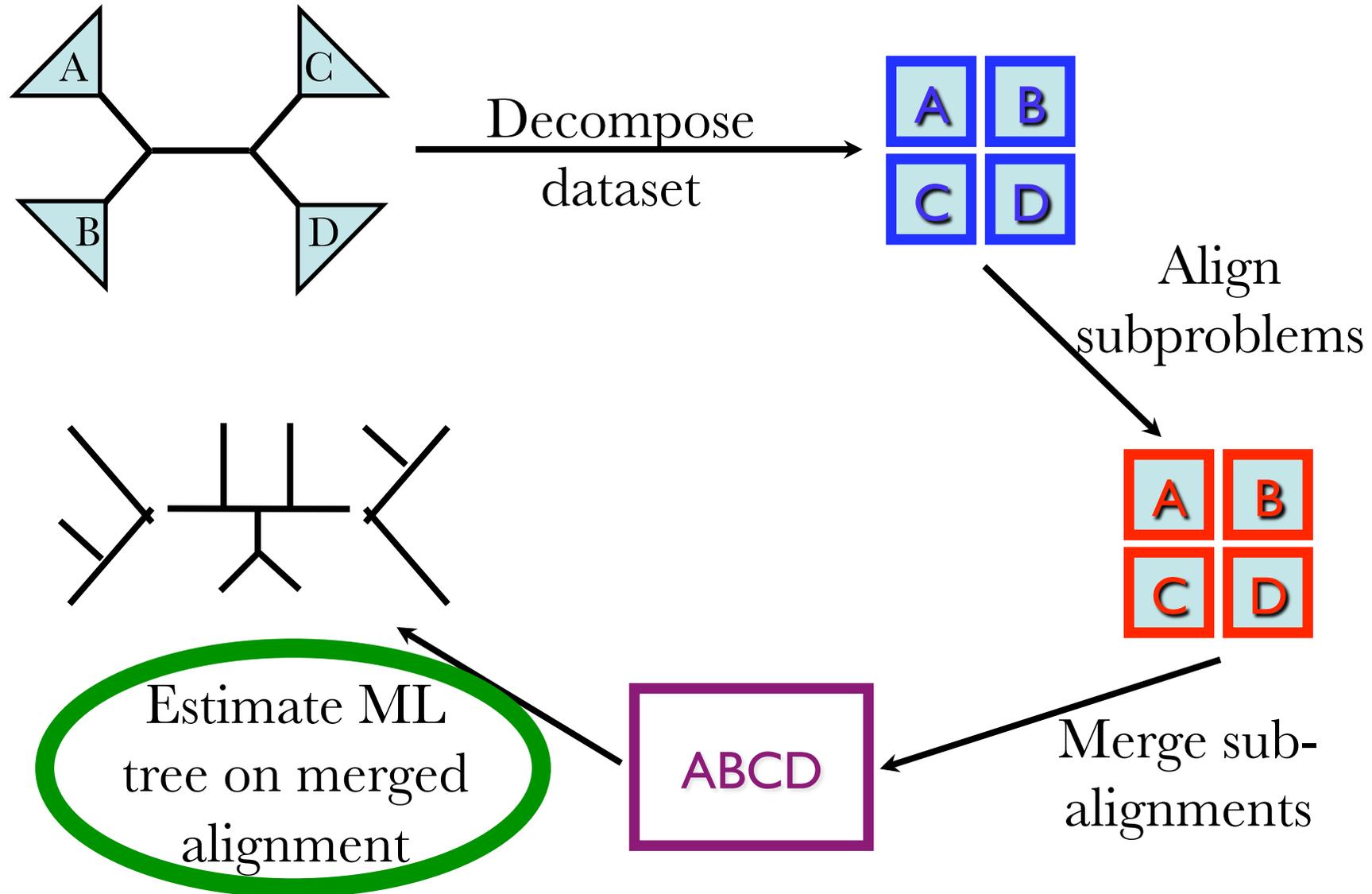
- Not because it finds the best alignment to optimize ML, treating gaps as missing data!
- Theorem: under Jukes-Cantor, if we allow the alignment to change arbitrarily, then the best alignment is “mono-typic”, and all trees are optimal.

Liu et al., Systematic Biology, in press

Each SATé iteration



Limitations of SATé-I and -II



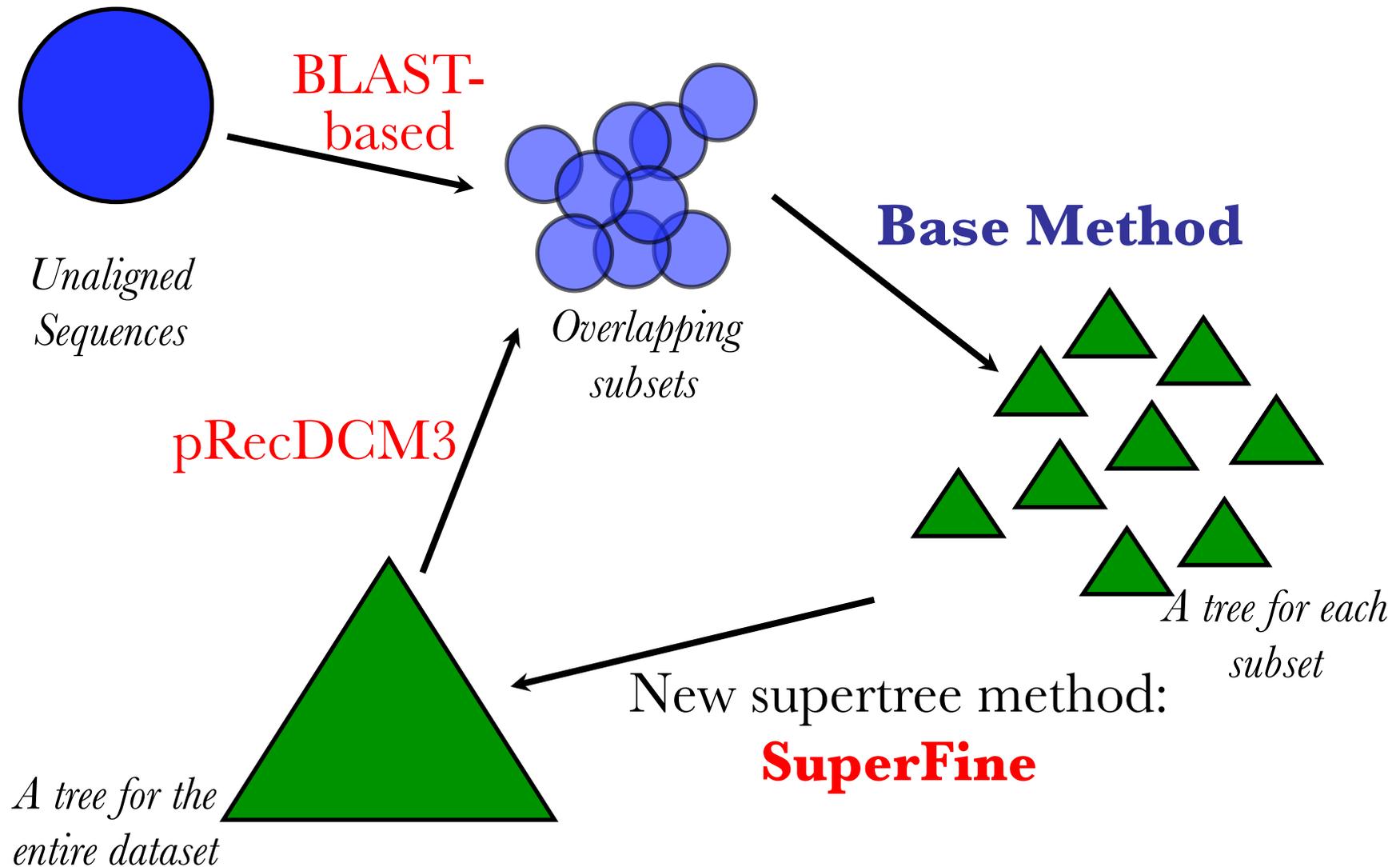
Part IV: DACTAL-boosting

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

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

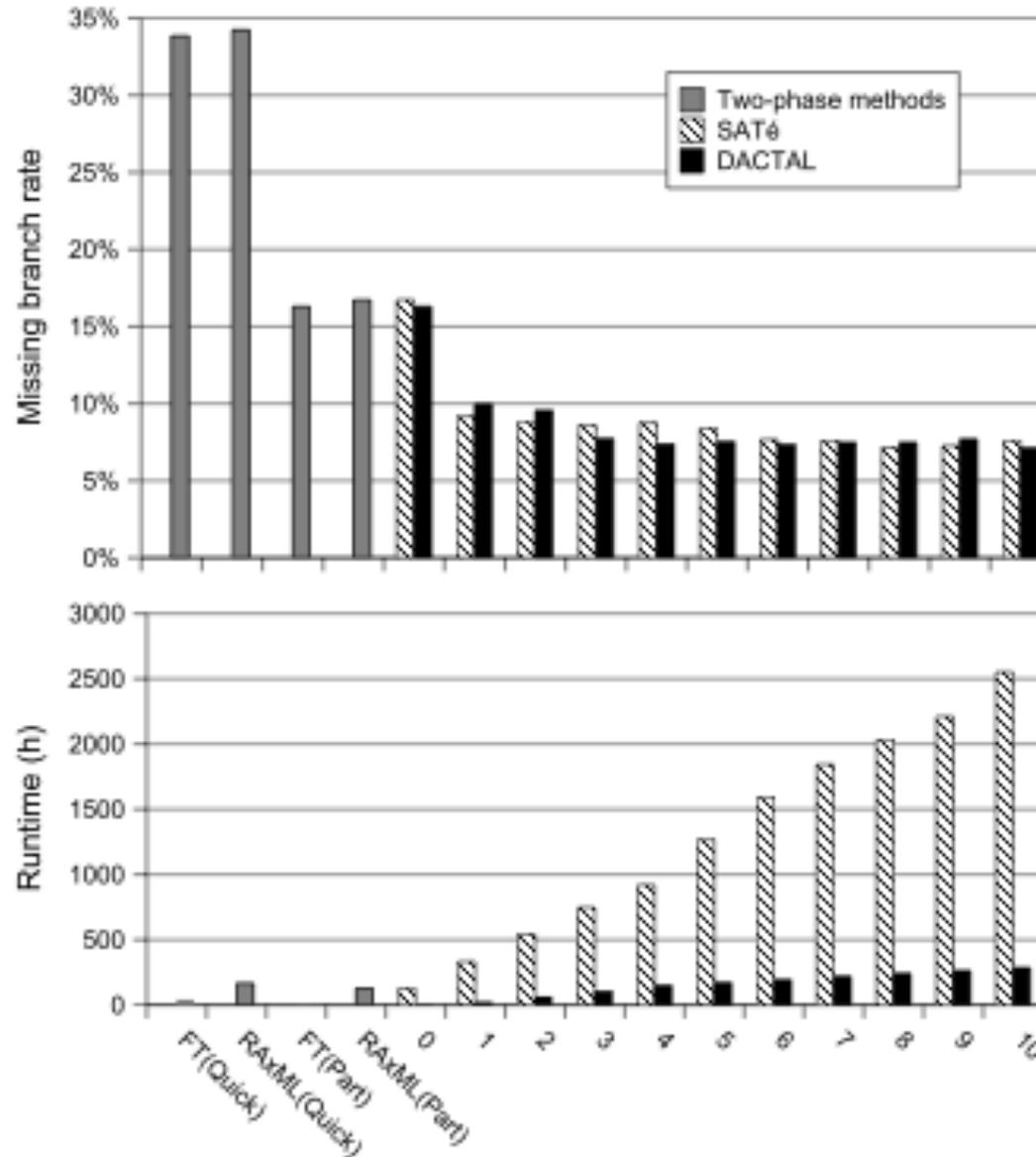
Nelesen, Liu, Wang, Linder, and Warnow, in preparation

DACTAL-boosting



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

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



DACTAL-boosting

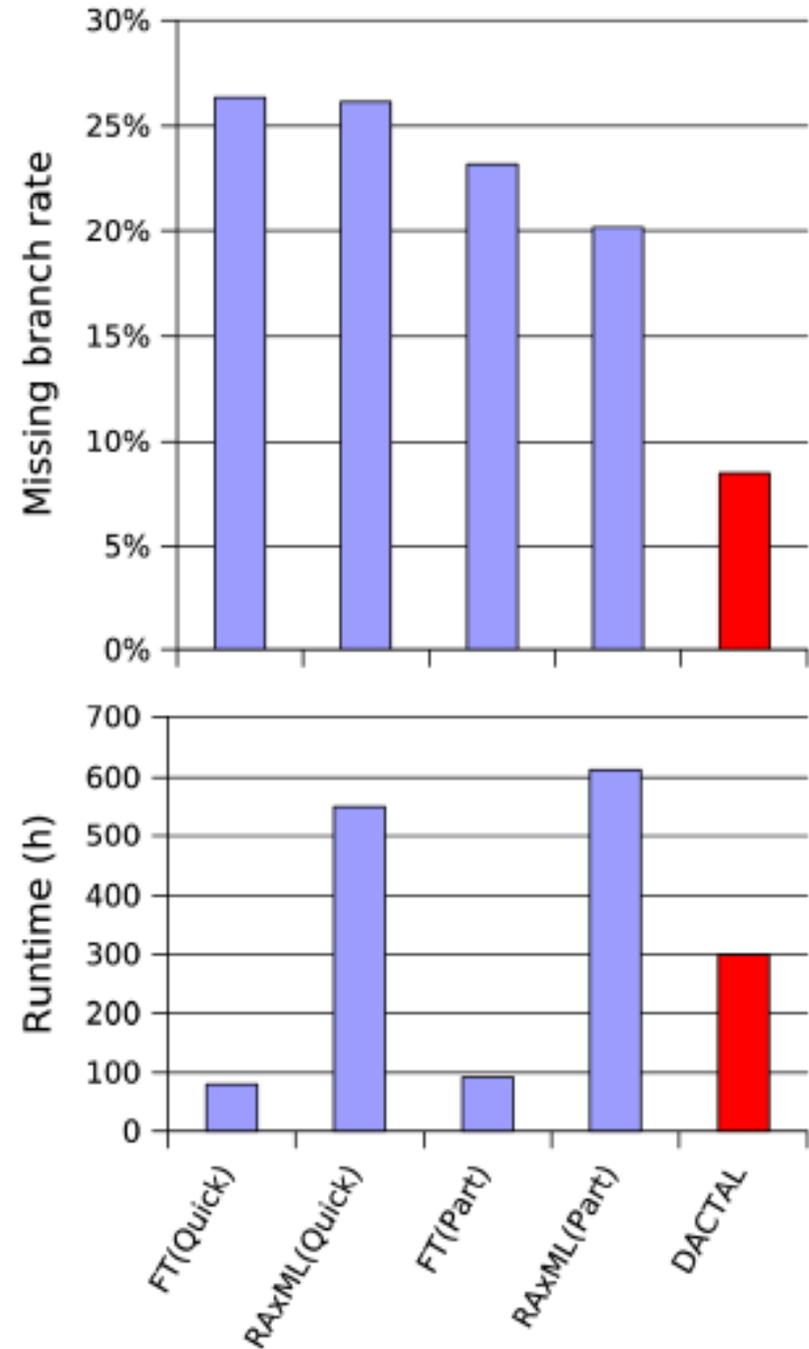
Base method: ML(MAFFT)

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



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

Software

- Superfine: available through Dryad, doi: 10.5061/dryad.879st
- SATé: <http://phylo.bio.ku.edu/software/sate/sate.html>
- DACTAL: in preparation

Acknowledgments

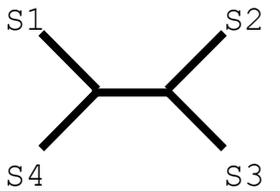
- Funding: Packard Foundation, NSF, Guggenheim Foundation, and Microsoft Research New England
- Collaborators: Mark Holder, Daniel Huson, Randy Linder, Kevin Liu, Bernard Moret, Luay Nakhleh, Serita Nelesen, Scott Nettles, Sindhu Raghavan, Usman Roshan, Alexis Stamatakis, Jerry Sun, Katherine St. John, Li-San Wang, and Jiaye Yu.

Simulation Studies

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

Unaligned
Sequences

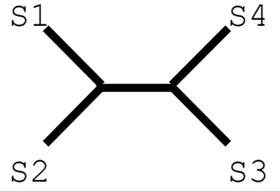
```
S1 = -AGGCTATCACCTGACCTCCA  
S2 = TAG-CTATCAC--GACCGC--  
S3 = TAG-CT-----GACCGC--  
S4 = -----TCAC--GACCGACA
```



A phylogenetic tree diagram showing the relationships between four sequences (S1, S2, S3, S4). S1 and S2 are sister taxa, and S3 and S4 are sister taxa. These two pairs are then joined together at a higher level. The labels S1, S2, S3, and S4 are placed at the tips of the branches.

True tree and
alignment

```
S1 = -AGGCTATCACCTGACCTCCA  
S2 = TAG-CTATCAC--GACCGC--  
S3 = TAG-C--T-----GACCGC--  
S4 = T---C-A-CGACCGA-----CA
```



A phylogenetic tree diagram showing the estimated relationships between four sequences (S1, S2, S3, S4). S1 and S4 are sister taxa, and S2 and S3 are sister taxa. These two pairs are then joined together at a higher level. The labels S1, S2, S3, and S4 are placed at the tips of the branches.

Estimated tree and
alignment

Compare

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)

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