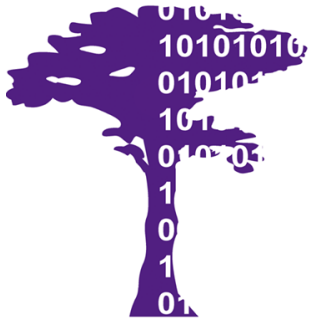
Simultaneous estimation of Trees and Alignments, or *Complexity and The Tree of Life*

Tandy Warnow

The University of Texas at Austin



CIPRES

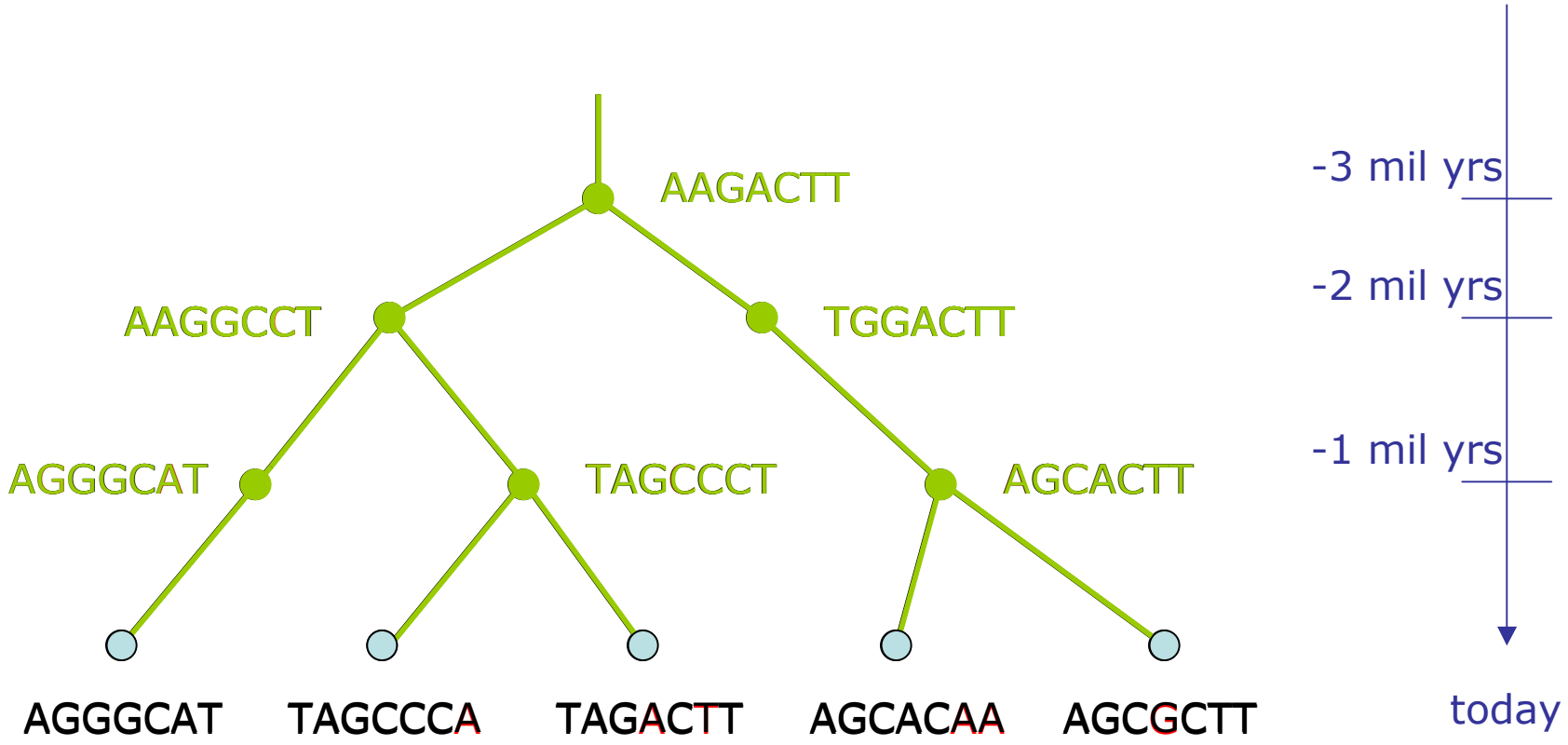
How did life evolve on earth?

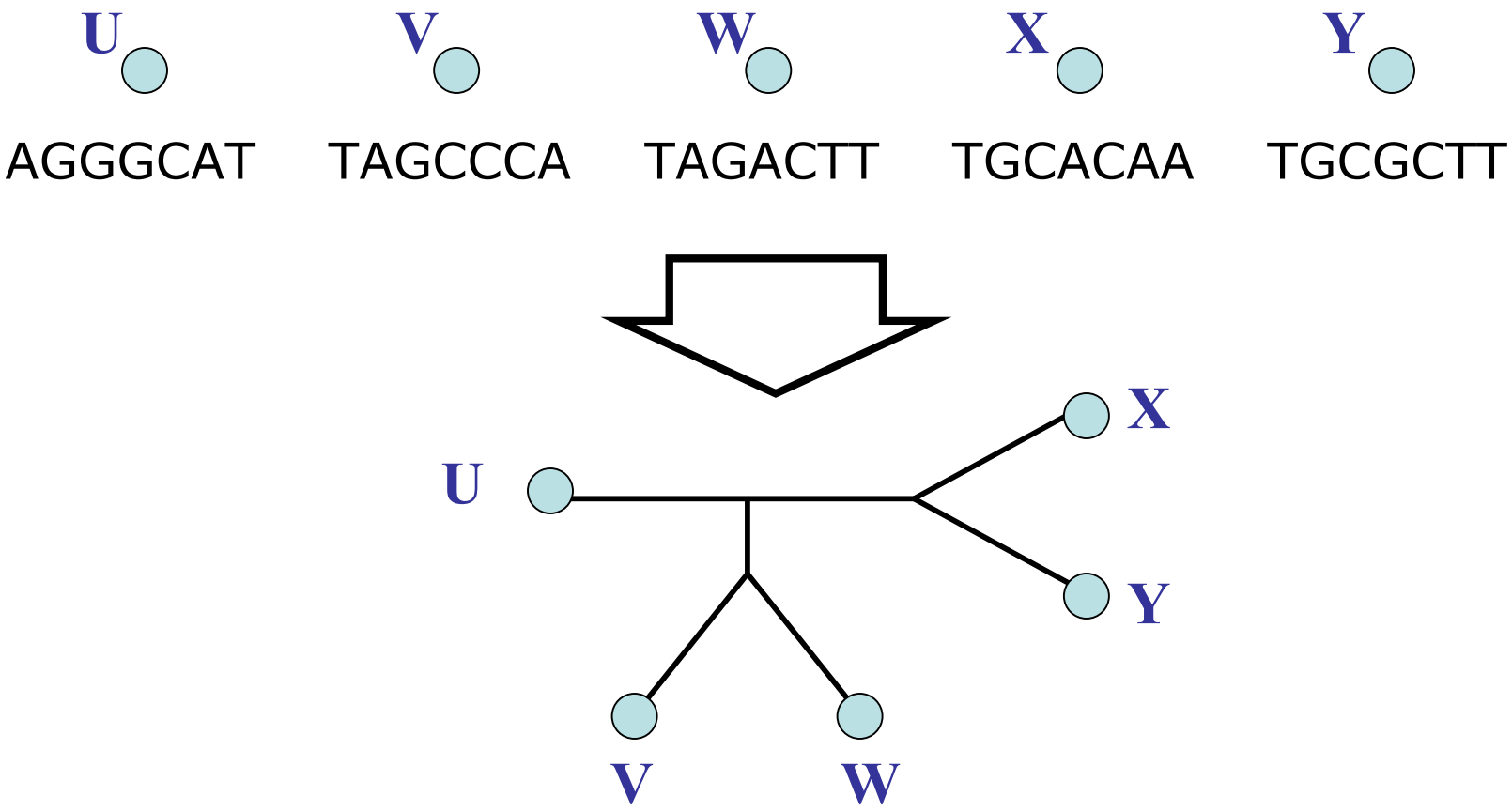
- Courtesy of the Tree of Life project

An international effort to understand how life evolved on earth

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

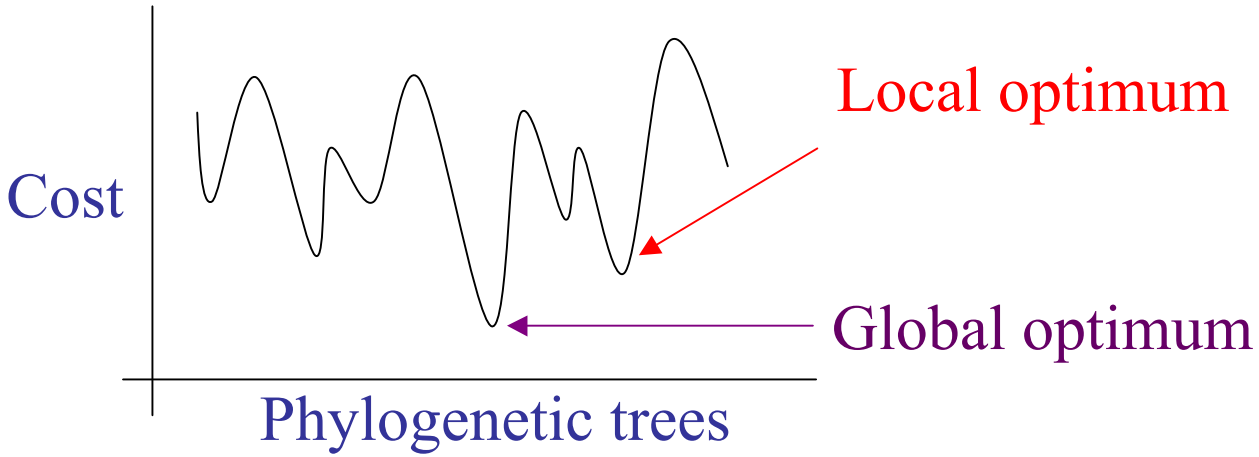
DNA Sequence Evolution




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

Phylogenetic reconstruction methods

1. Hill-climbing heuristics for hard optimization criteria (Maximum Parsimony and Maximum Likelihood)



2. Polynomial time distance-based methods, e.g. Neighbor Joining, FastME, UPGMA, etc.
3. Bayesian MCMC methods

Solving NP-hard problems exactly is … unlikely

- Number of (unrooted) binary trees on *n* leaves is (2n-5)!!
- If each tree on **1000** taxa could be analyzed in **0.001** seconds, we would find the best tree in **2890 millennia**

#leaves	#trees
4	3
5	15
6	105
7	945
8	10395
9	135135
10	2027025
20	2.2×10^{20}
100	4.5×10^{190}
1000	2.7×10^{2900}

Standard Markov models

- Sequences evolve just with substitutions
- Sites (i.e., positions) evolve *i.i.d.* (identically and independently), and have "rates of evolution" that are drawn from a common distribution (typically gamma)
- *Numerical parameters* describe the probability of substitutions of each type on each edge of the tree

Jukes-Cantor (simplest DNA model)

- DNA sequences (A,C,T,G) evolve just with substitutions
- Sites (i.e., positions) evolve *i.i.d.* (identically and independently)
- *If a site changes on an edge, it changes with equal probability to the remaining states (A,C,G,T)*
- *Numerical parameters p(e): for each edge in the tree, p(e) denotes the probability that each site changes on e*

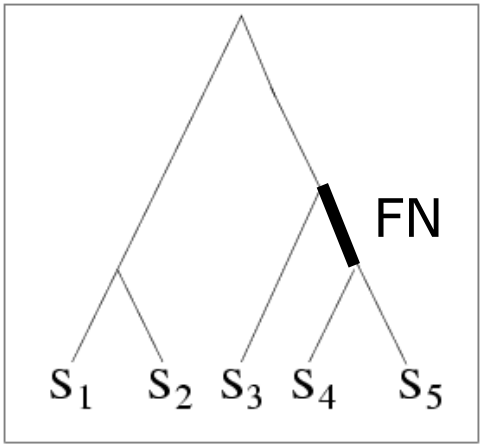
A Jukes-Cantor model tree is a pair (T,θ), where T is a tree and θ denotes the numerical parameters p(e), one for each edge e in the tree T.

Note that the JC model is *time-reversible*, so that without an assumption of the *molecular clock*, the root cannot be identified, even given infinite data

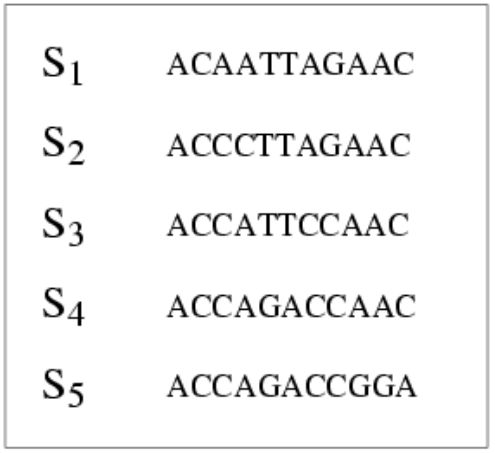
Questions

- *Statistical consistency*: Is the given phylogeny reconstruction method guaranteed to reconstruct the model tree when infinitely long sequences are available?
- *Convergence rate* (sample size complexity): How long do the sequences need to be for the method to be accurate with high probability?

Quantifying Error



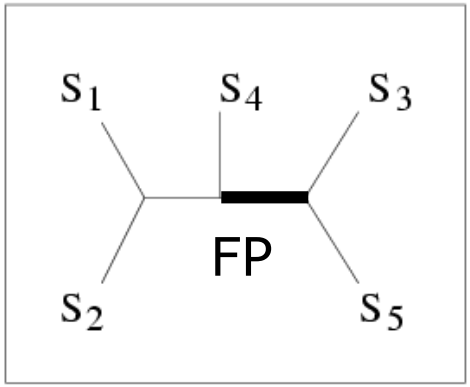
TRUE TREE



DNA SEQUENCES

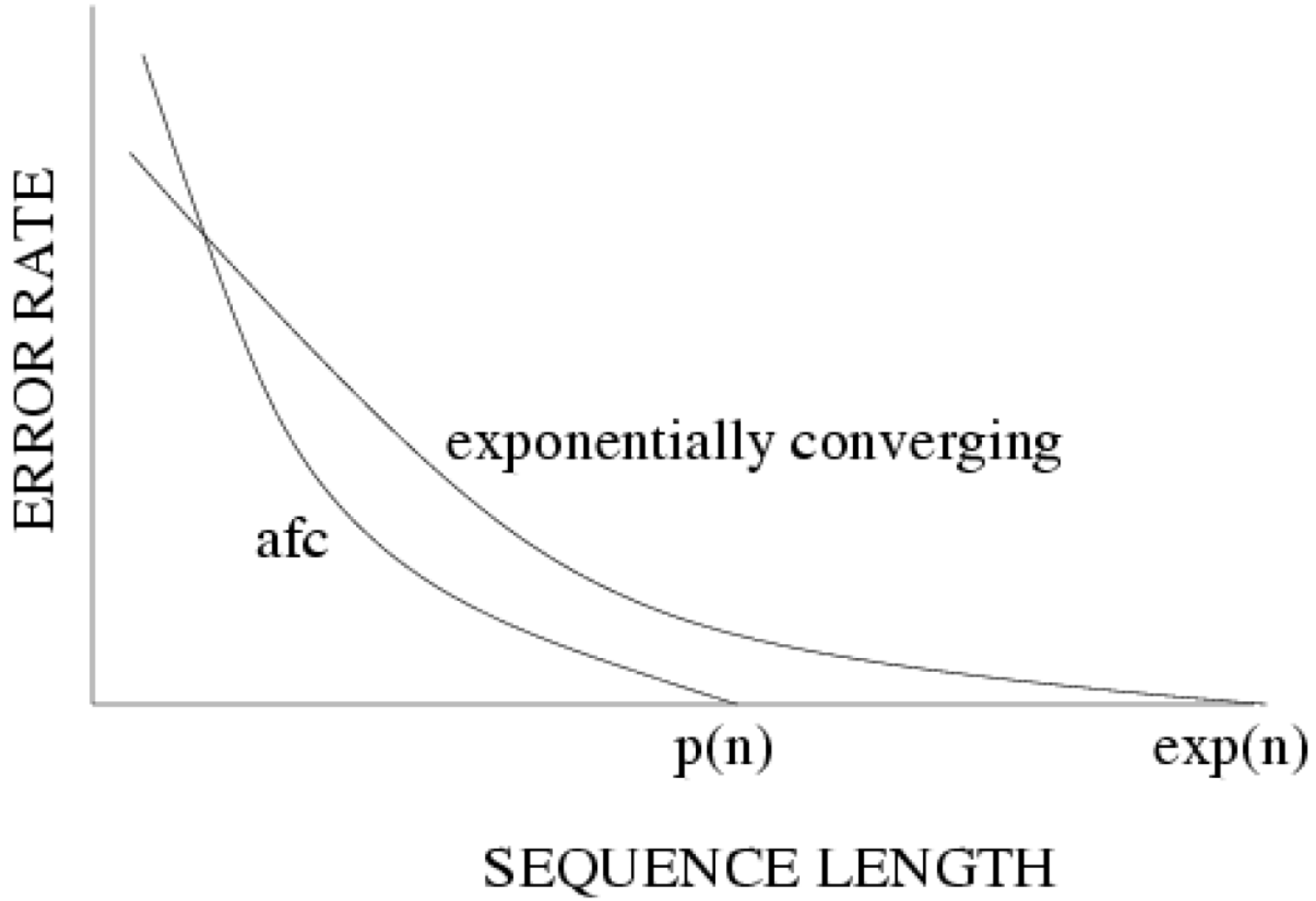
FN: false negative
(missing edge)
FP: false positive
(incorrect edge)

50% error rate



INFERRED TREE

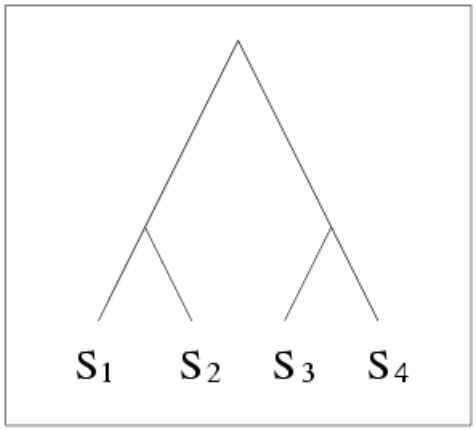
Statistical consistency, exponential convergence, and absolute fast convergence (afc)



Current state of knowledge

- We have established much of the statistical performance (consistency and convergence rates) of the major methods for phylogeny estimation.
- We have developed "fast converging" methods (guaranteed to reconstruct the true tree from polynomial length sequences) with excellent performance in practice.
- We have very fast methods for solving ***maximum likelihood*** and maximum parsimony, the major optimization problems, even for large datasets.

Distance-based Phylogenetic Methods (polynomial time)



TRUE TREE

S_1	ACAATTAGAAC
S_2	ACCCTTAGAAC
S_3	ACCATTCCAAC
S_4	ACCAGACCAAC

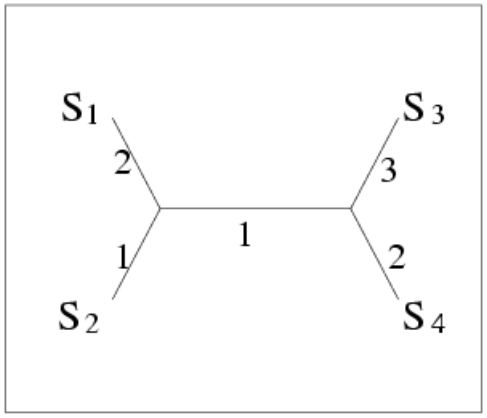
DNA SEQUENCES

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

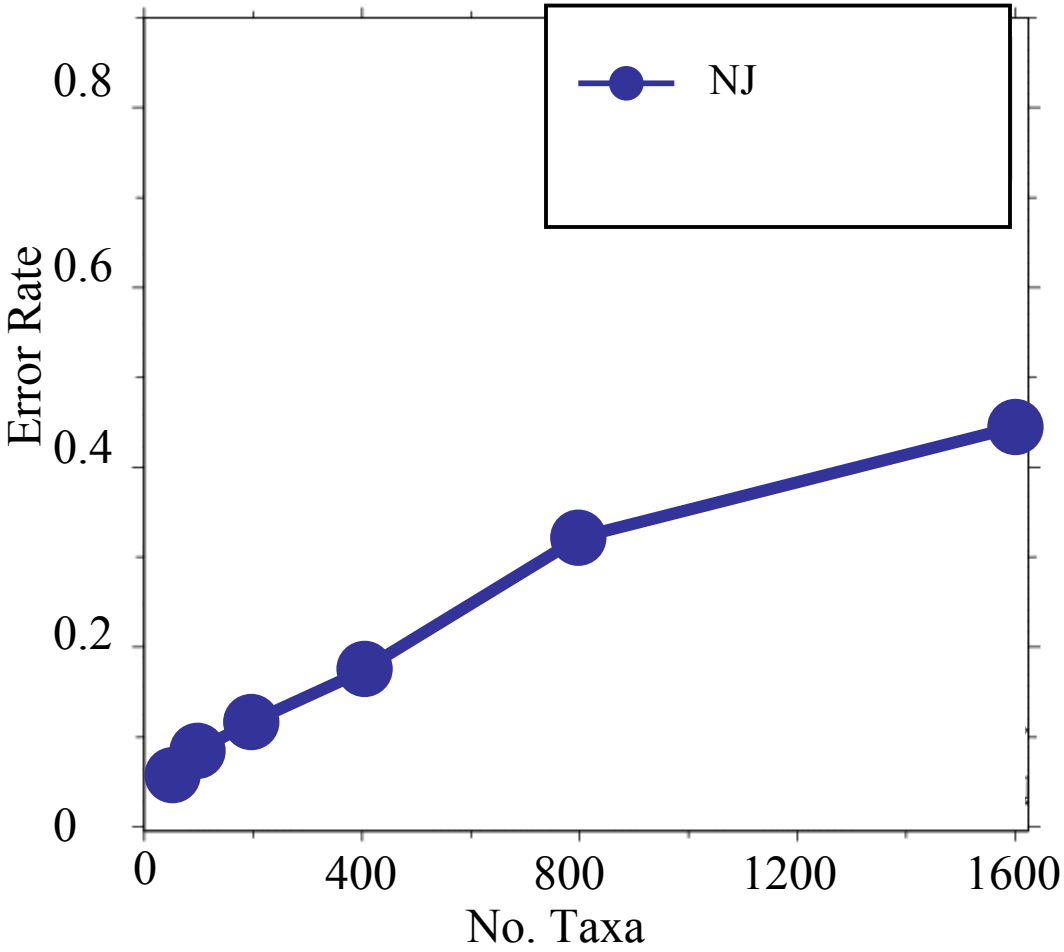


INFERRED TREE

Neighbor Joining's sequence length requirement is exponential!

- Atteson: Let T be a General Markov model tree defining distance matrix D. Then Neighbor Joining will reconstruct the true tree with high probability from sequences that are of length at least **$O(\lg n \, e^{\max Dij})$,** where n is the number of leaves in T.

Neighbor joining has poor performance on large diameter trees *[Nakhleh et al. ISMB 2001]*

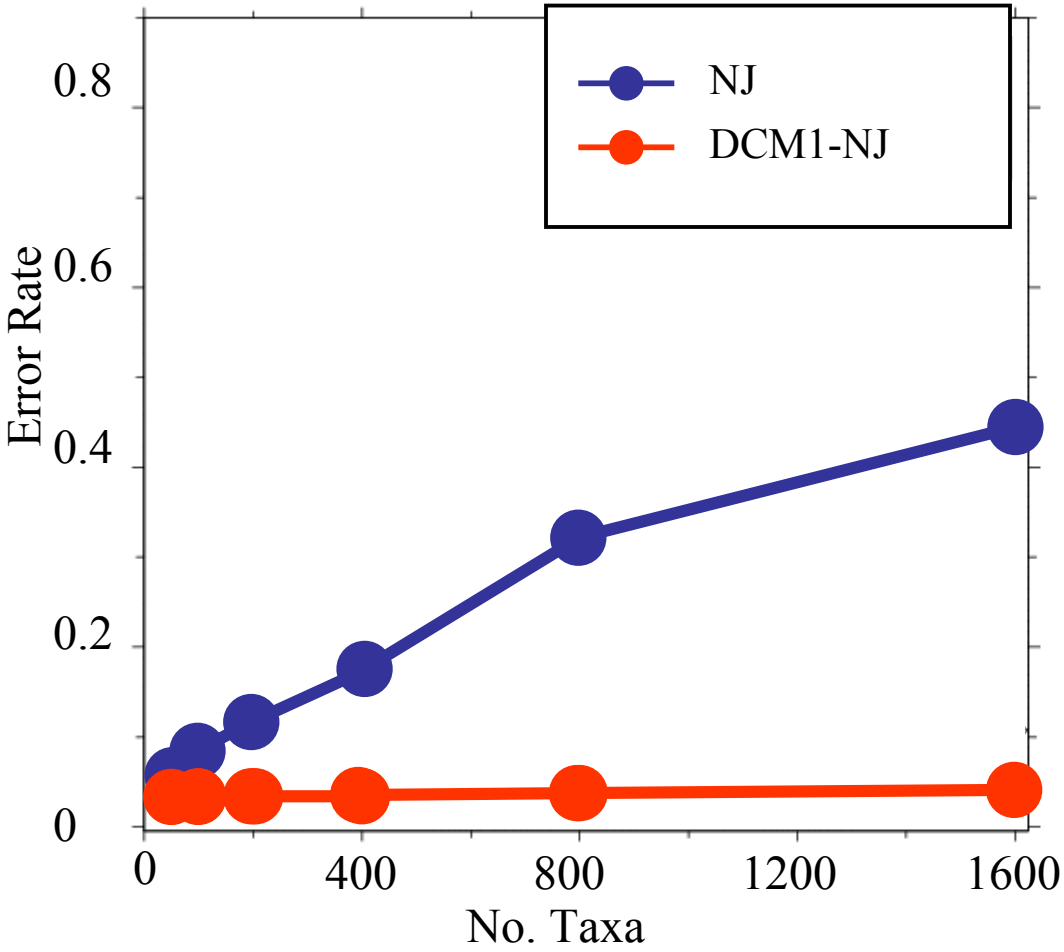


Simulation study based upon fixed edge lengths, K2P model of evolution, sequence lengths fixed to 1000 nucleotides.

Error rates reflect proportion of incorrect edges in inferred trees.

DCM1-boosting distance-based methods

[Nakhleh et al. ISMB 2001]



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

Maximum Likelihood (ML)

- Given: Set S of aligned DNA sequences, and a parametric model of sequence evolution (e.g., Jukes-Cantor)
- Objective: Find model tree (T,θ) to maximize $\Pr[S|T,\theta]$.

Maximum Likelihood (ML)

- Given: Set S of aligned DNA sequences, and a parametric model of sequence evolution (e.g., Jukes-Cantor)
- Objective: Find model tree (T,θ) to maximize $Pr[S|T,\theta]$.

Statistically consistent, but not known to be afc (best upper bound on sequence length requirement is exponential)

NP-hard

Excellent heuristics exist (e.g., RAxML) that *produce highly accurate trees*

Is phylogeny estimation *Solved*?

Is phylogeny estimation *Solved*?

- Much mathematical theory about convergence rates for phylogeny estimation methods

Is phylogeny estimation *Solved*?

- Much mathematical theory about convergence rates for phylogeny estimation methods
- Fast-converging polynomial time distance-based methods with excellent performance in simulation (DCM1-NJ and others).

Is phylogeny estimation *Solved*?

- Much mathematical theory about convergence rates for phylogeny estimation methods
- Fast-converging polynomial time distance-based methods with excellent performance in simulation (DCM1-NJ and others).
- Maximum likelihood: statistically consistent, with excellent heuristics, producing highly accurate trees (established using simulations) on large datasets

No, because standard Markov models *are too simple!*

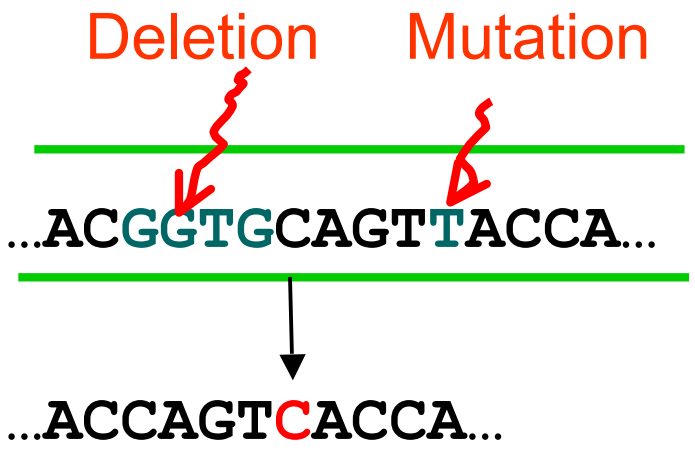
Simplifying assumptions:

- Sequences evolve just with substitutions
- Sites (i.e., positions) evolve identically and independently, and have "rates of evolution" that are drawn from a common distribution (typically gamma)

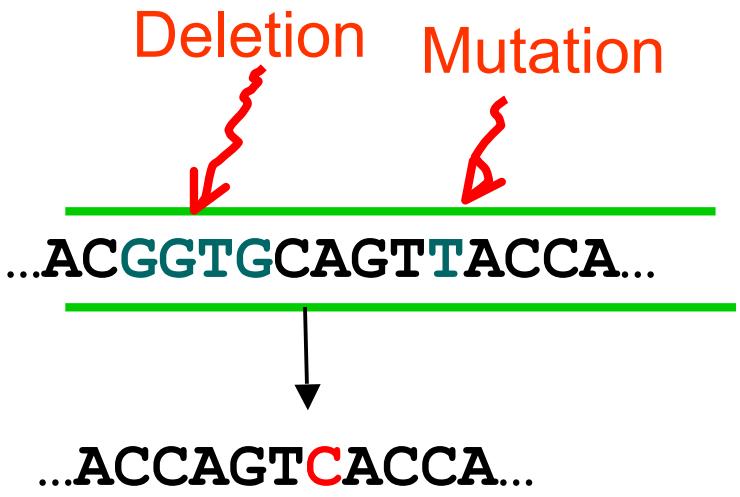
No, because standard Markov models *are too simple!*

Simplifying assumptions:

- Sequences evolve just with substitutions
- Sites (i.e., positions) evolve identically and independently, and have "rates of evolution" that are drawn from a common distribution (typically gamma)



indels (insertions and deletions) also occur!



The true pairwise alignment is:

...ACGGTGCAGTTACCA...

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

The **true multiple alignment** on a set of homologous sequences is obtained by tracing their evolutionary history, and extending the pairwise alignments on the edges to a multiple alignment on the leaf sequences.

Input: unaligned sequences

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

Phase 1: Multiple Sequence Alignment

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



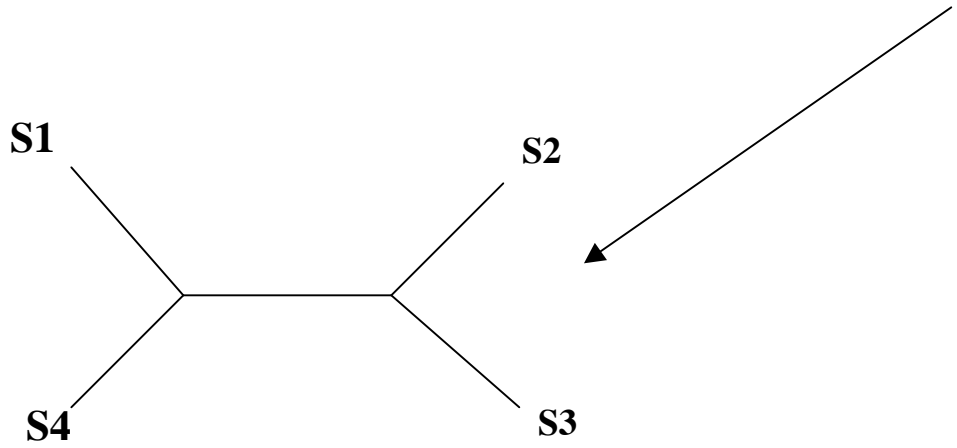
```
S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAC--GACCGC--
S3 = TAG-CT-------GACCGC--
S4 = -------TCAC--GACCGACA
```

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

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

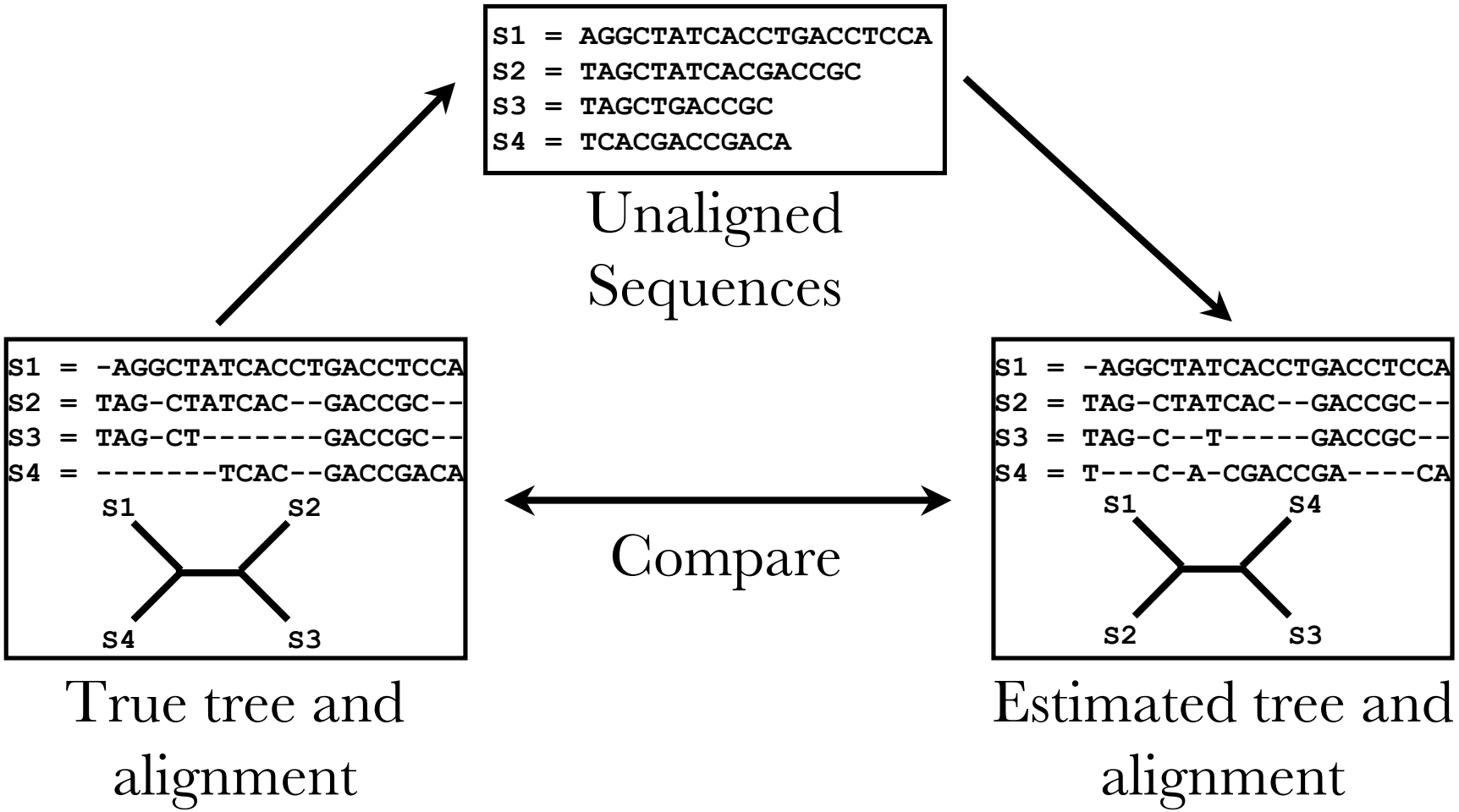
***RAxML**: best heuristic for large-scale ML optimization*

Question: How well do two-phase methods perform?

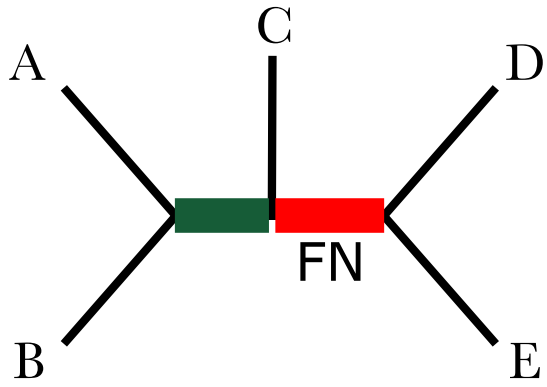
- ROSE simulation:
  - 1000, 500, and 100 sequences
  - Evolution with substitutions and indels
  - Varied gap lengths, rates of evolution
- Estimated alignments using leading methods
- Used RAxML to compute trees
- Recorded tree error (missing branch rate)
- Recorded alignment error

Liu et al., *Science* 2009

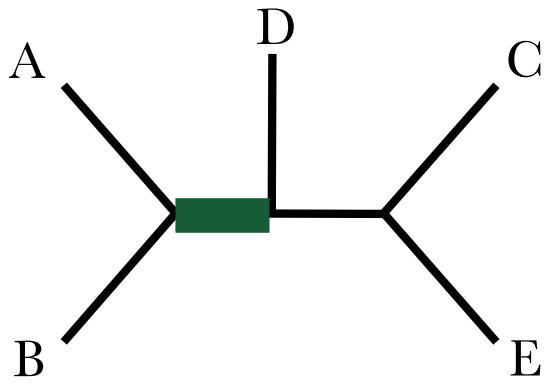
Simulation Studies



Quantifying Error

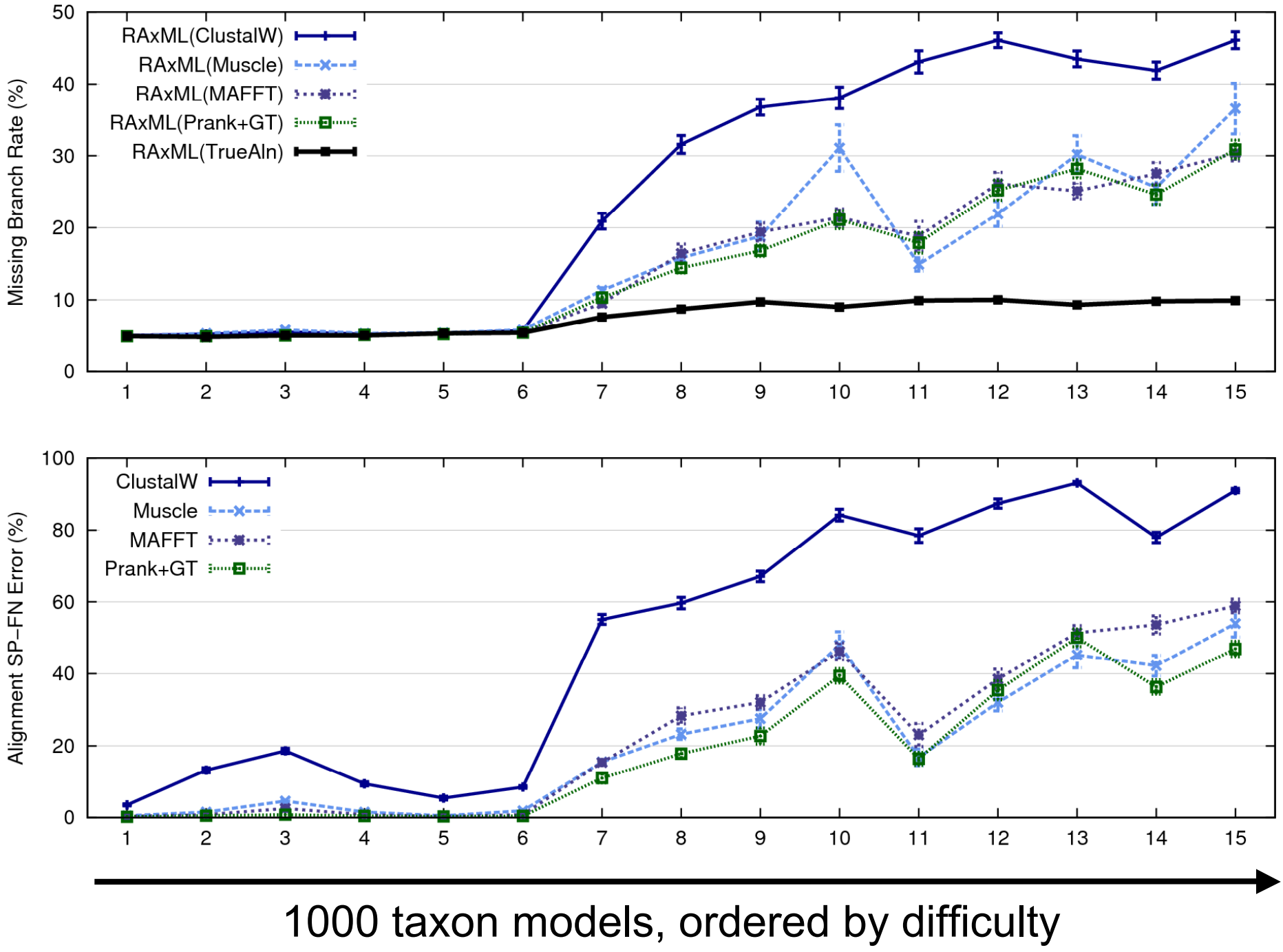


True Tree



Estimated Tree

False negative (FN) - aka “missing branch” :
An edge in the true tree missing from the estimated tree

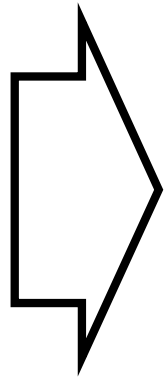

RAxML(ClustalW)
RAxML(Muscle)
RAxML(MAFFT)
RAxML(Prank+GT)
RAxML(TrueAln)
Missing Branch Rate (%)
0
10
20
30
40
50
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
ClustalW
Muscle
MAFFT
Prank+GT
Alignment SP–FN Error (%)
0
20
40
60
80
100
1000 taxon models, ordered by difficulty

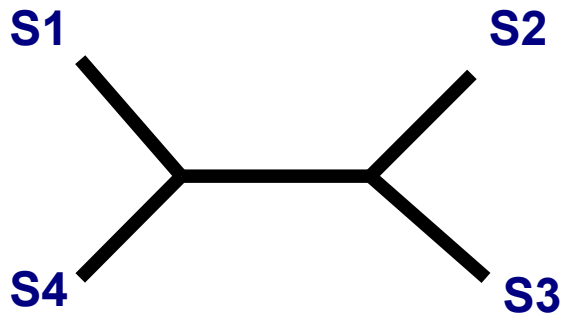
Problems with the two-phase approach

- 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)

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





and

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

Statistical simultaneous estimation methods (BALiPhy, Alifritz, Statalign) are not scalable.

POY and related methods are not more accurate than standard two-phase methods.

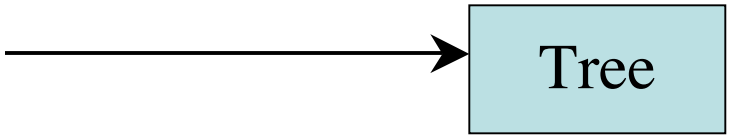
SATé:
(Simultaneous Alignment and Tree Estimation)

- 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, attempting to optimize the **GTR+Gamma maximum likelihood score.**
- Software at http://phylo.bio.ku.edu/software/sate/sate.html

- *Science*, 19 June 2009, pp. 1561-1564.

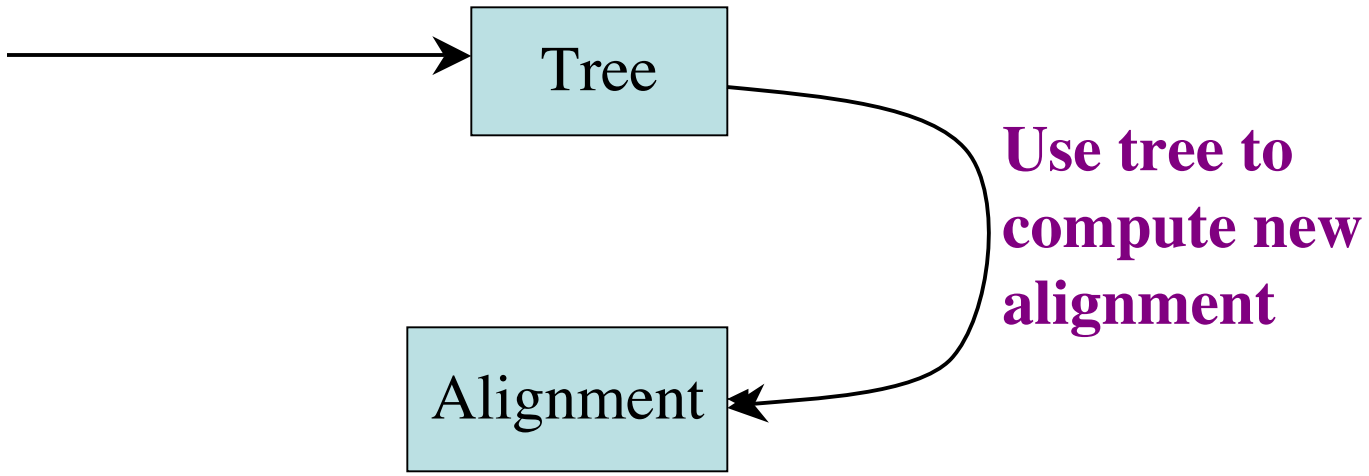
SATé Algorithm

Obtain initial alignment
and estimated ML tree



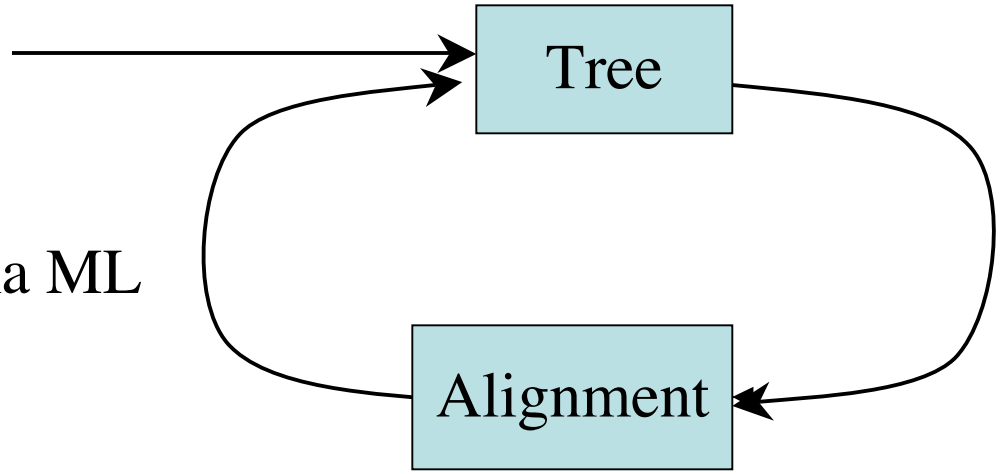
SATé Algorithm

Obtain initial alignment
and estimated ML tree



SATé Algorithm

Obtain initial alignment
and estimated ML tree

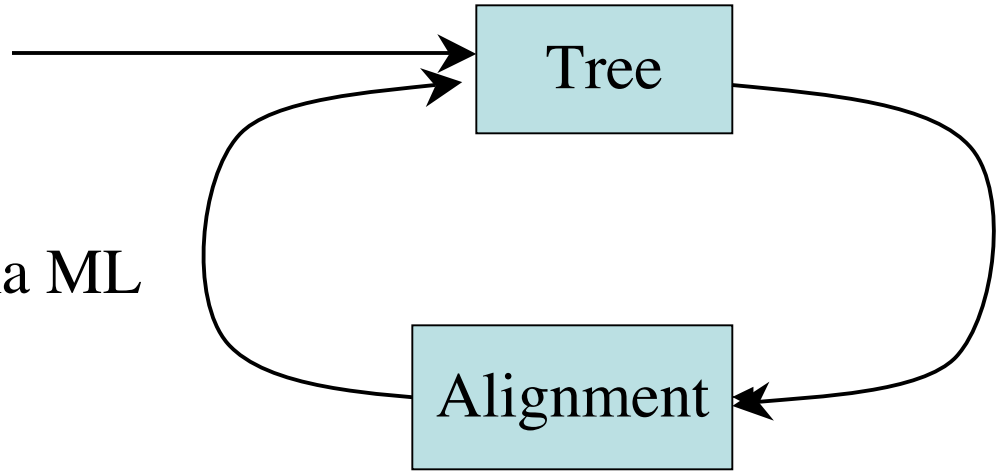


Estimate
GTR+Gamma ML
tree on new
alignment

Use tree to compute new alignment

SATé Algorithm

Obtain initial alignment
and estimated ML tree



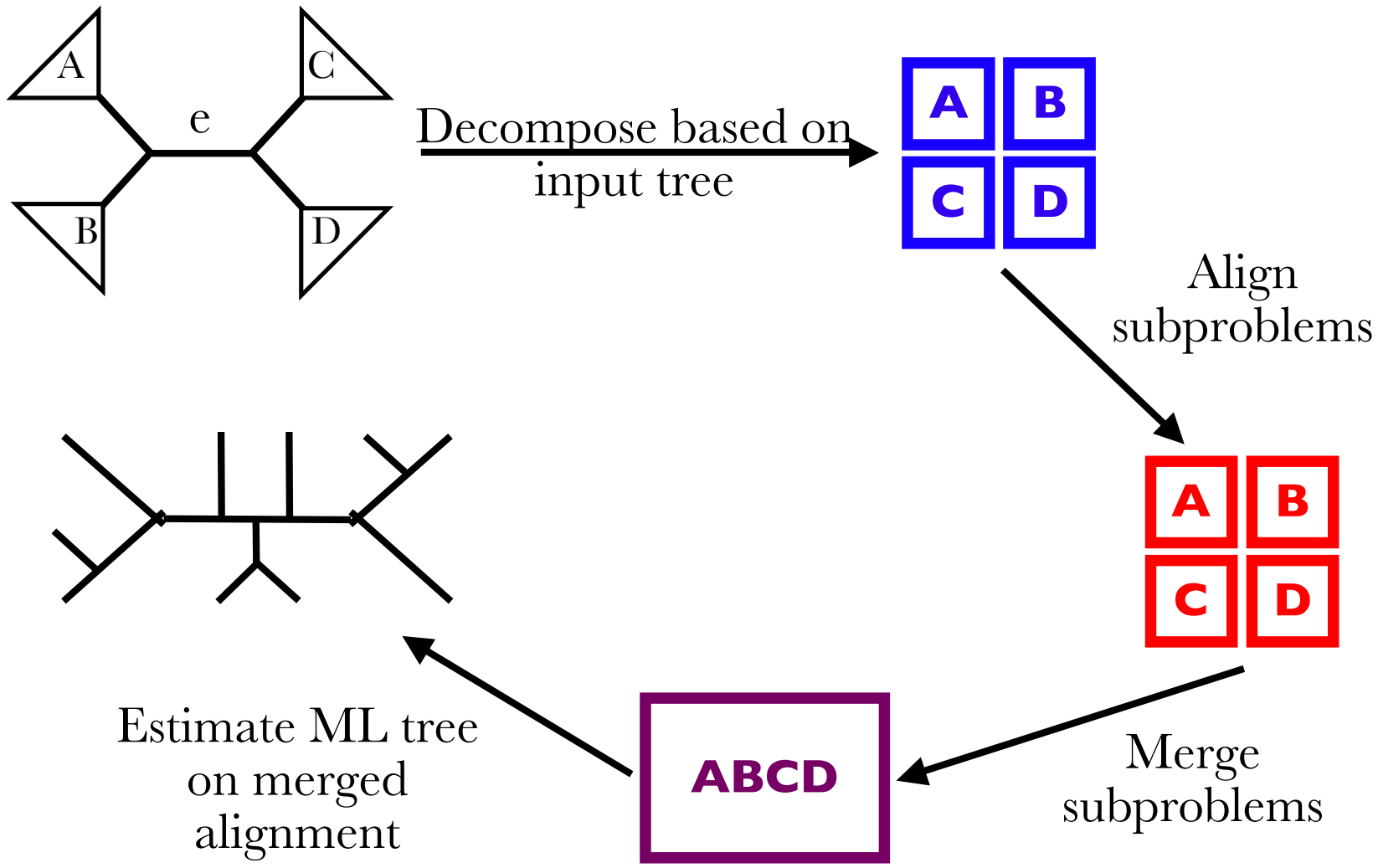
Estimate
GTR+Gamma ML
tree on new
alignment

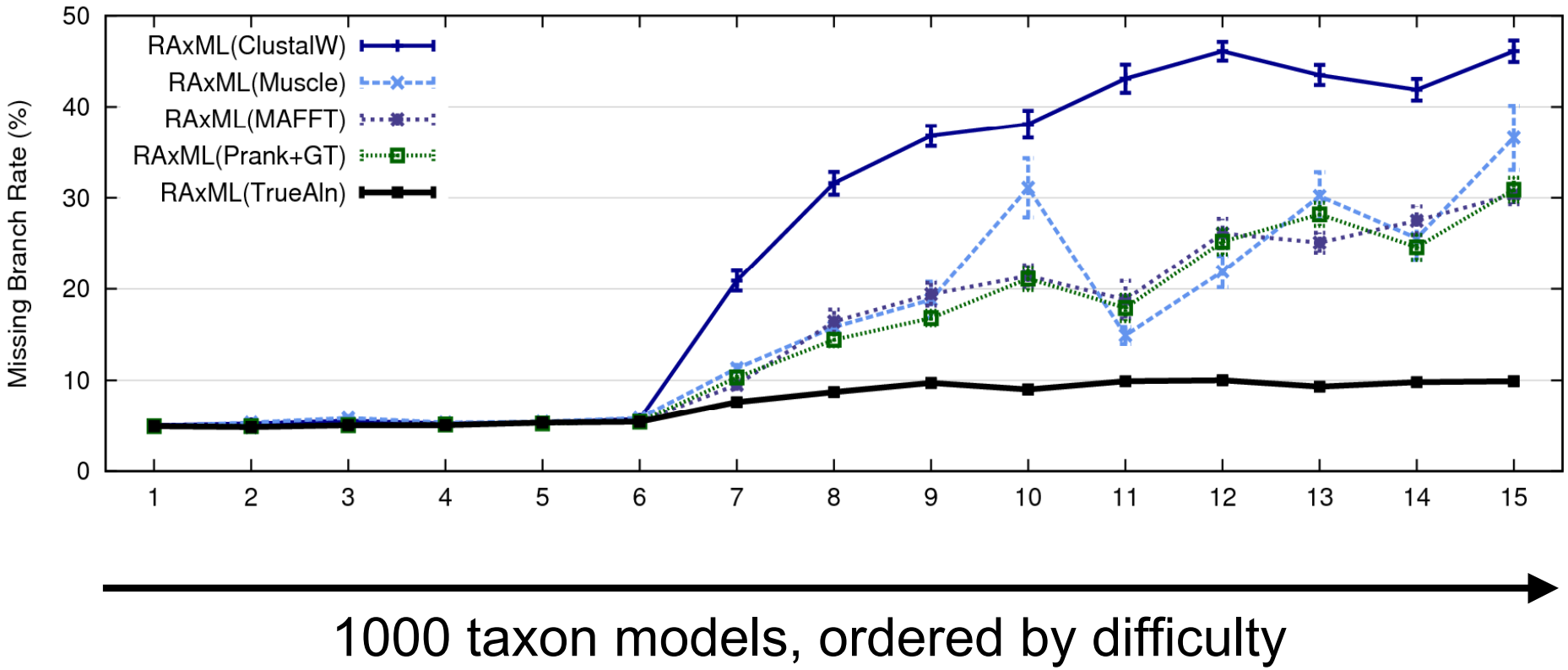
Use tree to compute new alignment

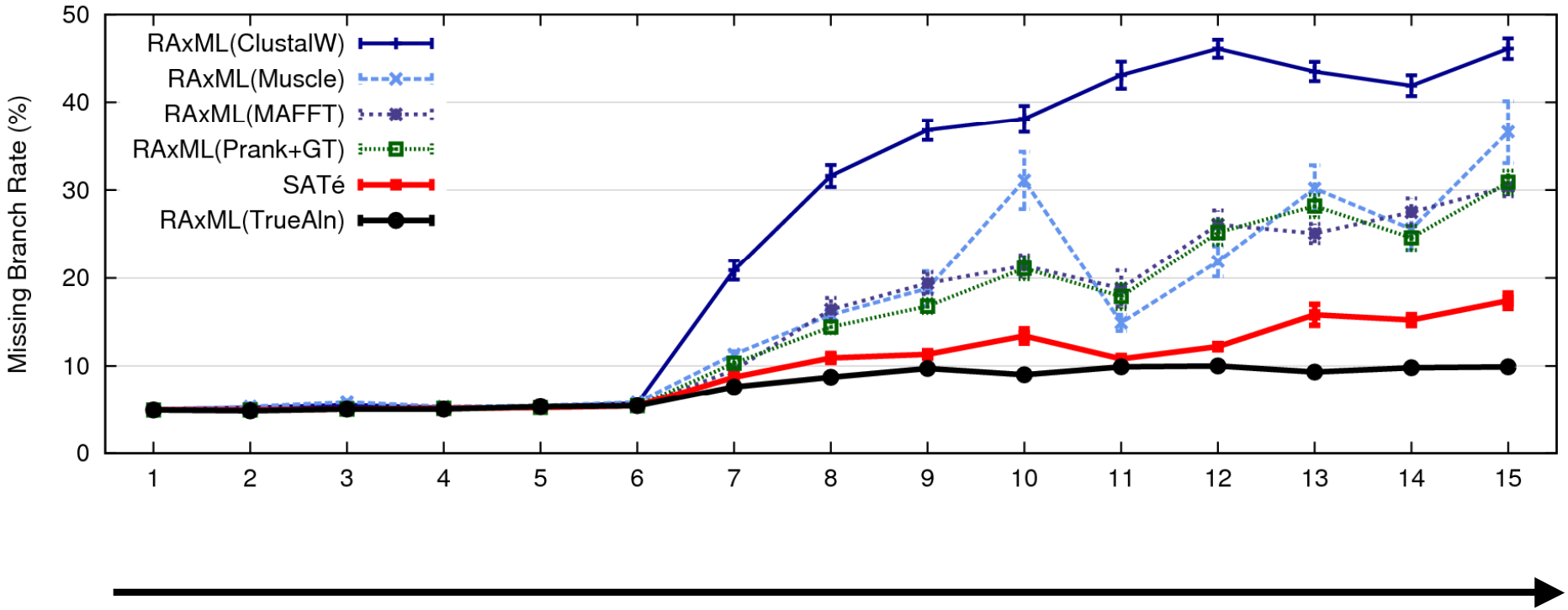
If new alignment/tree pair has worse GTR+Gamma ML score, realign using a different decomposition

Repeat until termination condition (typically, 24 hours)

SATé iteration
(Actual decomposition produces 32 subproblems)




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)

Why is SATé so accurate?

- *Not because it's good to optimize ML* under a model treating gaps as missing data - we **prove** that this optimization problem is a *bad* approach.

Extended Jukes-Cantor model

- Sites evolve *i.i.d*
- The state at the root is random
- Substitution probabilities p(e) on each edge e satisfy $0 \leq p(e) < 3/4$.
- If a substitution occurs on an edge e, the nucleotide changes to the remaining nucleotides with equal probability.

Negative result

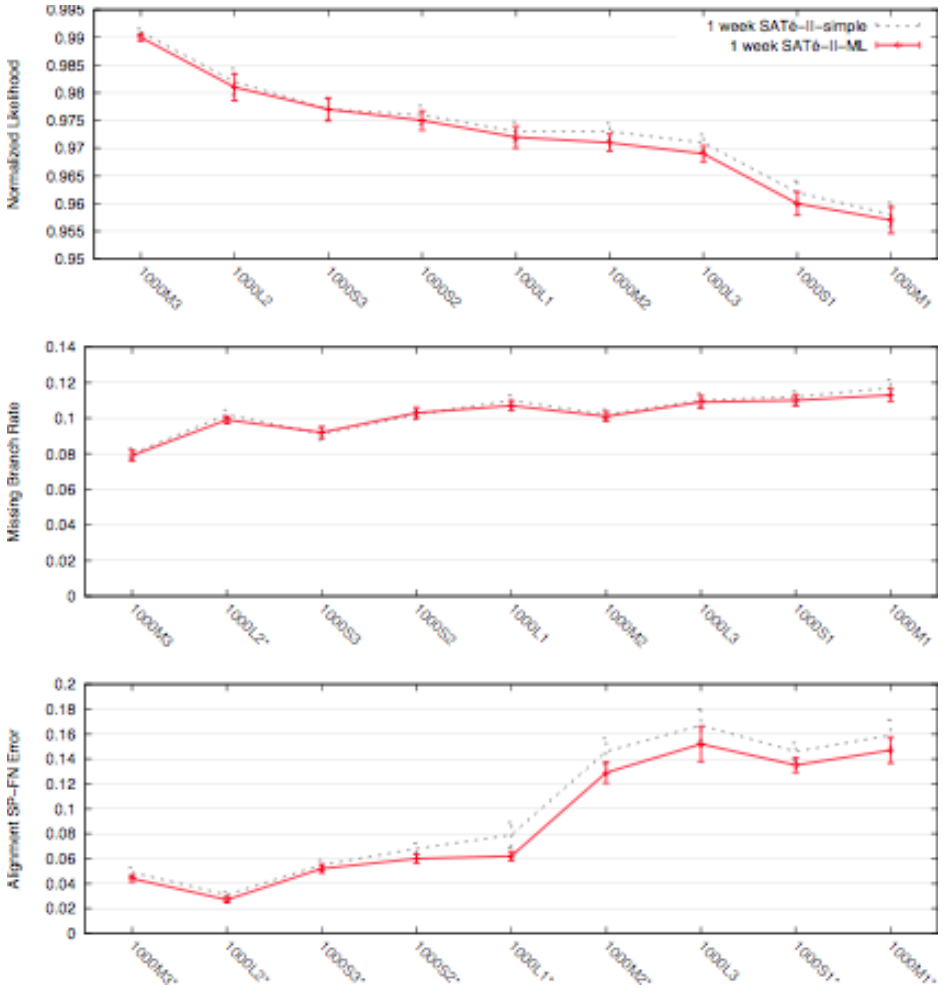
- Let S be a set of DNA sequences, and let Opt(S)= max Pr[A|T,θ], where A is an alignment on S and (T,θ) is an EJC model tree for S.
- Let BestTrees(S)={T: for some θ and some alignment A, Pr[A|T,θ]=Opt(S)}
- **Theorem: For all sets S of unaligned DNA sequences, BestTrees(S) contains *all* trees on S.**

Why is SATé so accurate?

- *Not because it's good to optimize ML* under GTR+Gamma

- Instead, the key is the *divide-and-conque*r technique used in the re-alignment strategy.

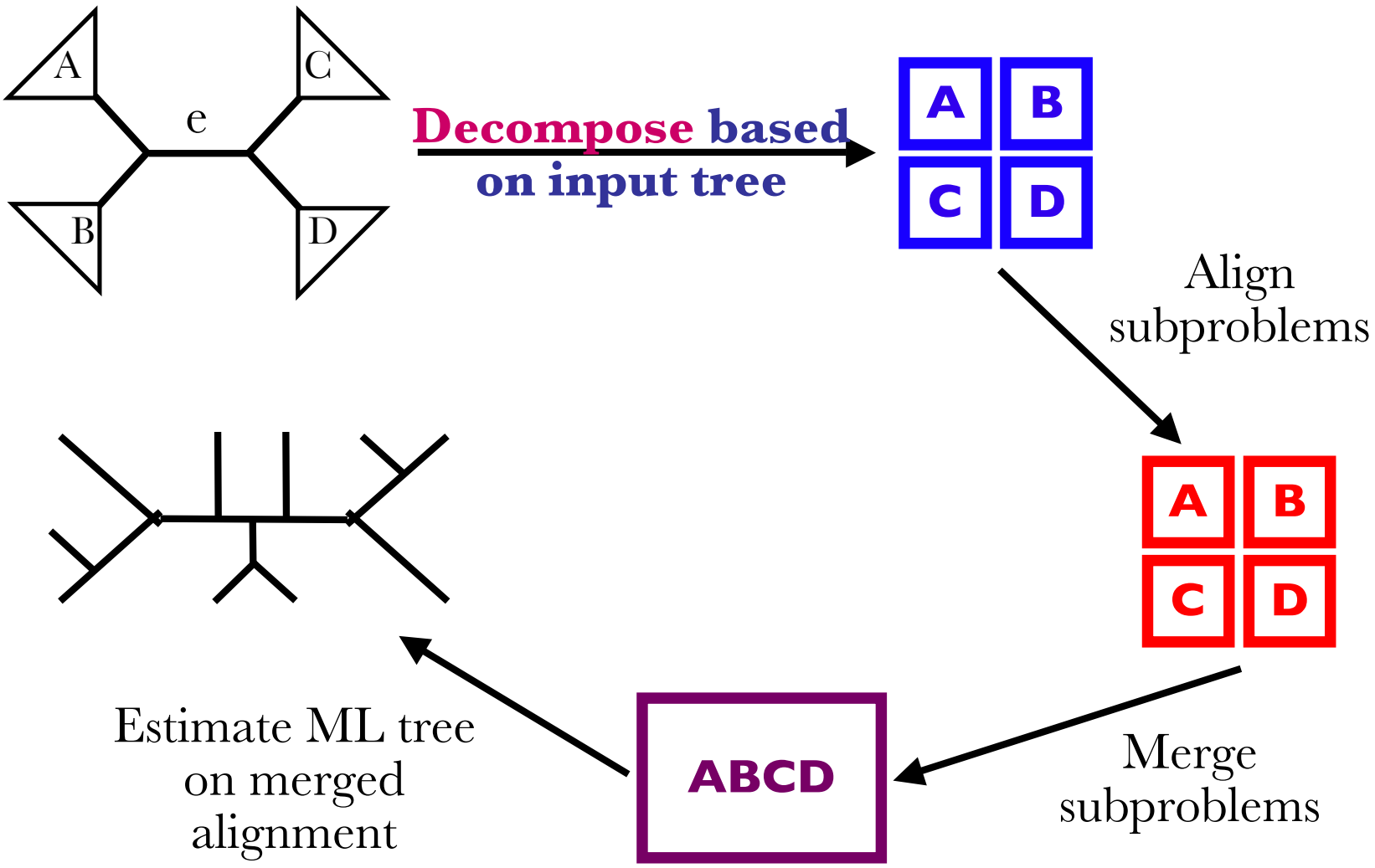
Does using ML help?

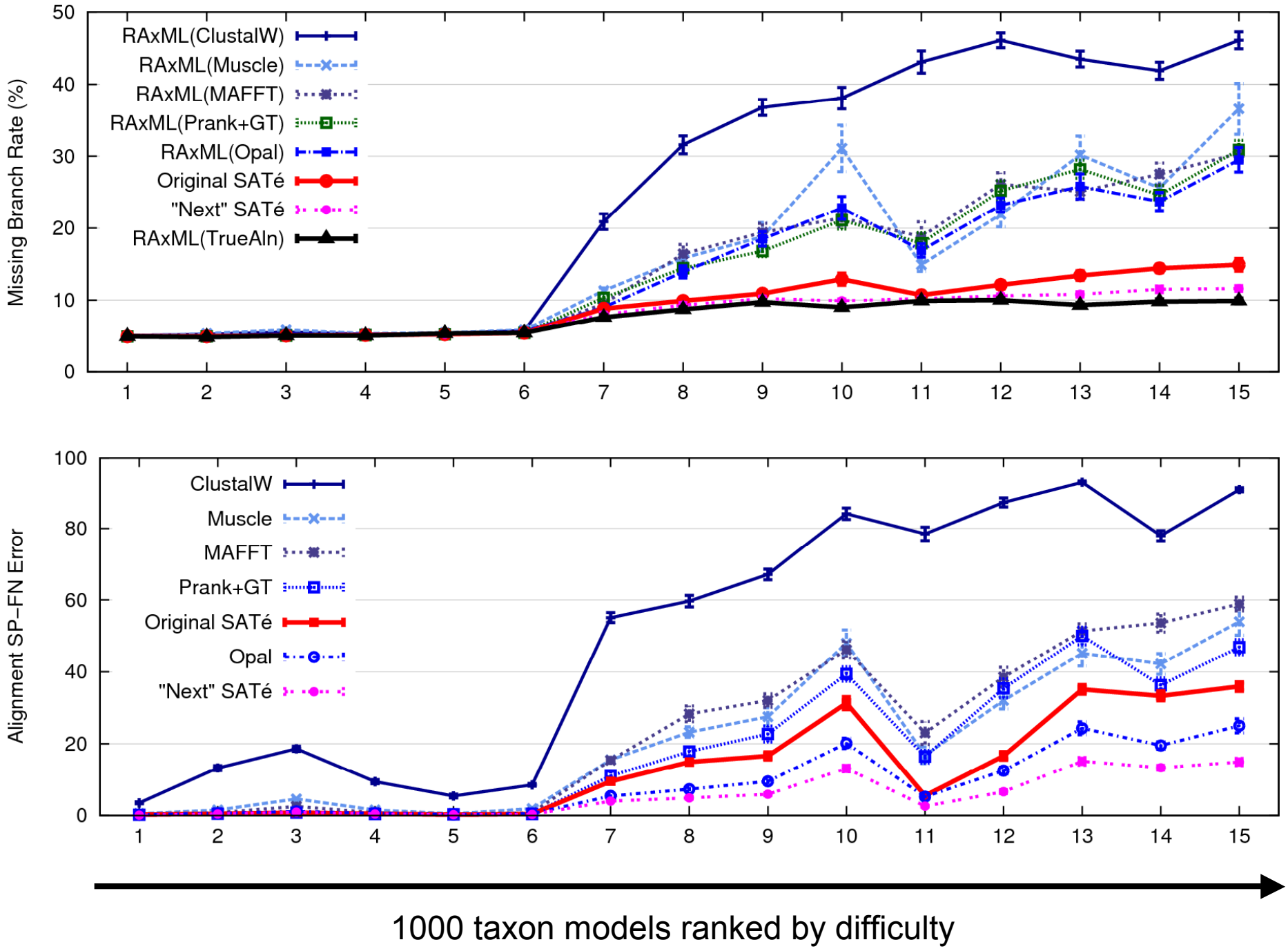


Since the *Science* paper: SATé-II:

- Uses a different re-alignment strategy, but same general algorithm design.
- More accurate than SATé and much faster!

SATé-II: same as SATé-I except for **decomposition**




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

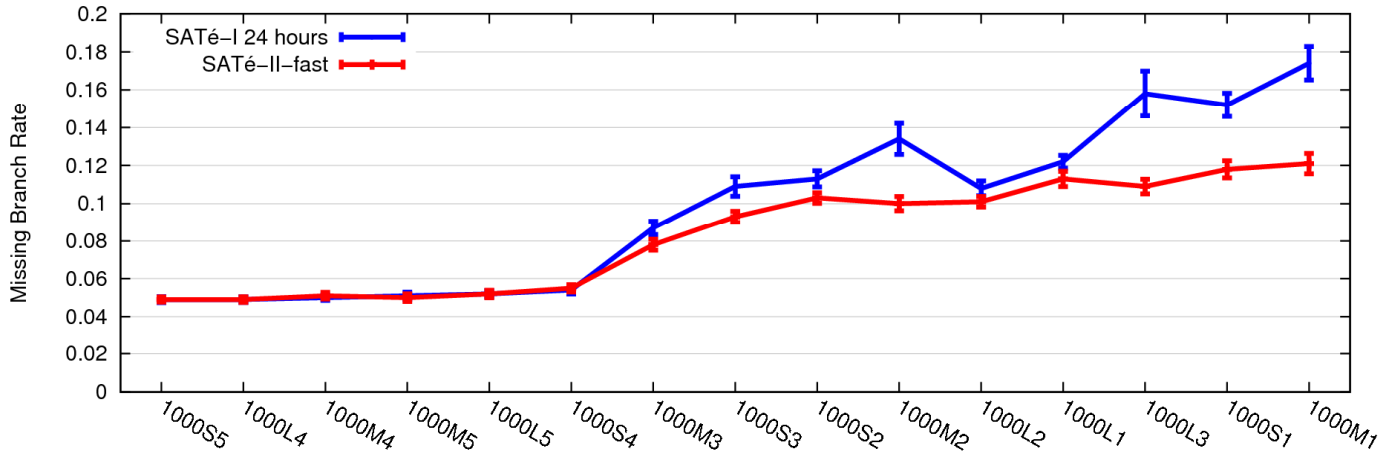
SATé-I

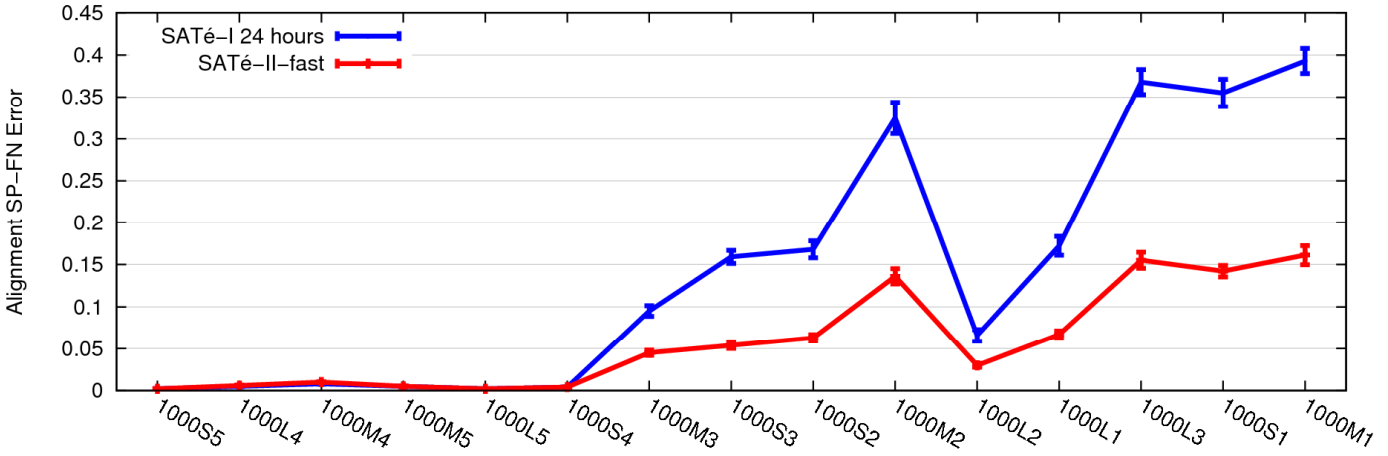
vs.

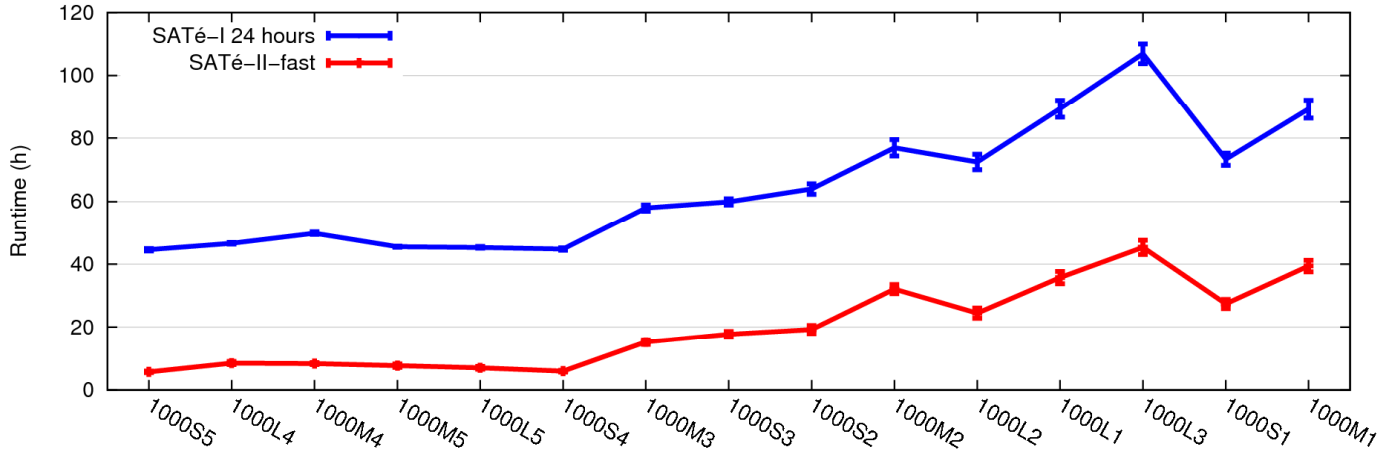
SATé-II

SATé-II

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







SATé Software

- Downloadable software (with user-friendly gui)
- Developers: Mark Holder and Jiaye Yu at the University of Kansas
- Webpage

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

Complexity *viz*. The Tree of Life

- **Algorithmic complexity** (e.g., running time and NP-hardness)
- **Sample size complexity** (e.g. how long do the sequences need to be to obtain a highly accurate reconstruction with high probability?)
- ***Stochastic model complexity*** *(i.e., how realistic are the models of evolution, and what are the consequences of making the models more realistic?)*

Take-home lessons

Take-home lessons

- Current models of evolution are simplistic

Take-home lessons

- Current models of evolution are simplistic
- More realistic models may not be "identifiable"

Take-home lessons

- Current models of evolution are simplistic
- More realistic models may not be "identifiable"
- Everything worth doing in phylogeny is NP-hard

Take-home lessons

- Current models of evolution are simplistic
- More realistic models may not be "identifiable"
- Everything worth doing in phylogeny is NP-hard
- But this doesn't mean we (mathematcians and computer scientists) can't make important contributions.

Acknowledgements

- Funding: NSF, The David and Lucile Packard Foundation, The Program in Evolutionary Dynamics at Harvard, and The Institute for Cellular and Molecular Biology at UT-Austin.
- Collaborators:
  - Fast-converging methods: Peter Erdös, Daniel Huson, Bernard Moret, Luay Nakhleh, Usman Roshan, Katherine St. John, Michael Steel, and Laszlo Székély
  - SATé: Randy Linder, Kevin Liu, Serita Nelesen, and Sindhu Raghavan

Thoughts

- Current models of sequence evolution are clearly too simple, and more realistic ones are not identifiable.
- The relative performance between methods can change as the models become more complex or as the number of taxa increases.
- We do not know how methods perform under realistic conditions (nor how long we need to let computationally intensive methods run).
- Therefore, *simulations should be done under very realistic (sufficiently complex) models, even if estimations are done under simpler models (and it is likely that estimations are best done under more realistic models, too)*.