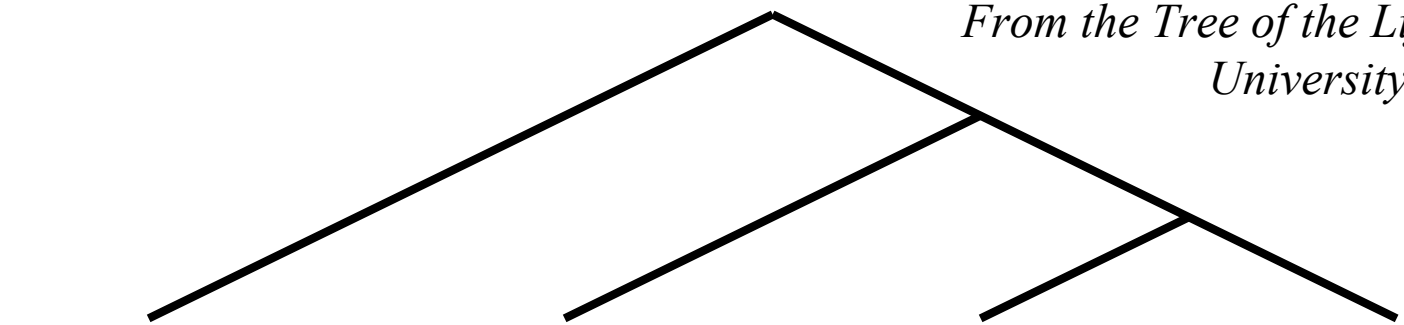
Approaching multiple sequence alignment from a phylogenetic perspective

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

Species phylogeny

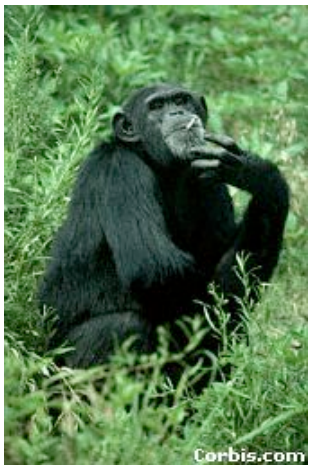
From the Tree of the Life Website,
University of Arizona



Orangutan

Gorilla

Chimpanzee



Human

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

Phylogenetic estimation is a "Grand Challenge": millions of taxa, NP-hard optimization problems

The CIPRES Project

(Cyber-Infrastructure for Phylogenetic Research)

www.phylo.org

This project is funded by the NSF under a Large ITR grant

- ***ALGORITHMS and SOFTWARE: scaling to millions of sequences (open source, freely distributed)***
- ***MATHEMATICS/PROBABILITY/STATISTICS: Obtaining better mathematical theory under complex models of evolution***
- ***DATABASES: Producing new database technology for structured data, to enable scientific discoveries***
- ***SIMULATIONS: The first million taxon simulation under realistically complex models***
- ***OUTREACH: Museum partners, K-12, general scientific public***
- ***PORTAL available to all researchers***

Step 1: Gather data

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

Step 2: Multiple Sequence Alignment

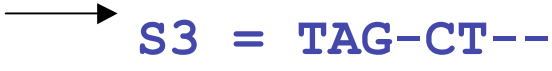
```
S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC
S3 = TAGCTGACCGC
S4 = TCACGACCGACA
```



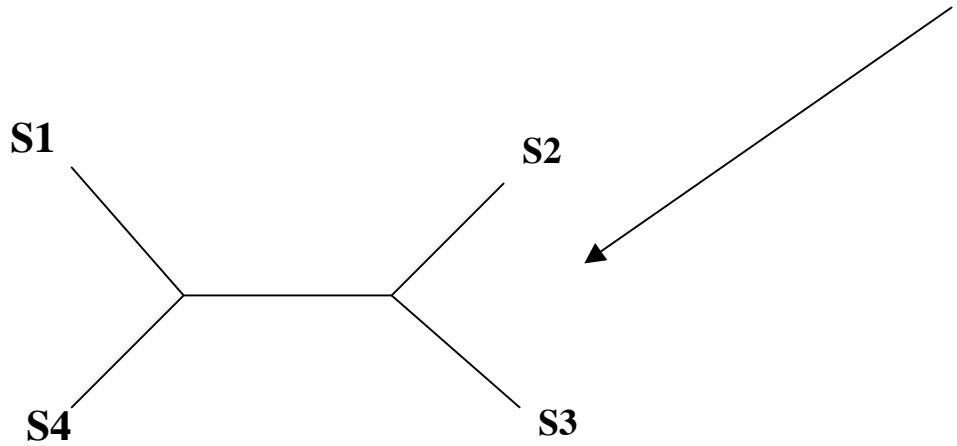
```
S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAC--GACCGC--
S3 = TAG-CT-------GACCGC--
S4 = -------TCAC--GACCGACA
```

Step 3: Construct tree

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

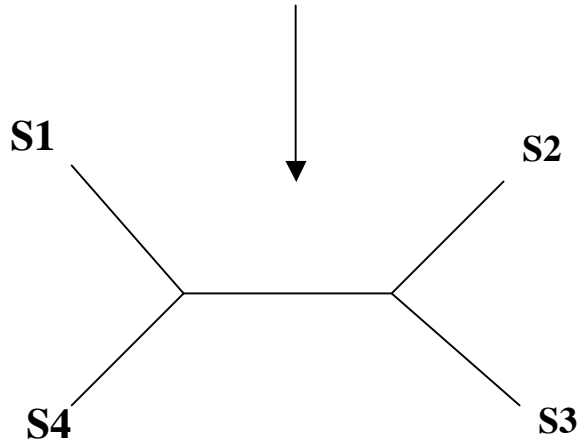


```
S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAC--GACCGC--
S3 = TAG-CT-------GACCGC--
S4 = -------TCAC--GACCGACA
```



But molecular phylogenetics assumes the alignment is given

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



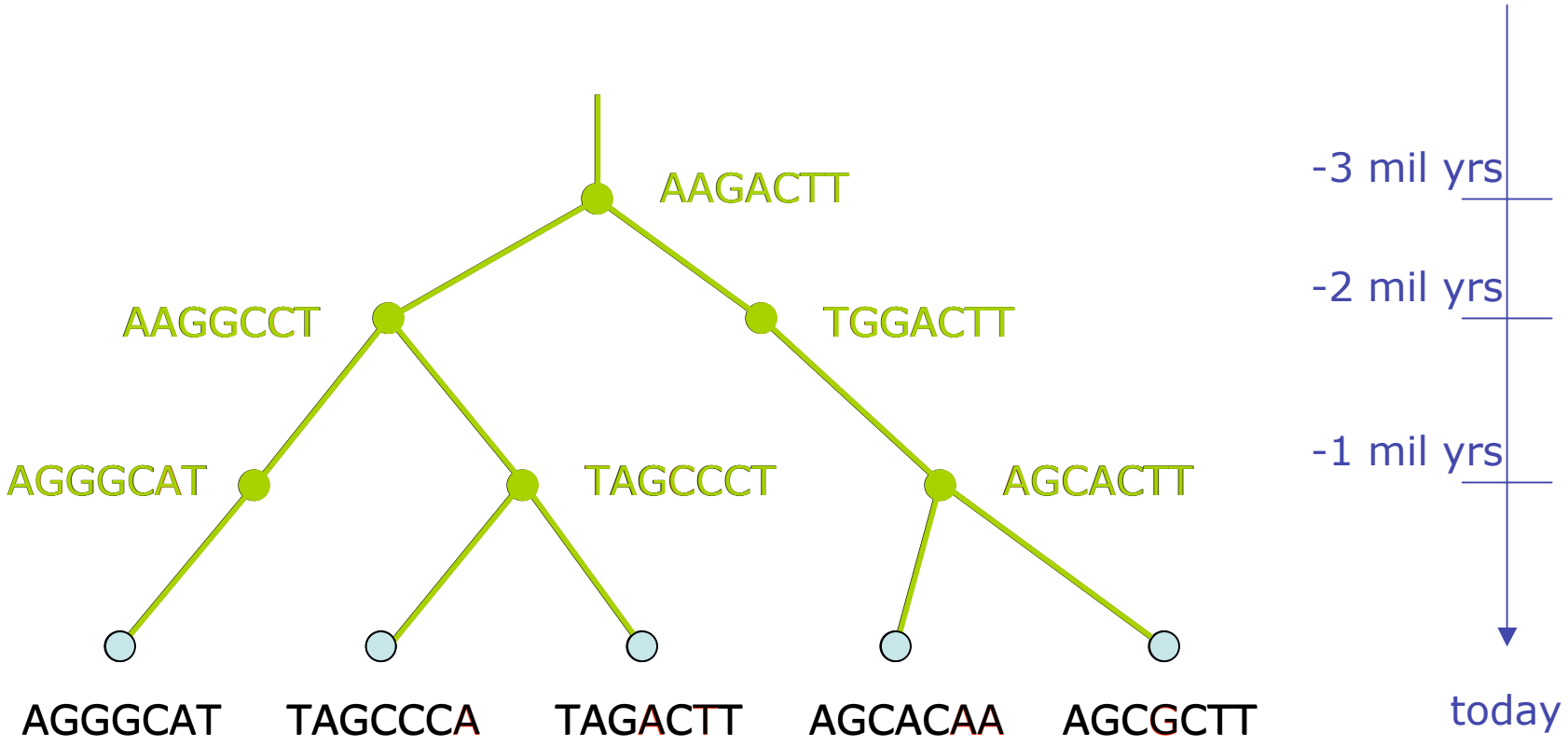
This talk

- **DCM-NJ**: Dramatic improvement in phylogeny estimation in terms of tree accuracy, and theoretical performance under Markov models of evolution
- **DCM-MP and DCM-ML**: Speeding up heuristics for large-scale phylogenetic estimation
- Simulation studies of two-phase methods (amino-acid and DNA sequences).
- **SATe**: A new technique for simultaneous estimation of trees and alignments

Performance criteria

- Estimated alignments are evaluated with respect to the ***true alignment***. Studied both in simulation and on real data.
- Estimated trees are evaluated for "topological accuracy" with respect to the ***true tree.*** Typically studied in simulation.
- Methods for these problems can also be evaluated with respect to an optimization criterion (e.g., maximum likelihood score) as a function of running time. Typically studied on real data. (Reasonably valid for phylogeny but not yet for alignment.)

DNA Sequence Evolution

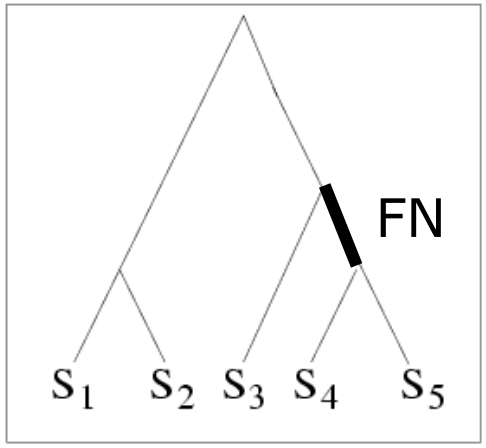


Markov models of single site evolution

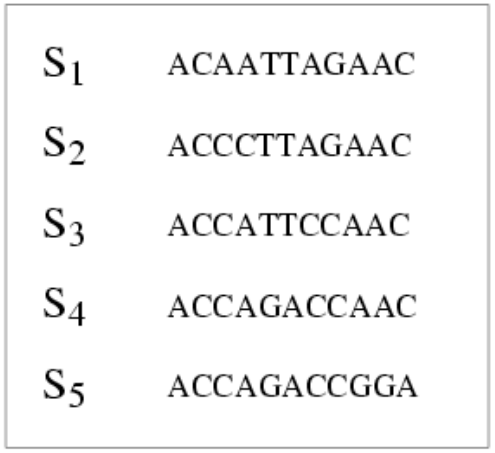
Simplest (Jukes-Cantor):

- The model tree is a pair (T,{e,p(e)}), where T is a rooted binary tree, and p(e) is the probability of a substitution on the edge e.
- The state at the root is random.
- If a site 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.



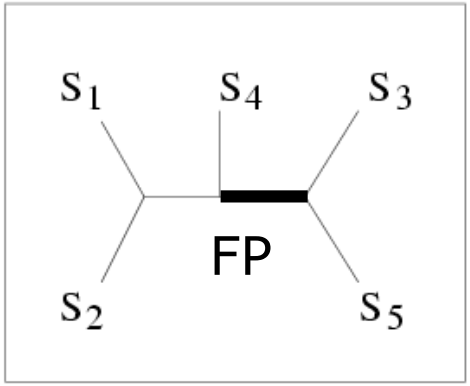
TRUE TREE



DNA SEQUENCES

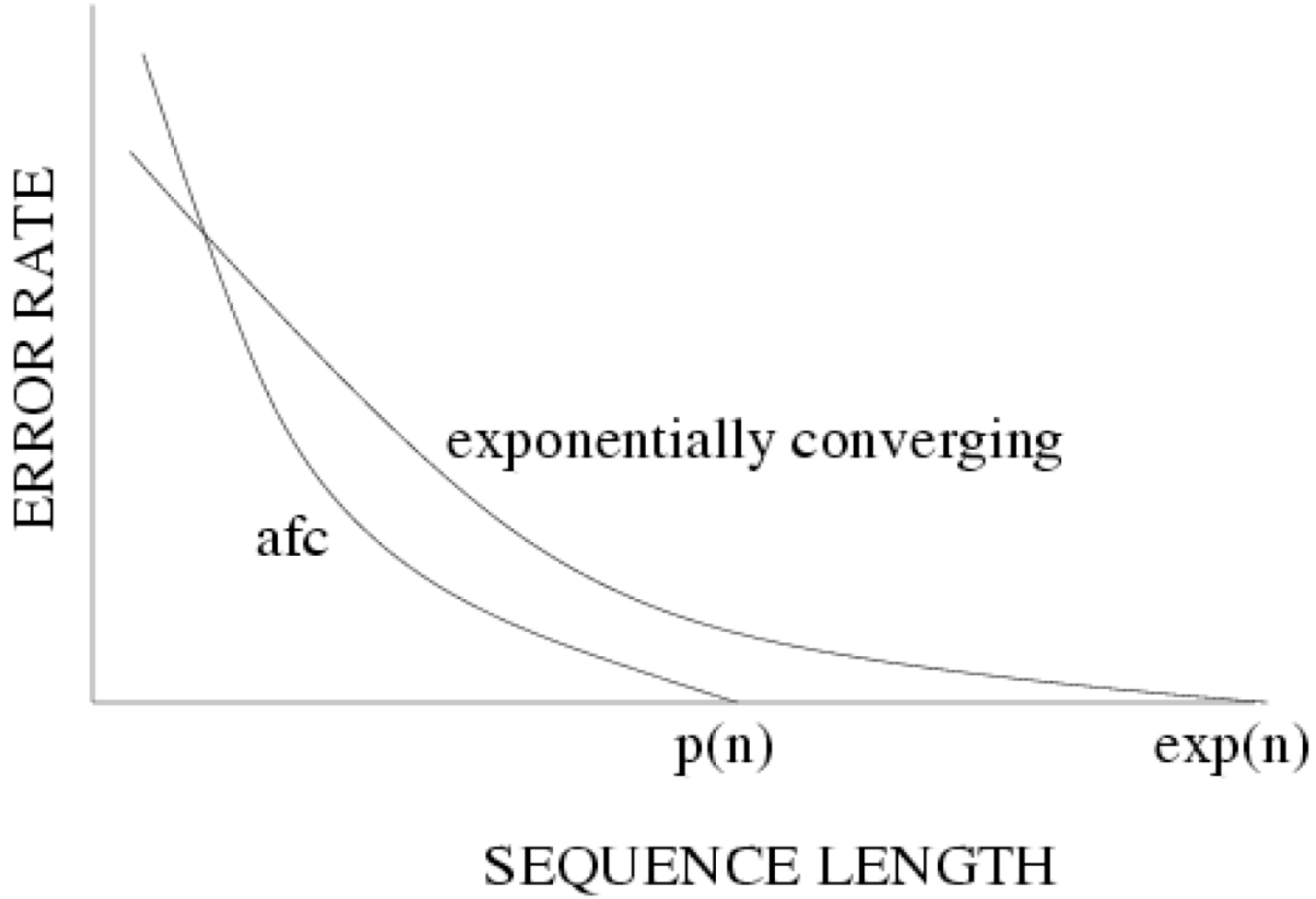
FN: false negative
(missing edge)
FP: false positive
(incorrect edge)

50% error rate

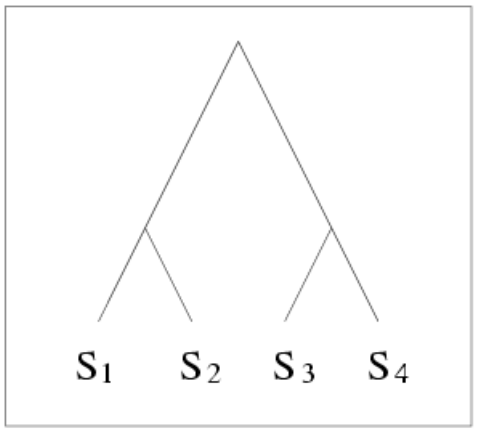


INFERRED TREE

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



Distance-based Phylogenetic Methods



TRUE TREE

S_1	ACAATTAGAAC
S_2	ACCCTTAGAAC
S_3	ACCATTCCAAC
S_4	ACCAGACCAAC

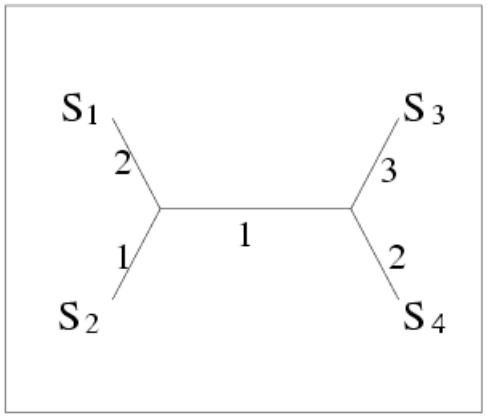
DNA SEQUENCES

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

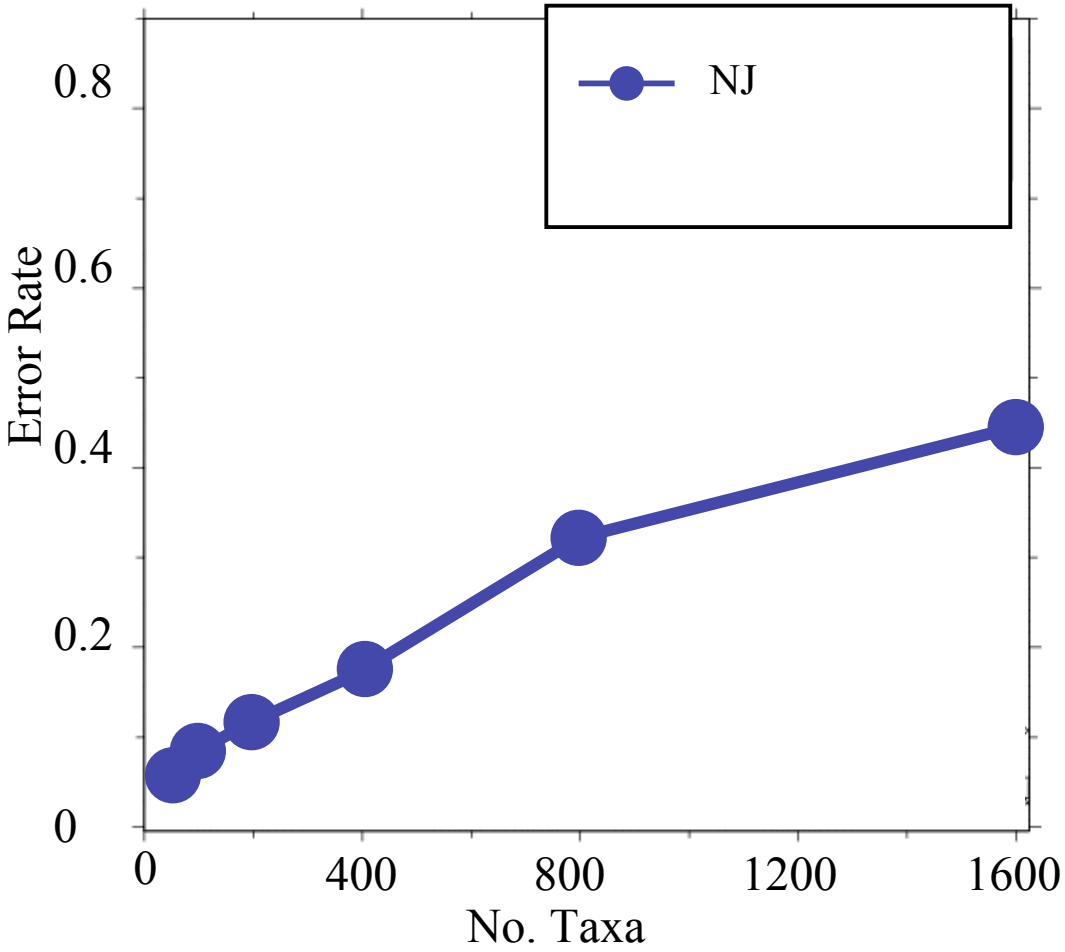


INFERRED TREE

- Theorem (Erdos, Szekely, Steel and Warnow 1997, Atteson 1997): Neighbor joining (and some other distance-based methods) will return the true tree with high probability provided sequence lengths are **exponential** in the diameter of the tree.

Neighbor joining's performance

[Nakhleh et al. ISMB 2001]

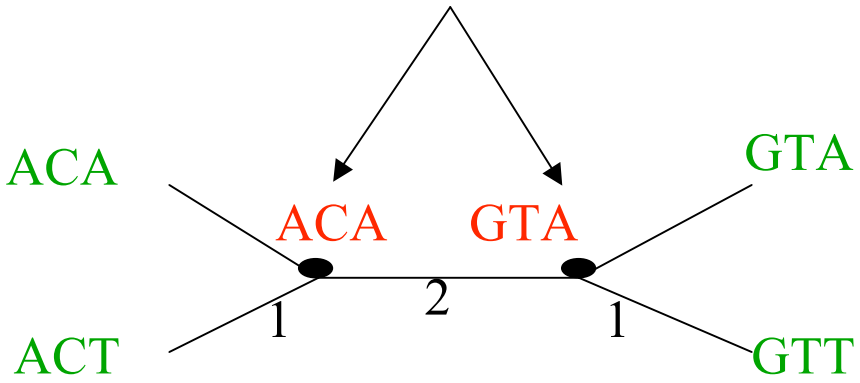


Simulation study

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

Maximum Parsimony:

Optimal labeling on a fixed tree can be computed in linear time O(nk)



MP score = 4

MP is not statistically consistent.
Finding the optimal MP tree is **NP-hard.**

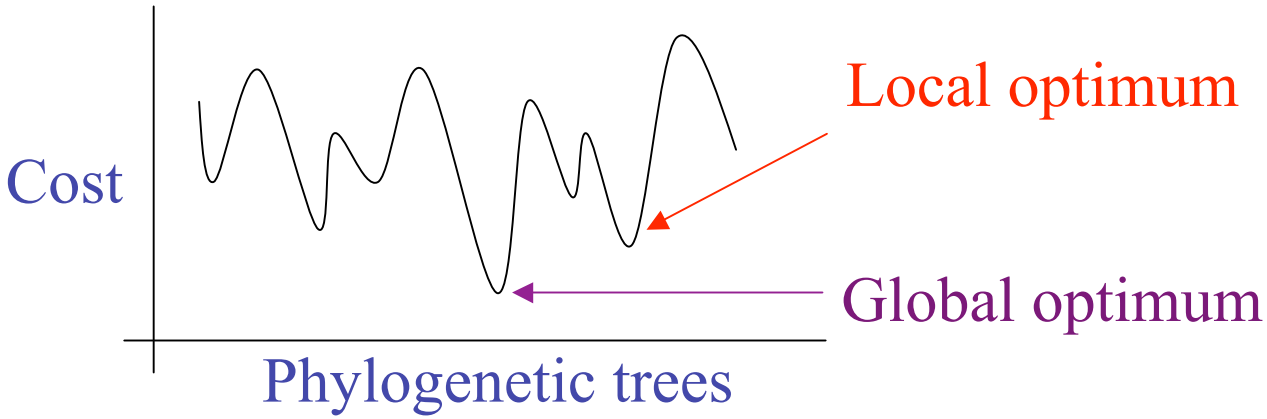
Maximum Likelihood (ML)

- Given: stochastic model of sequence evolution (e.g. Jukes-Cantor, or GTR+Gamma+I) and a set S of sequences
- Objective: Find tree T and parameter values so as to maximize the probability of the data.

NP-hard, but statistically consistent. Preferred by many systematists, but even harder than MP in practice. (Steel and Szekely proved that exponential sequence lengths suffice for accuracy with high probability.)

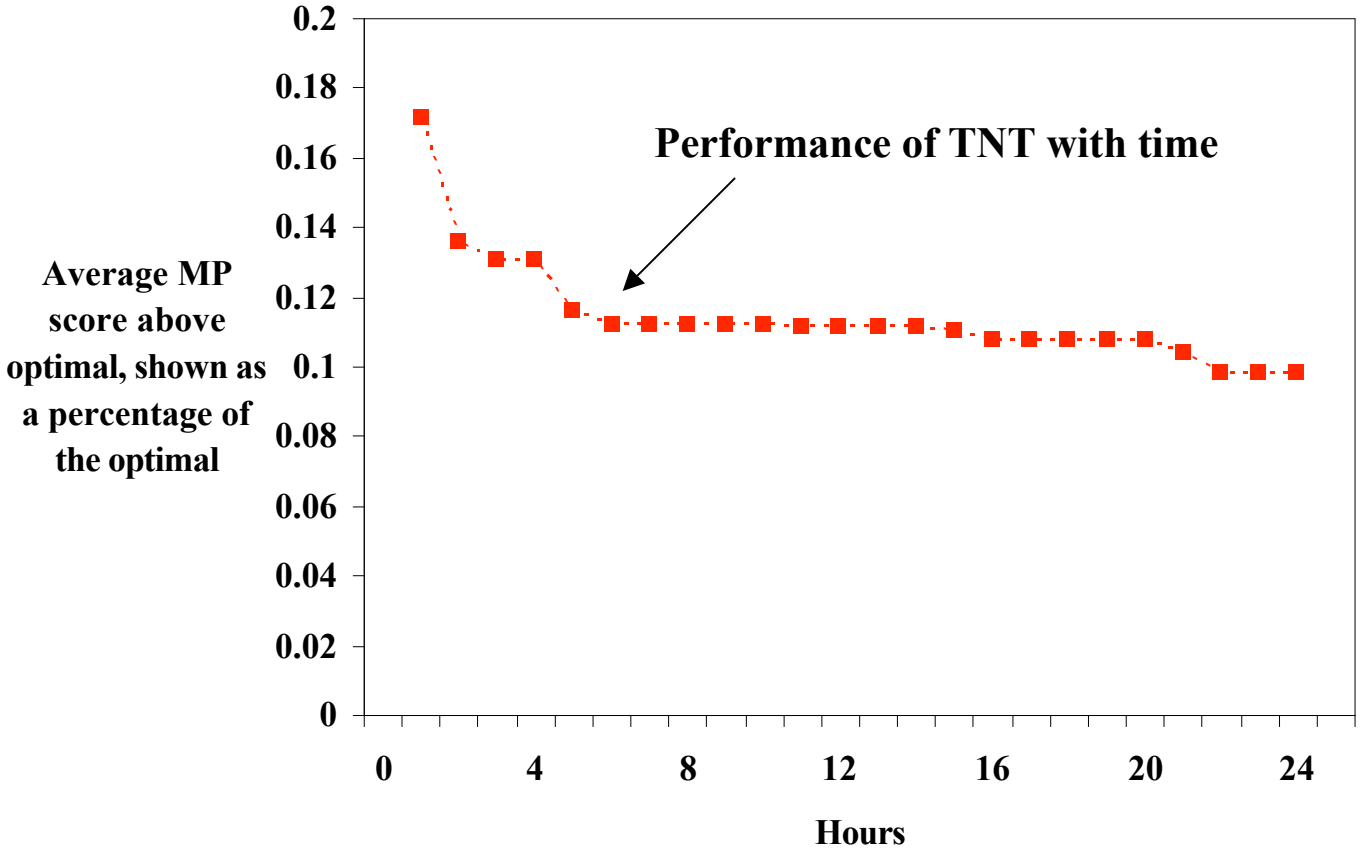
Approaches for "solving" MP and ML (and other NP-hard problems in phylogeny)

1. Hill-climbing heuristics (which can get stuck in local optima)
2. Randomized algorithms for getting out of local optima
3. Approximation algorithms for MP (based upon Steiner Tree approximation algorithms) -- however, the approx. ratio that is needed is probably 1.01 or smaller!



Problems with techniques for MP and ML

Shown here is the performance of a very good heuristic (TNT) for maximum parsimony analysis on a real dataset of almost 14,000 sequences. ("Optimal" here means ***best score to date***, using any method for any amount of time.) Acceptable error is below 0.01%.



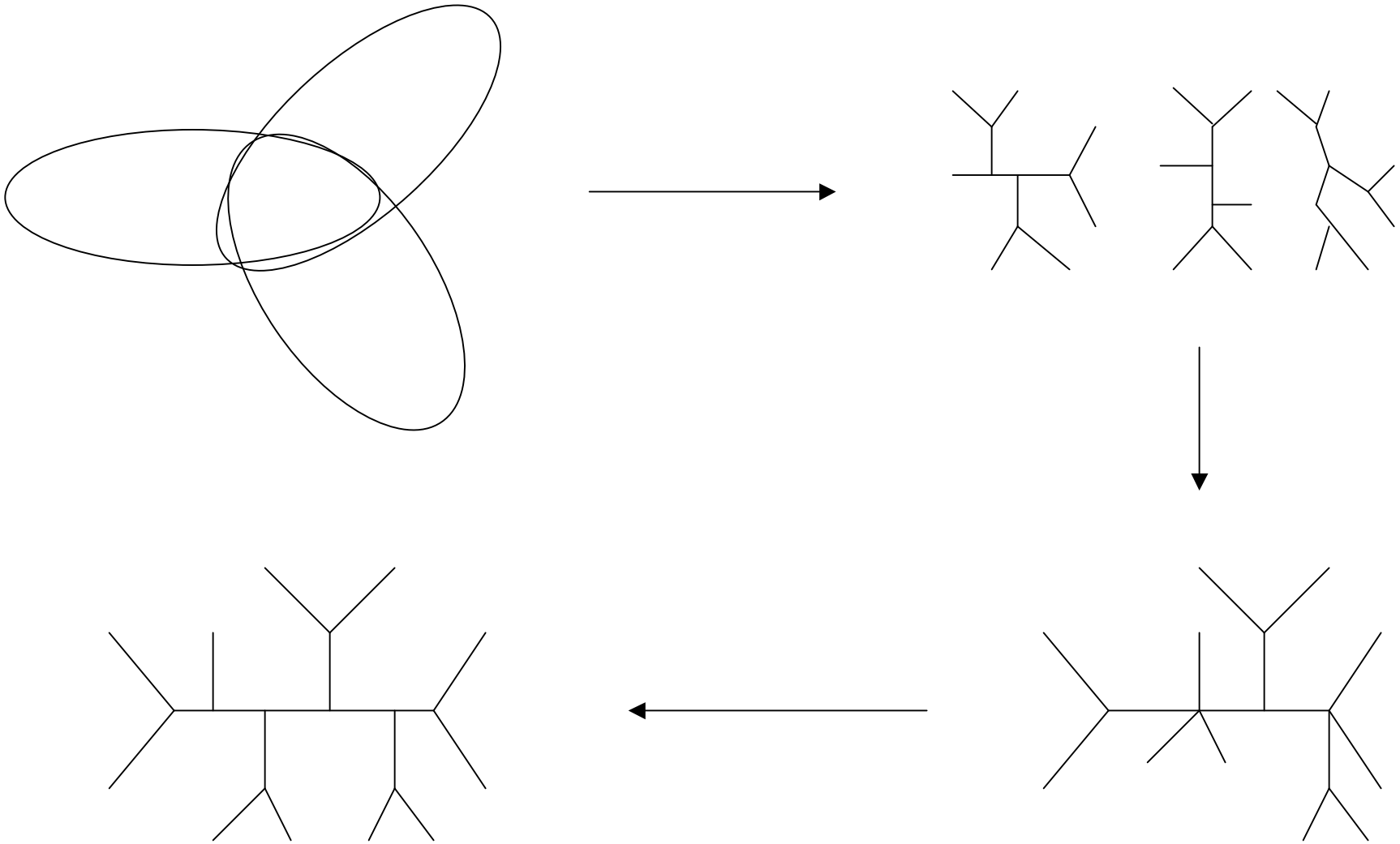
Problems with existing phylogeny reconstruction methods

- Polynomial time methods (generally based upon distances) have poor accuracy with large diameter datasets.
- Heuristics for NP-hard optimization problems take too long (months to reach acceptable local optima).

Warnow et al.: Meta-algorithms for phylogenetics

- Basic technique: determine the conditions under which a phylogeny reconstruction method does well (or poorly), and design a divide-and-conquer strategy (specific to the method) to improve its performance
- Warnow et al. developed a class of divide-and-conquer methods, collectively called DCMs (**Disk-Covering Methods**). These are based upon *chordal graph theory* to give fast decompositions and provable performance guarantees.

Disk-Covering Method (DCM)

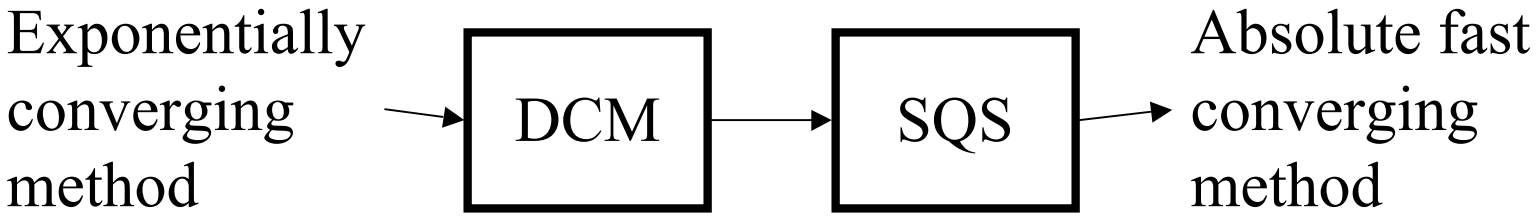


Improving phylogeny reconstruction methods using DCMs

- Improving the theoretical convergence rate and performance of polynomial time distance-based methods using DCM1
- Speeding up heuristics for NP-hard optimization problems (Maximum Parsimony and Maximum Likelihood) using Rec-I-DCM3

DCM1

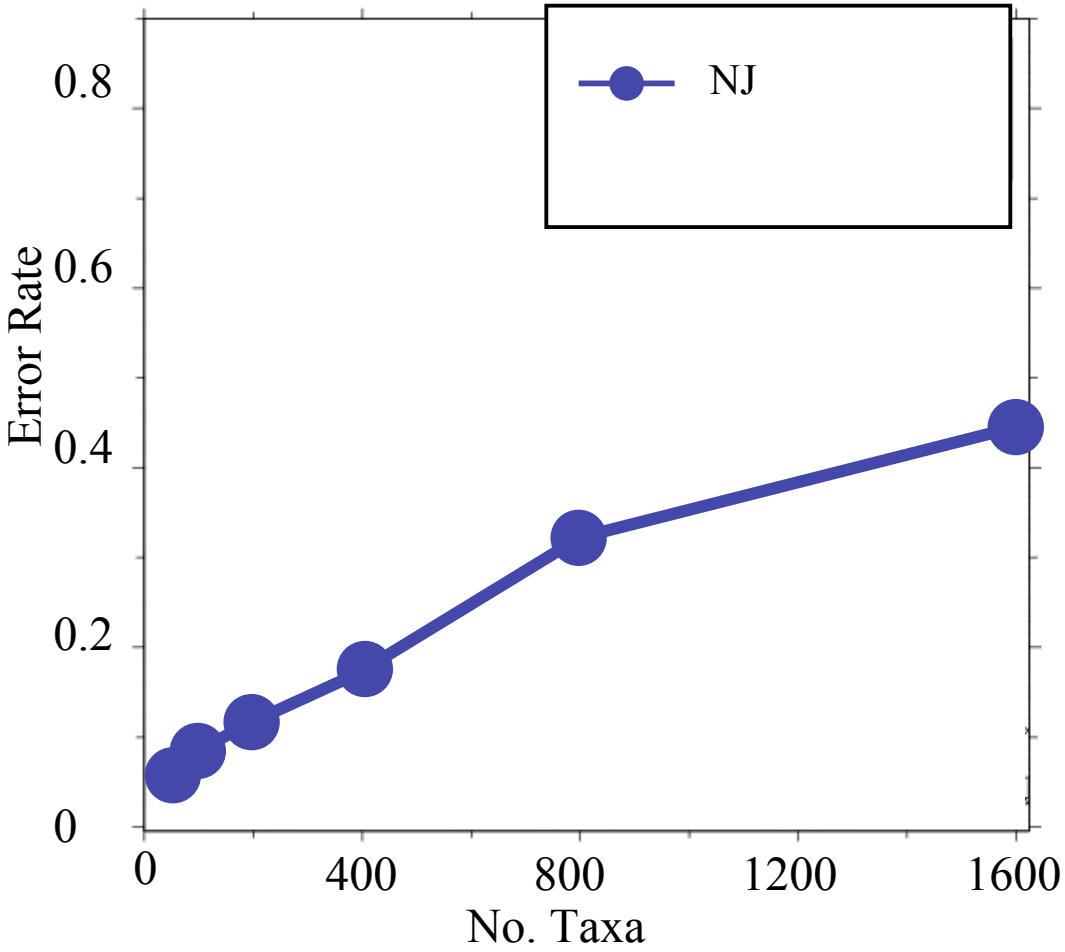
Warnow, St. John, and Moret, SODA 2001



- A two-phase procedure which reduces the sequence length requirement of methods. The DCM phase produces a collection of trees, and the SQS phase picks the "best" tree.
- The "base method" is applied to subsets of the original dataset. When the base method is NJ, you get DCM1-NJ.

Neighbor joining (although statistically consistent) has poor performance on large diameter trees

[Nakhleh et al. ISMB 2001]

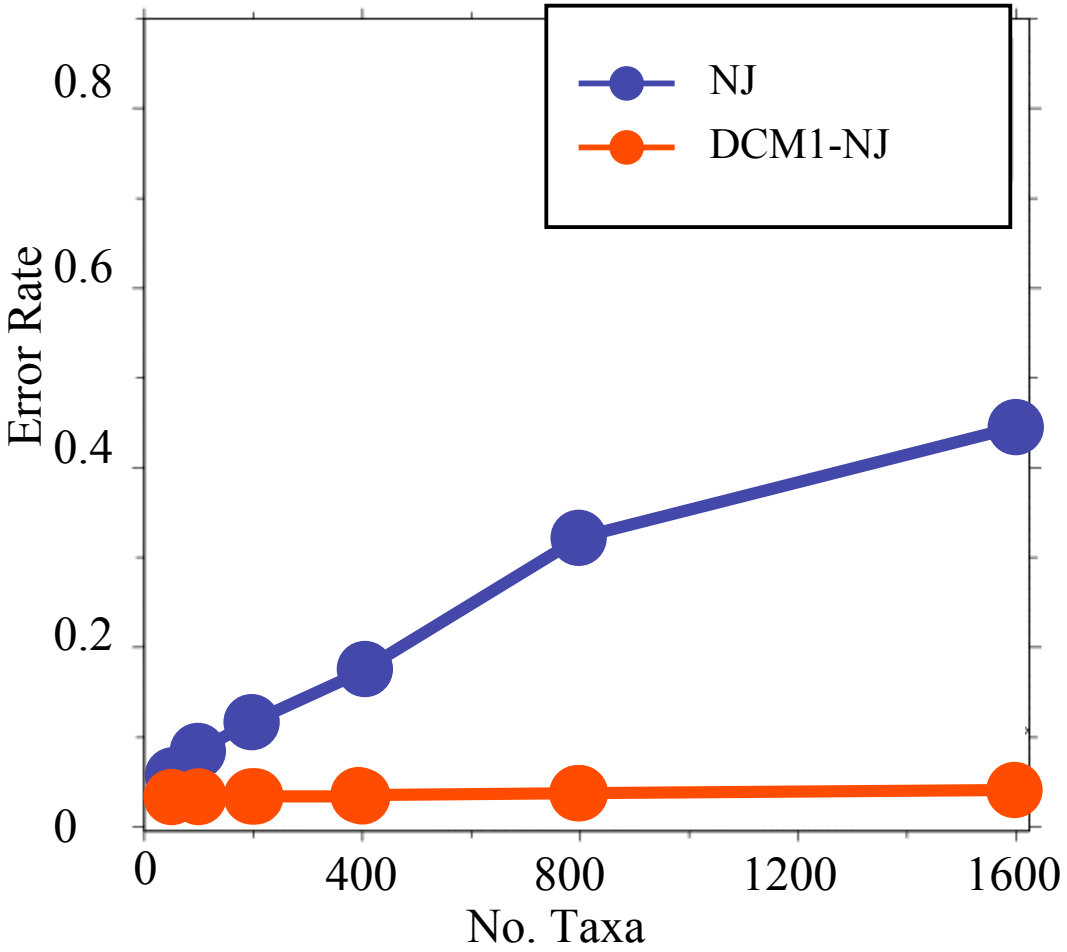


Simulation study based upon fixed edge lengths, K2P model of evolution, sequence lengths fixed to 1000 nucleotides.

Error rates reflect proportion of incorrect edges in inferred trees.

DCM1-boosting distance-based methods

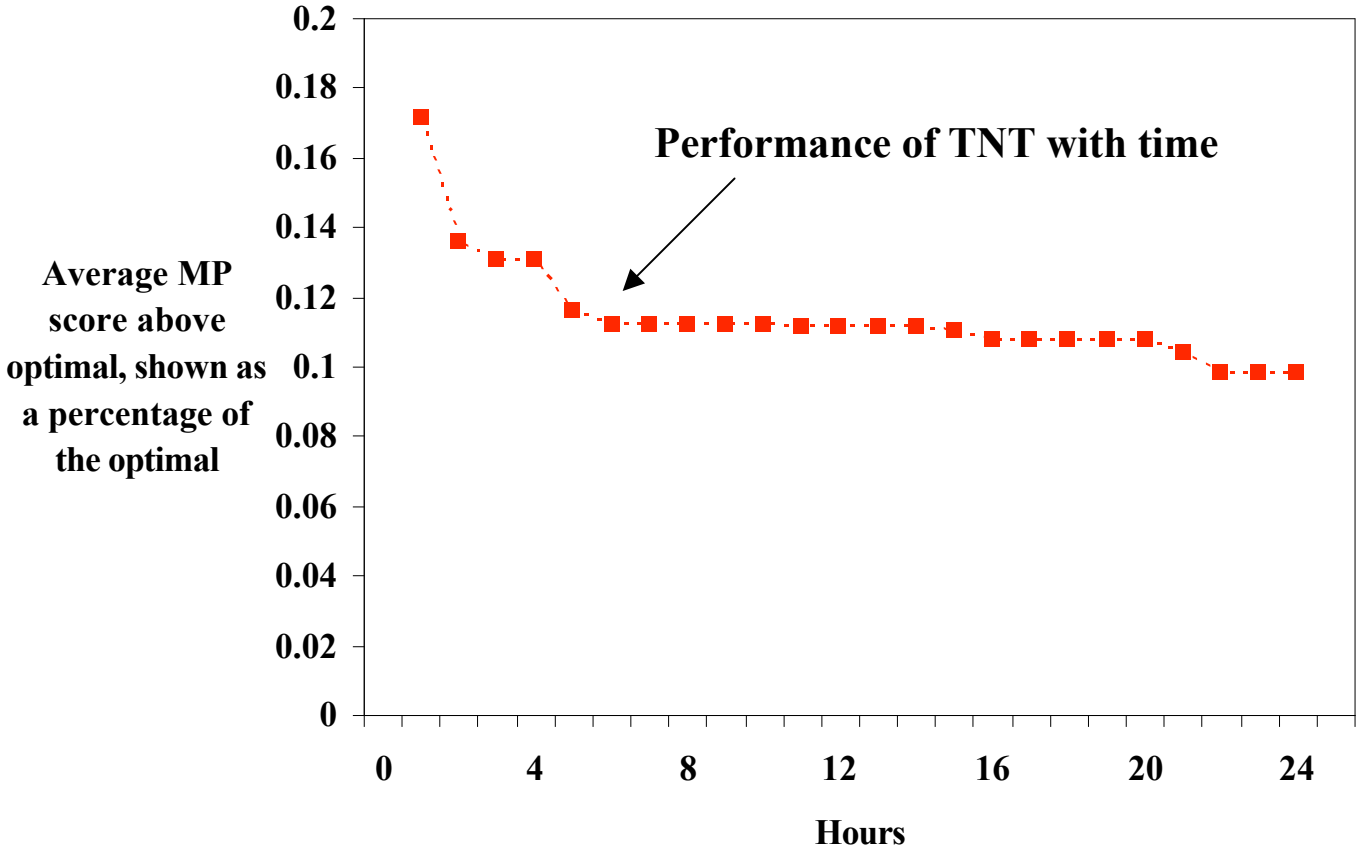
[Nakhleh et al. ISMB 2001]



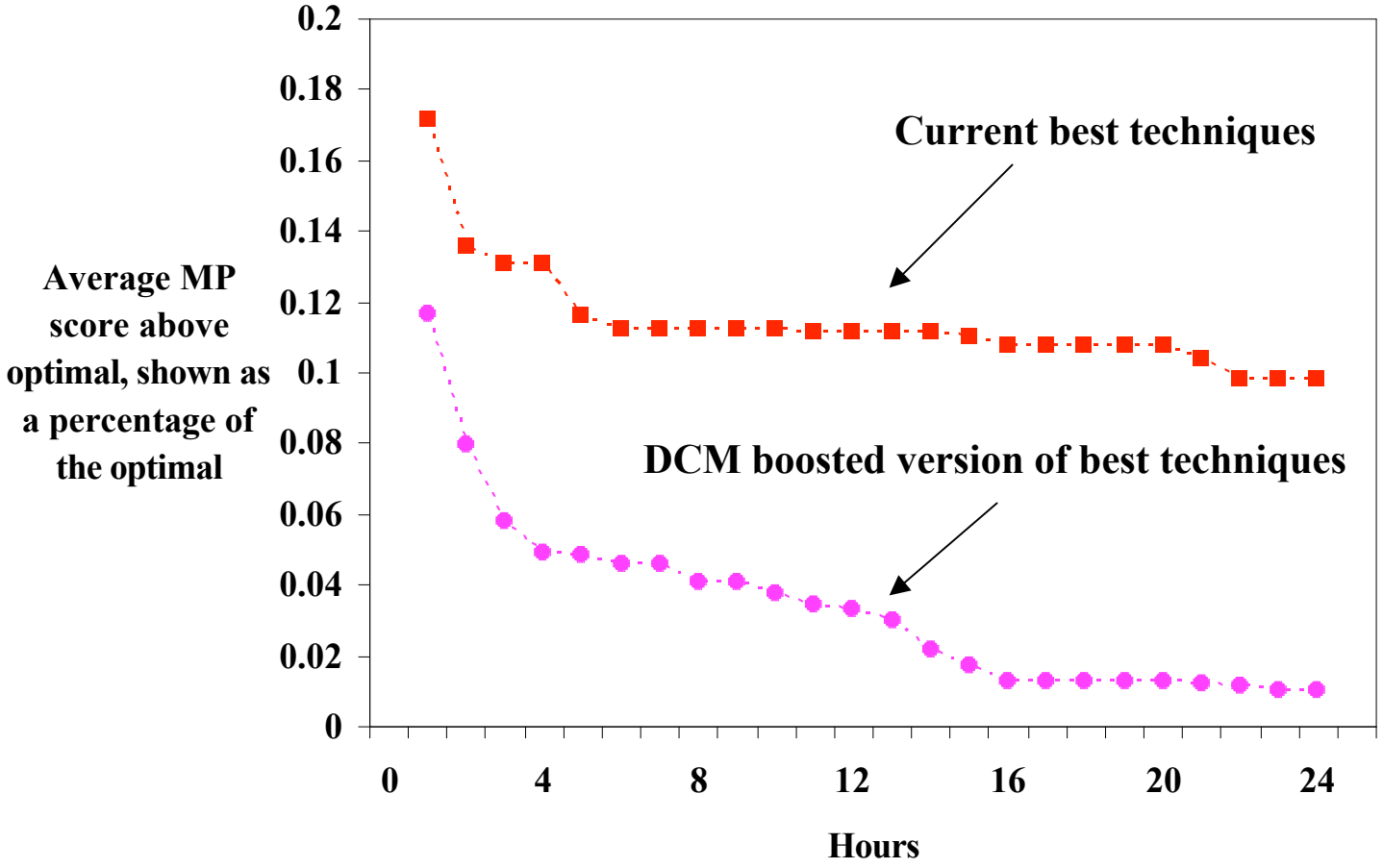
Theorem:
DCM1-NJ converges to the true tree from polynomial length sequences

Problems with techniques for MP and ML

Shown here is the performance of a TNT heuristic maximum parsimony analysis on a real dataset of almost 14,000 sequences. ("Optimal" here means ***best score to date***, using any method for any amount of time.) Acceptable error is below 0.01%.



Rec-I-DCM3 significantly improves performance (Roshan et al. CSB 2004)

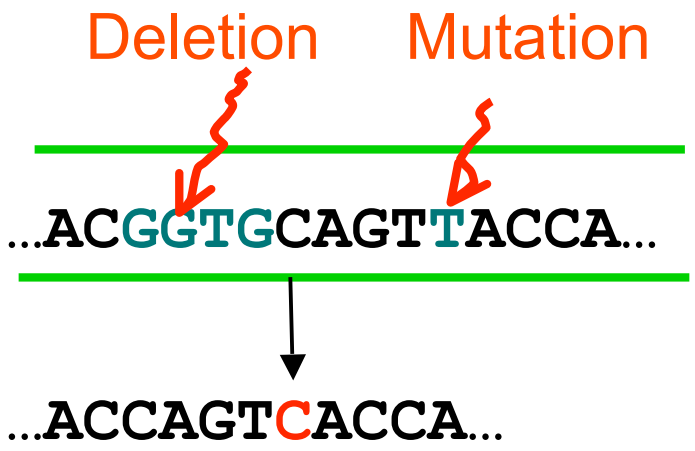


Comparison of TNT to Rec-I-DCM3(TNT) on one large dataset.

Similar improvements obtained for RAxML (maximum likelihood).

Very nice, but…

- ***Evolution is not as simple as these models assert!***



indels (insertions and deletions) also occur!

Basic Questions

- Does improving the alignment lead to an improved phylogeny?
- Are we getting good enough alignments from MSA methods?
- Are we getting good enough trees from the phylogeny reconstruction methods?
- Can we improve these estimations, perhaps through **simultaneous estimation** of trees and alignments?

Multiple Sequence Alignment

```
AGGCTATCACCTGACCTCCA    -AGGCTATCACCTGACCTCCA
TAGCTATCACGACCGC        TAG-CTATCAC--GACCGC--
TAGCTGACCGC             TAG-CT-------GACCGC--
```

Notes:

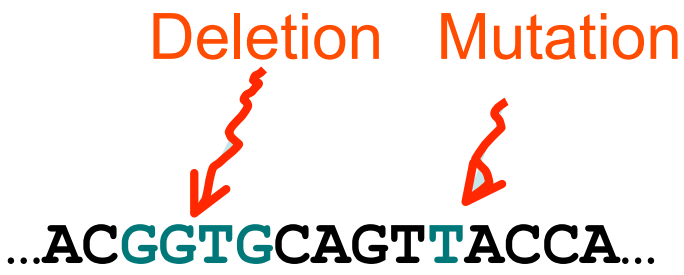
1. We insert gaps (dashes) to each sequence to make them “line up”.
2. Nucleotides in the same column are presumed to have a common ancestor (i.e., they are “homologous”).

Indels and substitutions at the DNA level

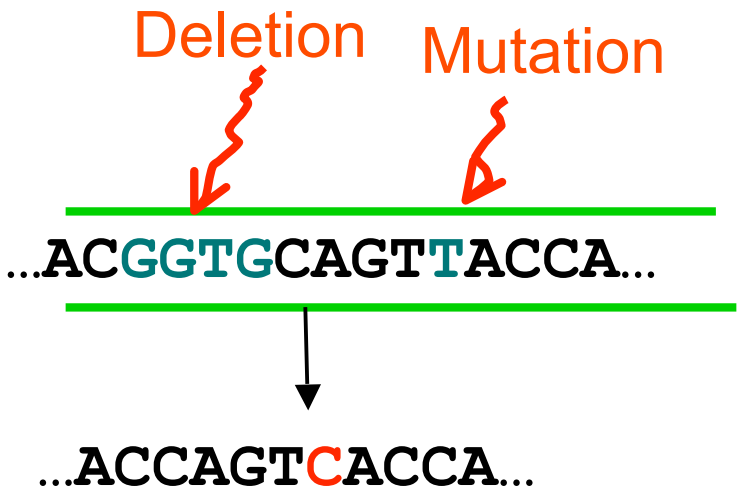
...**ACGGTGCAGTTACCA**...

Indels and substitutions at the DNA level

Indels and substitutions at the DNA level



...ACCAGTCACCA...



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.

Basics about alignments

- The standard alignment method for phylogeny is Clustal (or one of its derivatives), but many new alignment methods have been developed by the protein alignment community.
- Alignments are generally evaluated in comparison to the "true alignment", using the SP-score (percentage of truly homologous pairs that show up in the estimated alignment).
- On the basis of SP-scores (and some other criteria), methods like **ProbCons, Mafft,** and **Muscle** are generally considered "better" than Clustal.

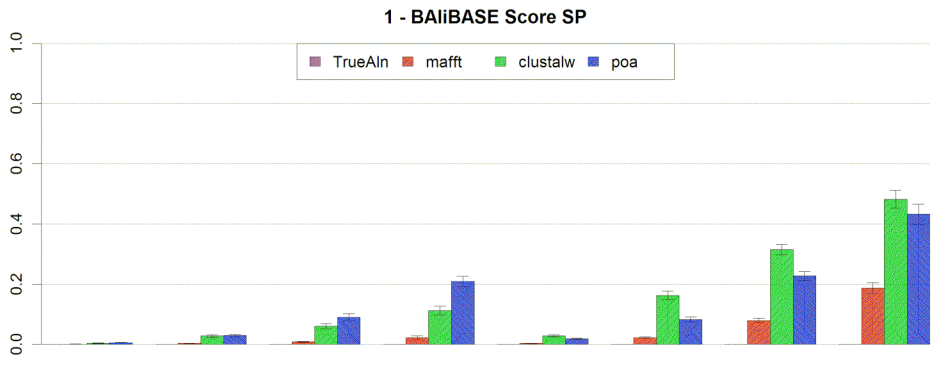
Questions

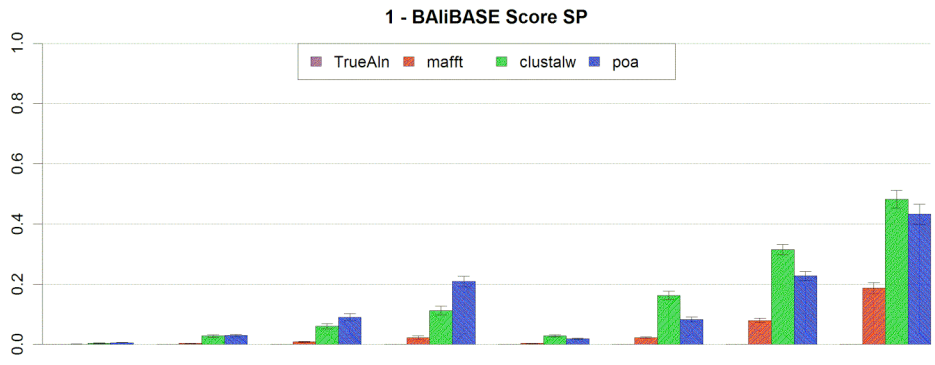
- Many new MSA methods improve on ClustalW on biological benchmarks (e.g., BaliBASE) and in simulation. Does this lead to improved phylogenetic estimations?
- The phylogeny community has tended to assume that alignment has a big impact on final phylogenetic accuracy. But does it? Does this depend upon the model conditions?
- What are the best two-phase methods?

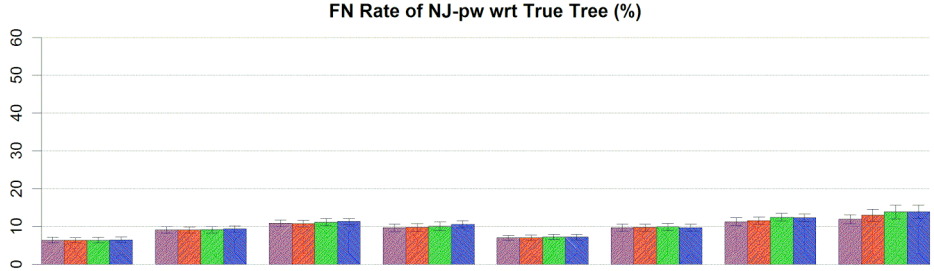
Our simulation studies (using ROSE*)

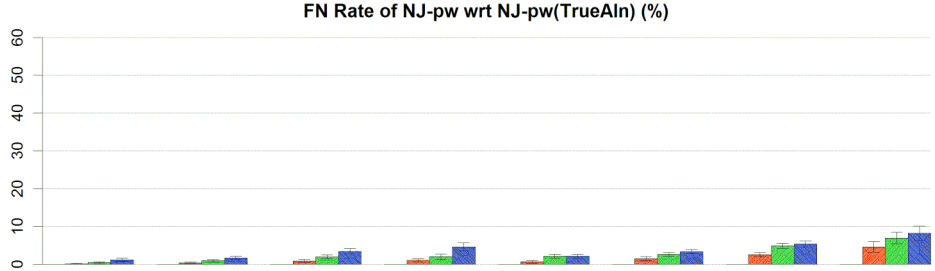
- Amino-acid evolution (Wang et al., unpublished):
  - BaliBase and birth-death model trees, 12 taxa to 100 taxa.
  - Average gap length 3.4.
  - Average identity 23% to 57%.
  - Average gappiness 3% to 60%.
- DNA sequence evolution (Liu et al., unpublished):
  - Birth-death trees, 25 to 500 taxa.
  - Two gap length distributions (short and long).
  - Average p-distance 43% to 63%.
  - Average gappiness 40% to 80%.

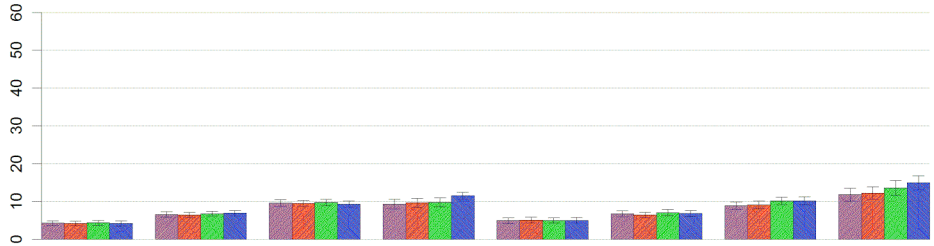
ROSE has limitations!

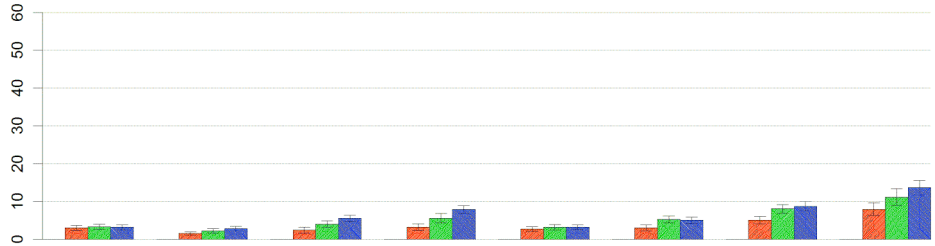

1 - BAliBASE Score SP
TrueAln
mafft
clustalw
poa
0.0
0.2
0.4
0.6
0.8
1.0

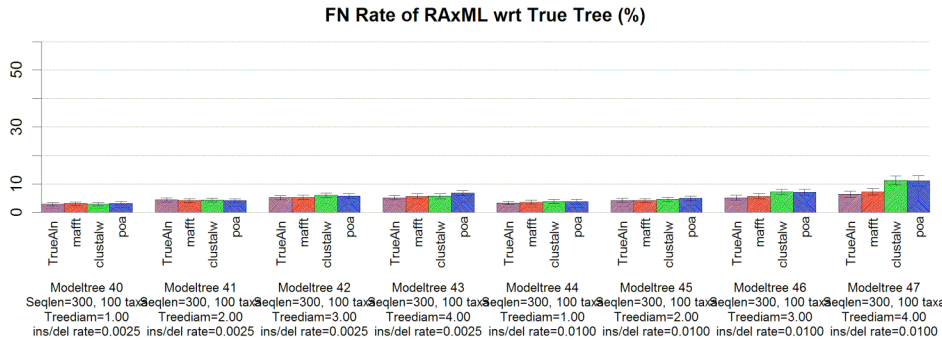

1 - BAliBASE Score SP
TrueAln
mafft
clustalw
poa
0.0
0.2
0.4
0.6
0.8
1.0

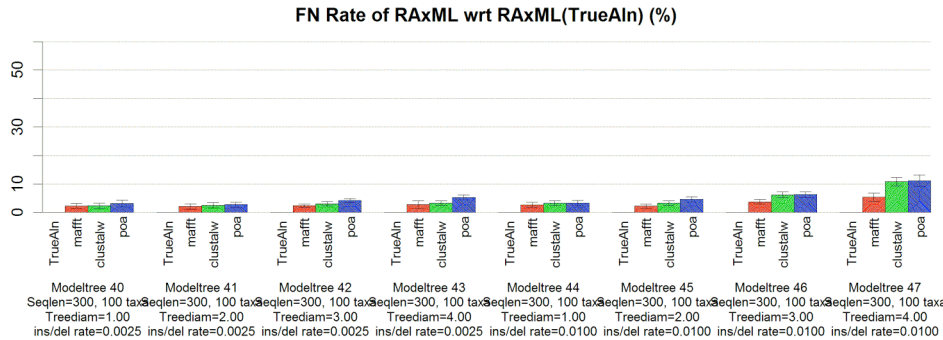

FN Rate of NJ-pw wrt True Tree (%)
0
10
20
30
40
50
60


FN Rate of NJ-pw wrt NJ-pw(TrueAln) (%)
0
10
20
30
40
50
60

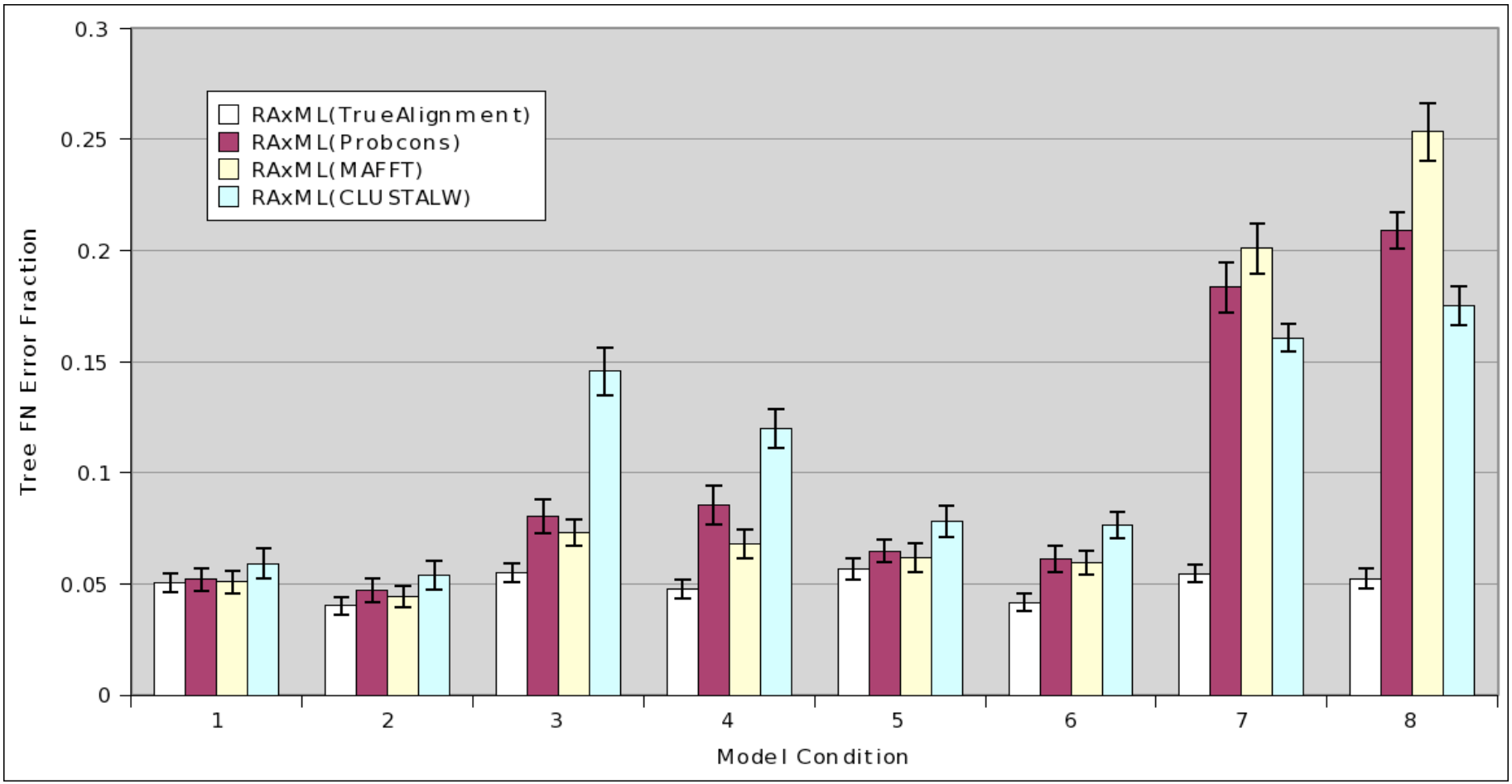

FN Rate of Maj-PAMwRatchet(InfAln) wrt True Tree (%)
0
10
20
30
40
50
60


FN Rate of Maj-PAMwRatchet(InfAln) wrt Maj-PAMwRatchet(TrueAln) (%)
0
10
20
30
40
50
60


FN Rate of RAxML wrt True Tree (%)
0
10
30
50
TrueAln
mafft
clustalw
poa
Modeltree 40
Seqlen=300, 100 taxa
Treediam=1.00
ins/del rate=0.0025
Modeltree 41
Seqlen=300, 100 taxa
Treediam=2.00
ins/del rate=0.0025
Modeltree 42
Seqlen=300, 100 taxa
Treediam=3.00
ins/del rate=0.0025
Modeltree 43
Seqlen=300, 100 taxa
Treediam=4.00
ins/del rate=0.0025
Modeltree 44
Seqlen=300, 100 taxa
Treediam=1.00
ins/del rate=0.0100
Modeltree 45
Seqlen=300, 100 taxa
Treediam=2.00
ins/del rate=0.0100
Modeltree 46
Seqlen=300, 100 taxa
Treediam=3.00
ins/del rate=0.0100
Modeltree 47
Seqlen=300, 100 taxa
Treediam=4.00
ins/del rate=0.0100


FN Rate of RAxML wrt RAxML(TrueAln) (%)
0
10
30
50
TrueAln
mafft
clustalw
poa
Modeltree 40
Seqlen=300, 100 taxa
Treediam=1.00
ins/del rate=0.0025
Modeltree 41
Seqlen=300, 100 taxa
Treediam=2.00
ins/del rate=0.0025
Modeltree 42
Seqlen=300, 100 taxa
Treediam=3.00
ins/del rate=0.0025
Modeltree 43
Seqlen=300, 100 taxa
Treediam=4.00
ins/del rate=0.0025
Modeltree 44
Seqlen=300, 100 taxa
Treediam=1.00
ins/del rate=0.0100
Modeltree 45
Seqlen=300, 100 taxa
Treediam=2.00
ins/del rate=0.0100
Modeltree 46
Seqlen=300, 100 taxa
Treediam=3.00
ins/del rate=0.0100
Modeltree 47
Seqlen=300, 100 taxa
Treediam=4.00
ins/del rate=0.0100

Non-coding DNA evolution



Models 1-4 have “long gaps”, and models 5-8 have “short gaps”

Observations

- Phylogenetic tree accuracy is positively correlated with alignment accuracy (measured using SP), but the degree of improvement in tree accuracy is ***much smaller***.
- The best two-phase methods are generally (but not always!) obtained by using either ProbCons or MAFFT, followed by Maximum Likelihood.
- However, even the best two-phase methods don't do well enough.

Two problems with two-phase methods

- All current methods for multiple alignment have high error rates when sequences evolve with many indels and substitutions.
- All current methods for phylogeny estimation treat indel events inadequately (either treating as missing data, or giving too much weight to each gap).

Simultaneous estimation?

- Statistical methods (e.g., AliFritz and BaliPhy) cannot be applied to datasets above ~20 sequences.
- POY (Wheeler et al.) attempts to find tree/alignment pairs of minimum total edit distance. POY can be applied to larger datasets, but has not performed as well as the best two-phase methods.

SATe:

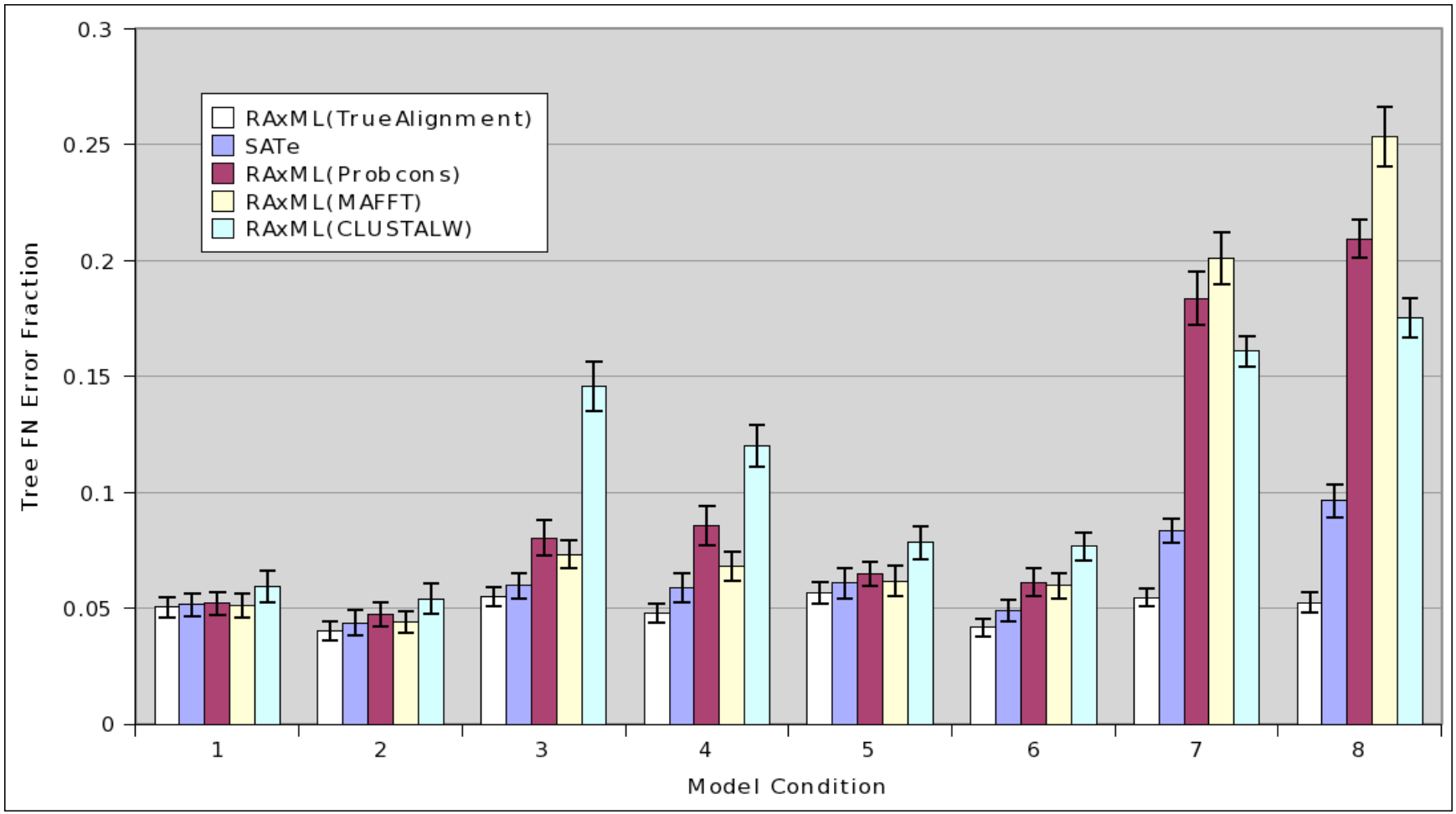
(Simultaneous Alignment and Tree Estimation)

- Developers: Warnow, Linder, Liu, Nelesen, and Zhao.
- Technique: search through tree space, and align sequences on each tree by ***heuristically estimating ancestral sequences*** and compute ML trees on the resultant multiple alignments.
- **SATe** returns the alignment/tree pair that optimizes maximum likelihood under GTR+Gamma+I.

Simulation study

- 100 taxon model trees (generated by r8s and then modified, so as to deviate from the molecular clock).
- DNA sequences evolved under ROSE (indel events of blocks of nucleotides, plus HKY site evolution). The root sequence has 1000 sites.
- We vary the gap length distribution, probability of gaps, and probability of substitutions, to produce 8 model conditions: models 1-4 have “long gaps” and 5-8 have “short gaps”.

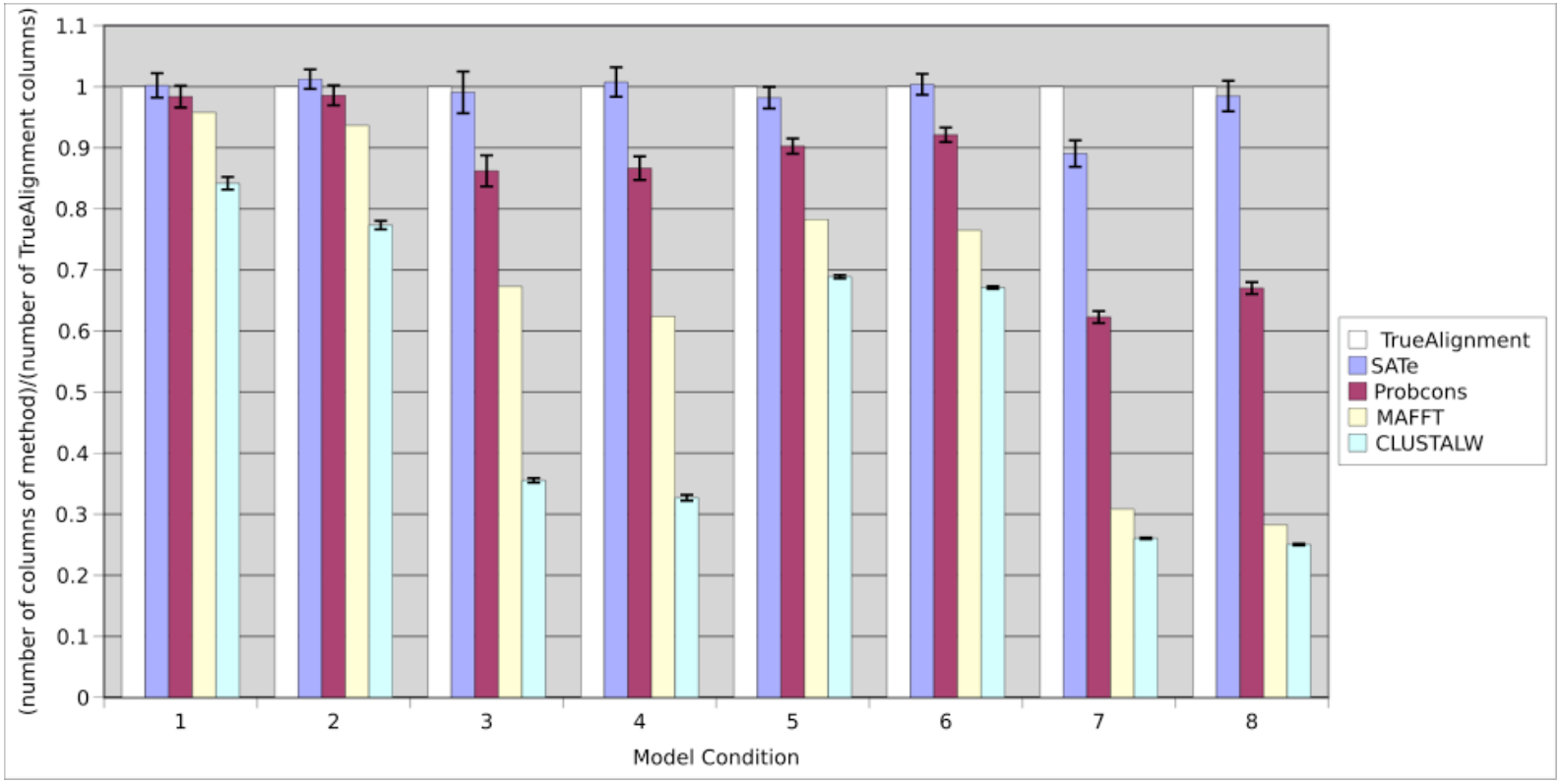
Our method (SATe) vs. other methods



- Long gap models 1-4, Short gap models 5-8

Alignment length accuracy

- Normalized number of columns in the estimated alignment relative to the true alignment.



Summary

- SATe improves upon the two-phase techniques we studied with respect to tree accuracy, and with respect to alignment length.
- SATe's performance depends upon how long you run it (these experiments limited to 48 hours).
- SATe is under development!

Note: SATe's algorithmic strategy is very different from most other alignment methods.

The **CIPRES Portal** contains Rec-I-DCM3 versions of parsimony and maximum likelihood, and we plan to add SATe.

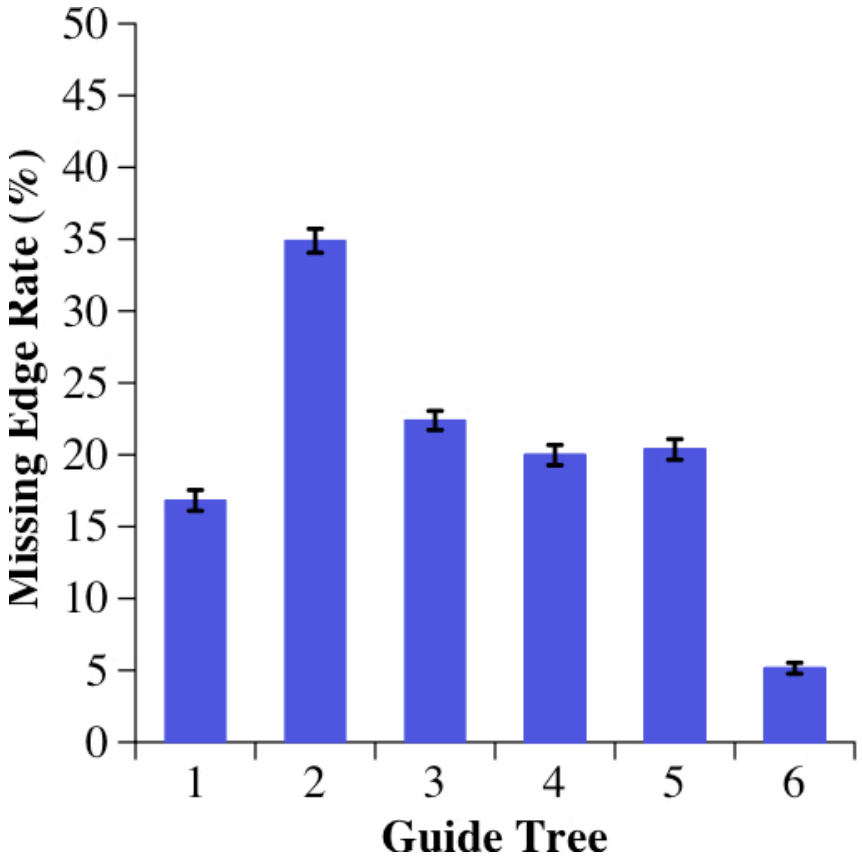
Future work

- *Better models and better simulators!!! (ROSE is limited)*
- *Extension of SATe-ML to models that include gap events (indels, duplications, and rearrangements)*
- Better metrics for alignment accuracy that are predictive of phylogenetic accuracy
- New data structures and visualization tools for representing homologies

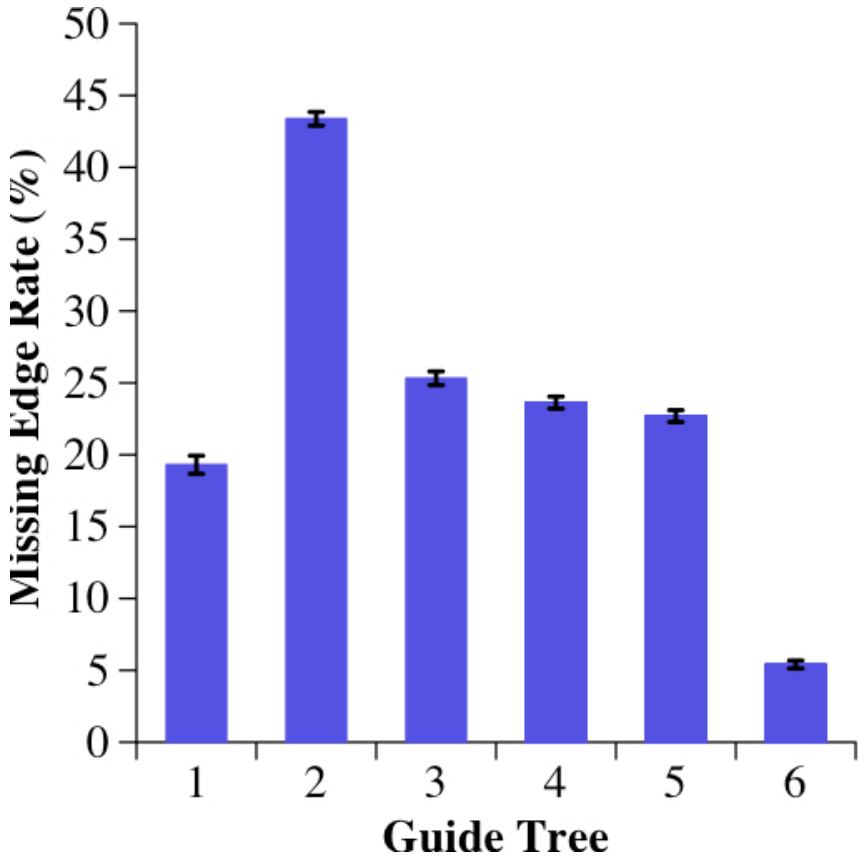
Acknowledgements

- 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: Claude de Pamphilis, Peter Erdos, Daniel Huson, Jim Leebens-Mack, Randy Linder, Kevin Liu, Bernard Moret, Serita Nelesen, Usman Roshan, Mike Steel, Katherine St. John, Laszlo Szekely, Li-San Wang, Tiffani Williams, and David Zhao.
- Thanks also to Li-San Wang and Serafim Batzoglou (slides)

Guide Tree Accuracy



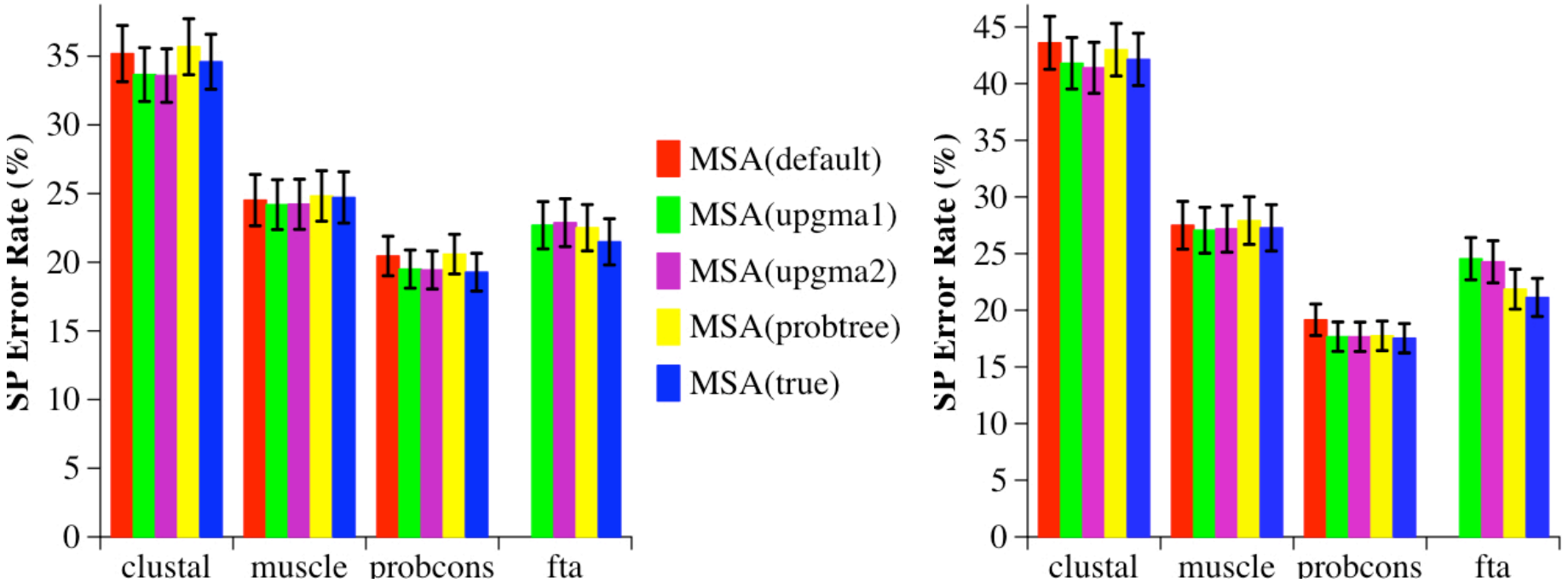
25 taxa



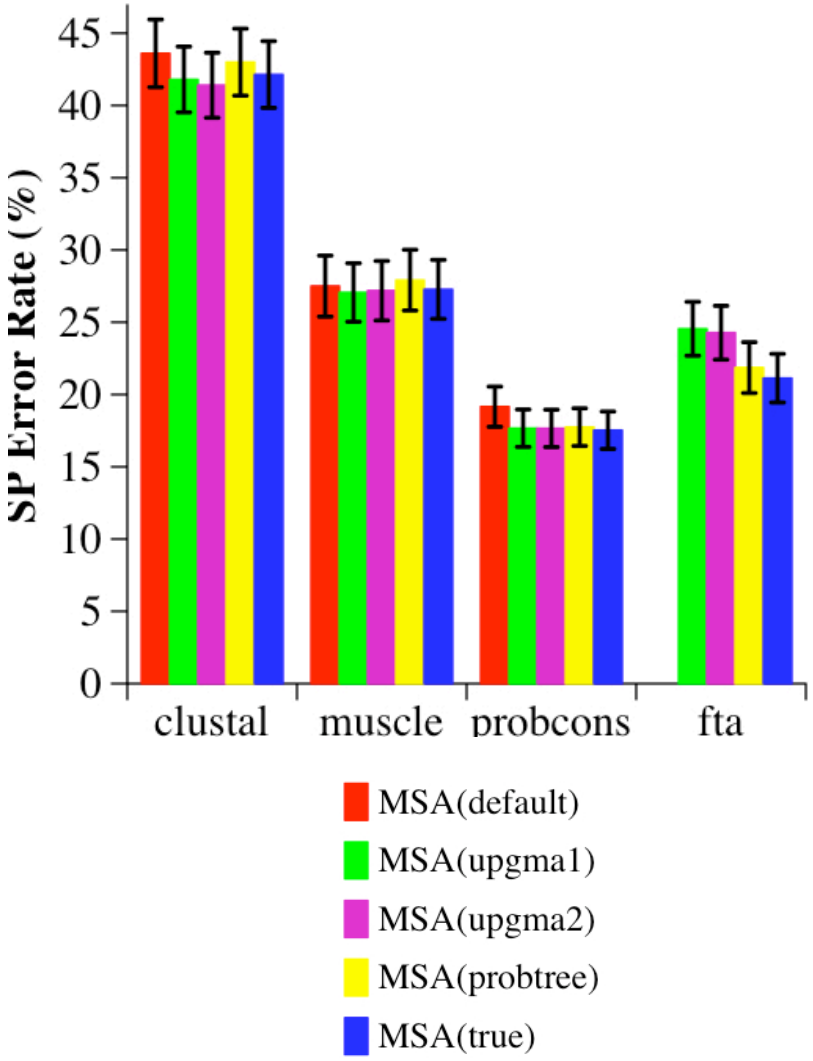
100 taxa

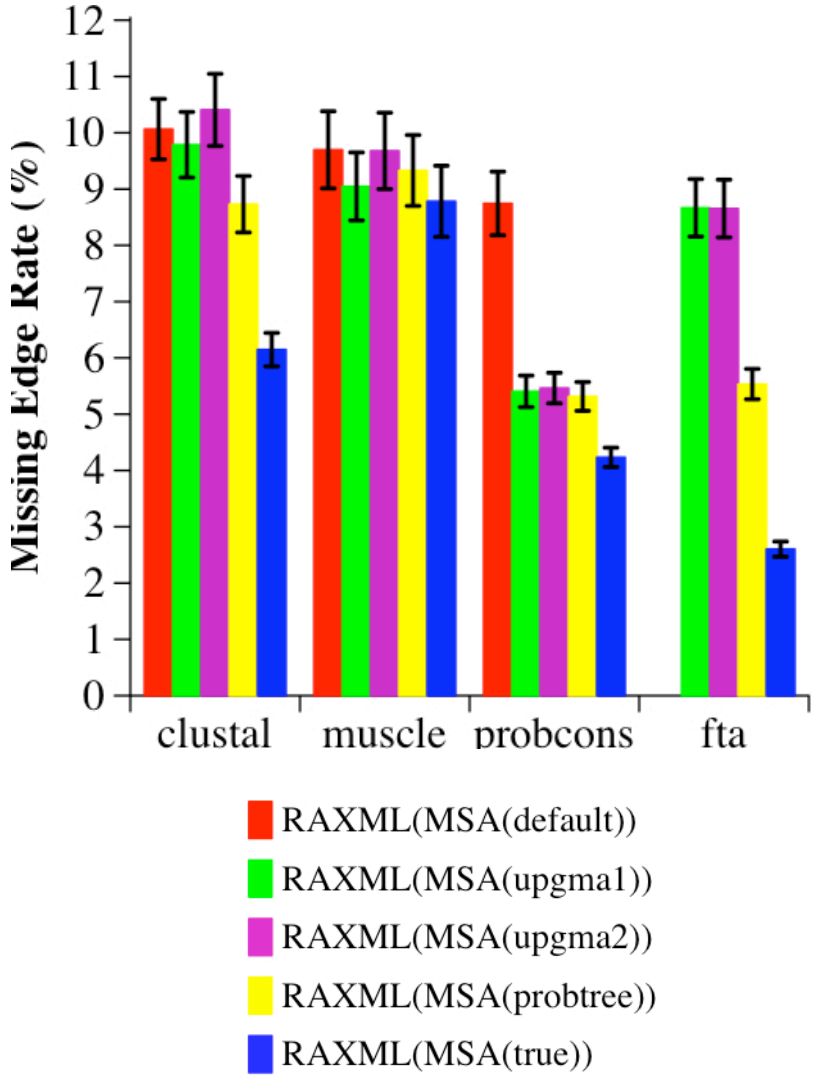
1-Clustal default, 2- ProbCons default, 3-Muscle default, 4-UPGMA1, 5-UPGMA2, 6-ProbTree

SP-Error Rates



Error Rates (100 Taxa)

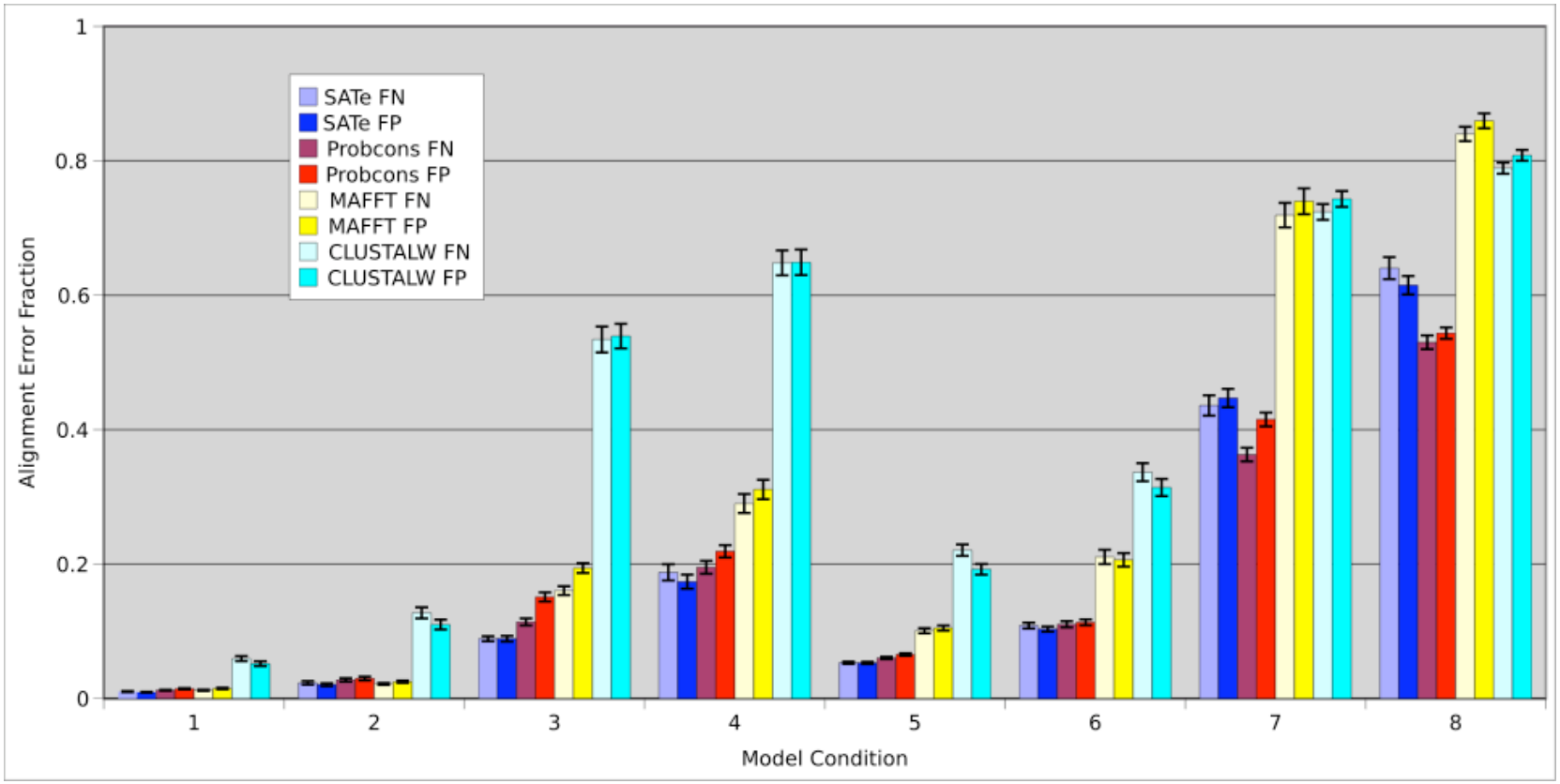




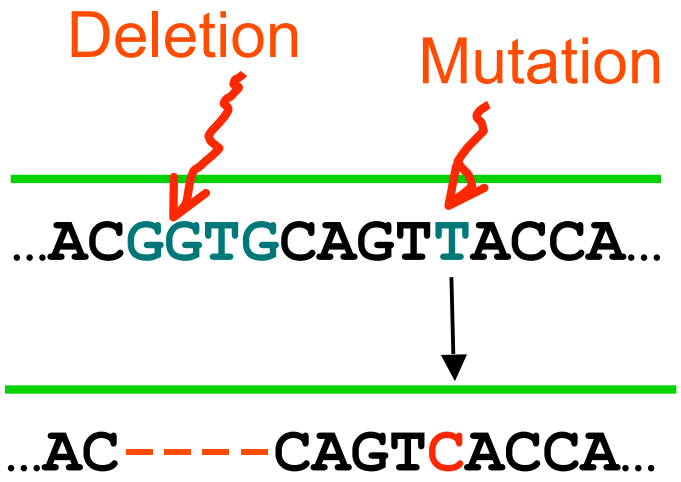
Alignment accuracy

FN: proportion of correctly homologous pairs of nucleotides *missing* from the estimated alignment (i.e., 1-SP score).

FP: proportion of incorrect pairings of nucleotides in the estimated alignment.

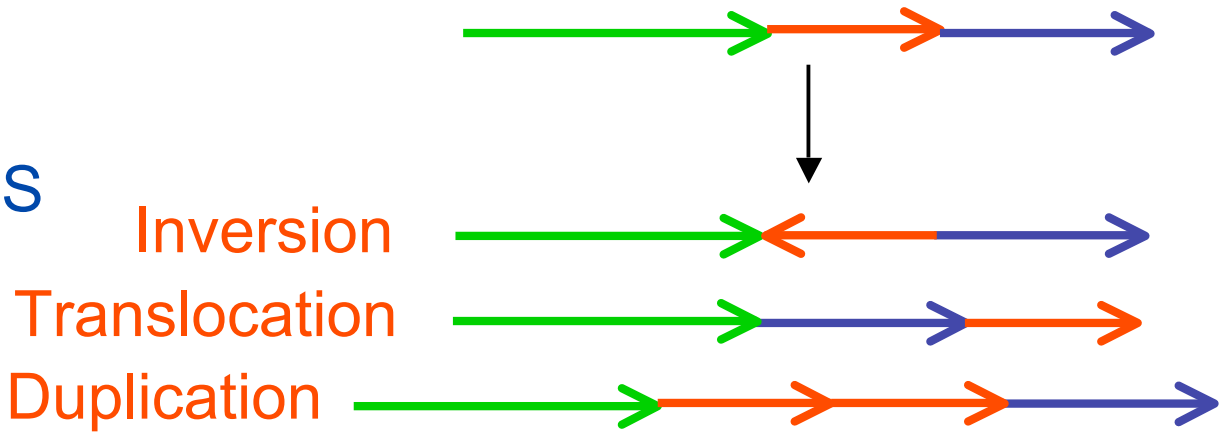


(but evolution is more complicated than that!)

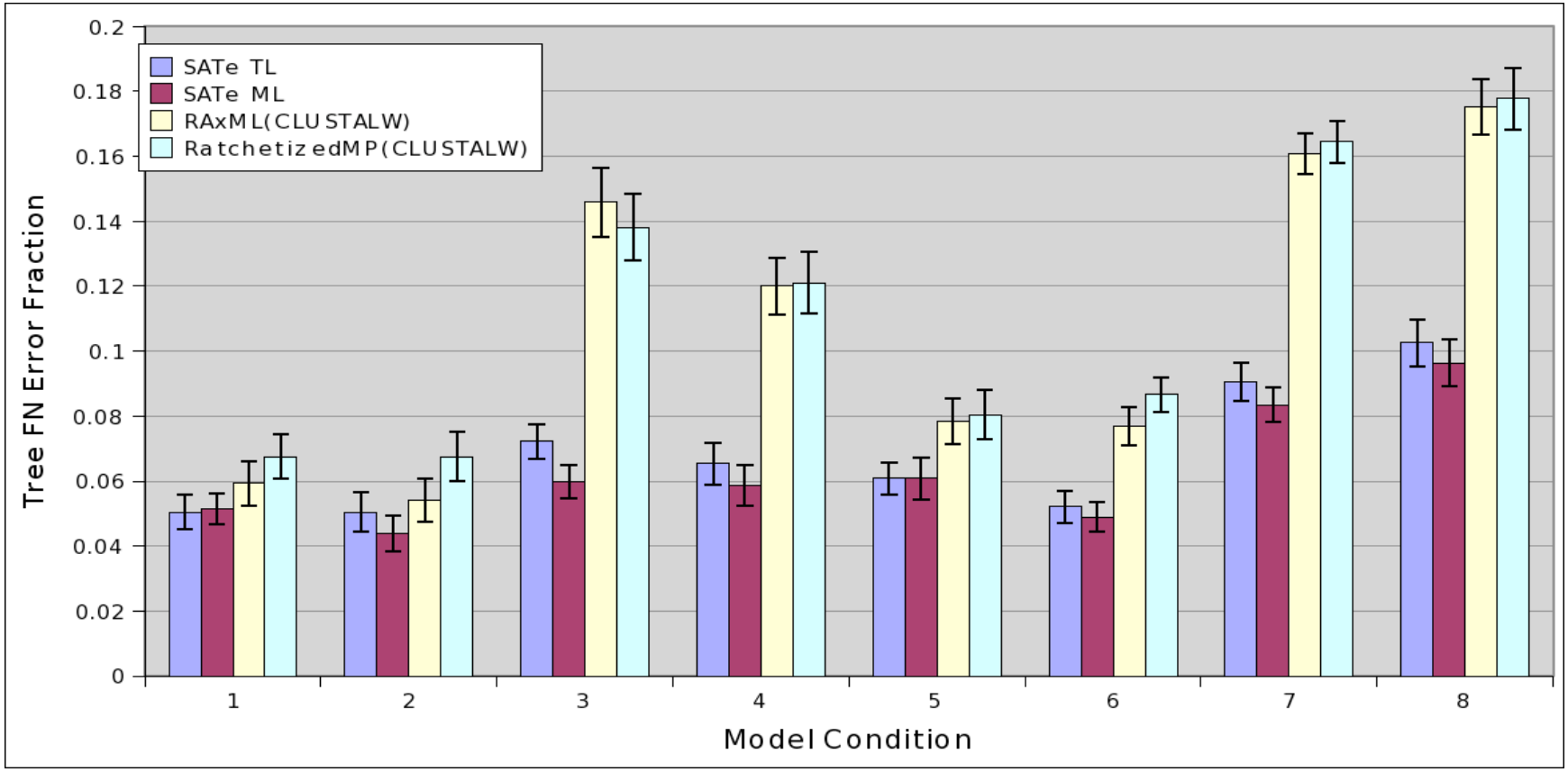


SEQUENCE EDITS

REARRANGEMENTS



SATe-TL vs. SATe-ML vs. Clustal



- Model conditions 1-4 have long gaps (100 taxa)
- Model conditions 5-8 have short gaps (100 taxa)