New methods for simultaneous estimation of trees and alignments

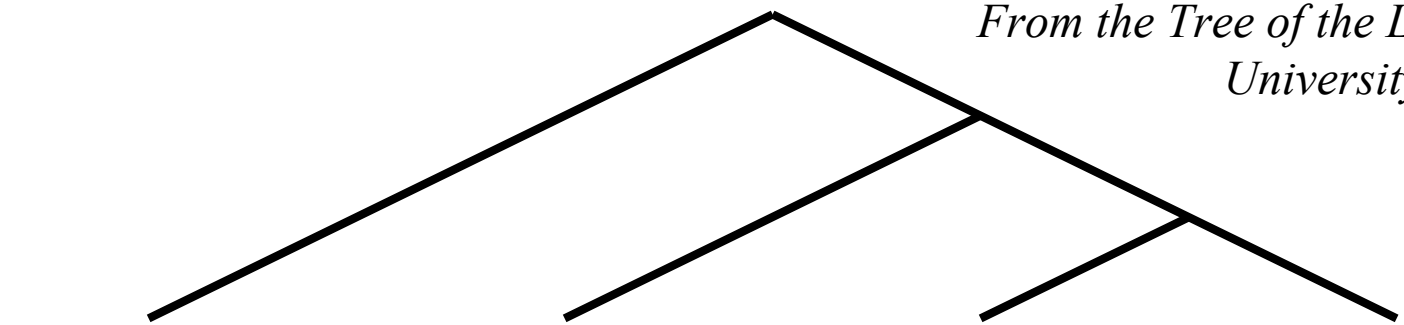
Tandy Warnow

The University of Texas at Austin

CIPRES

Species phylogeny

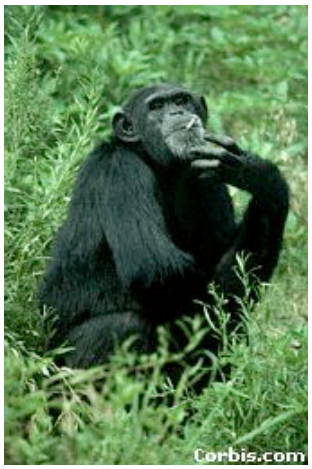
From the Tree of the Life Website,
University of Arizona



Orangutan

Gorilla

Chimpanzee



Human

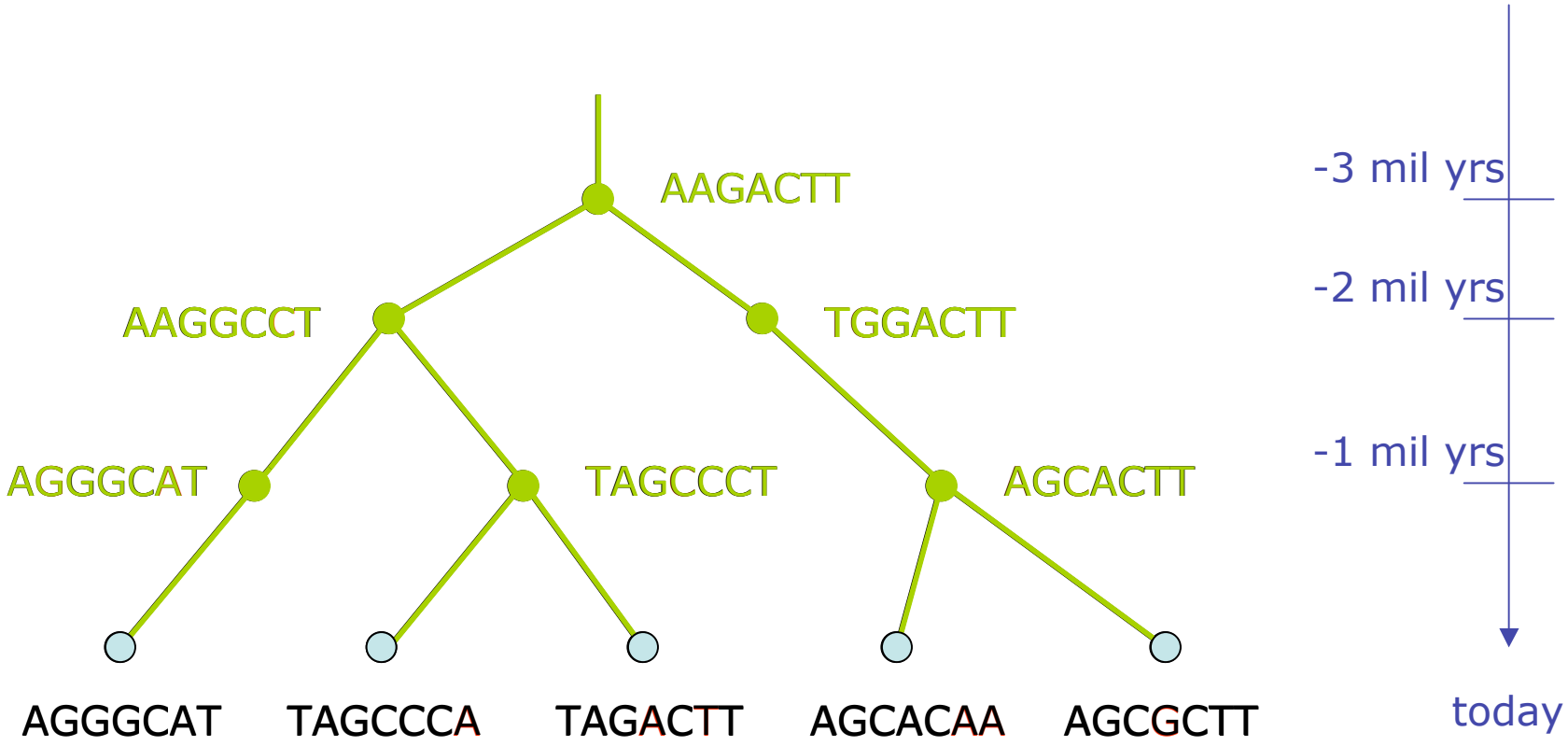
How did life evolve on earth?

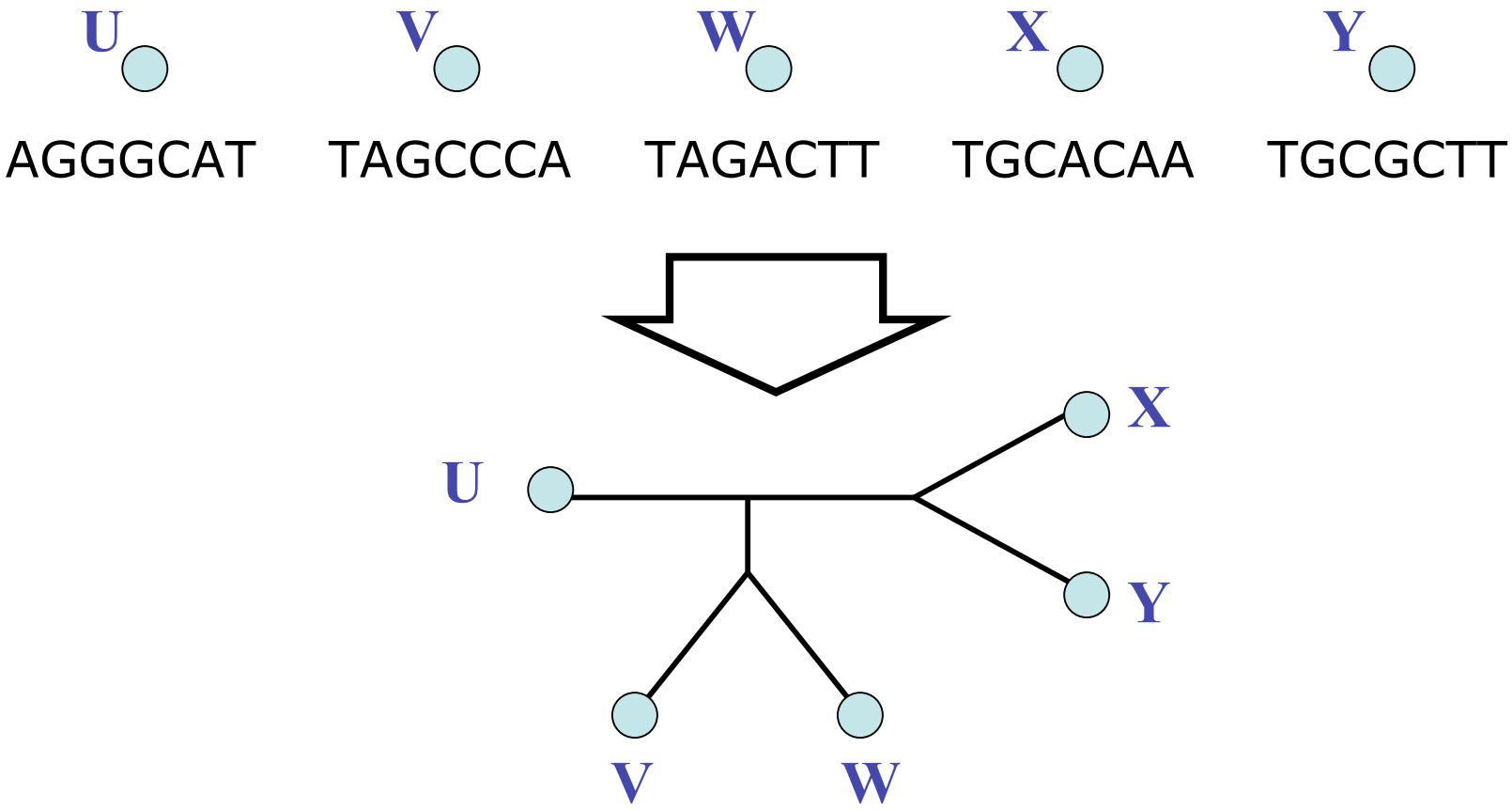
- Courtesy of the Tree of Life project

An international effort to understand how life evolved on earth

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

DNA Sequence Evolution




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

Standard Markov models

- 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)
- Numerical parameters describe the probability of substitutions of each type on each edge of the tree

Maximum Likelihood (ML)

- Given: Set S of aligned DNA sequences, and a parametric model of sequence evolution
- Objective: Find tree T and numerical parameter values (e.g, substitution probabilities) so as to maximize the probability of the data.

NP-hard

Statistically consistent for standard models if solved exactly

But solving this problem exactly is … unlikely

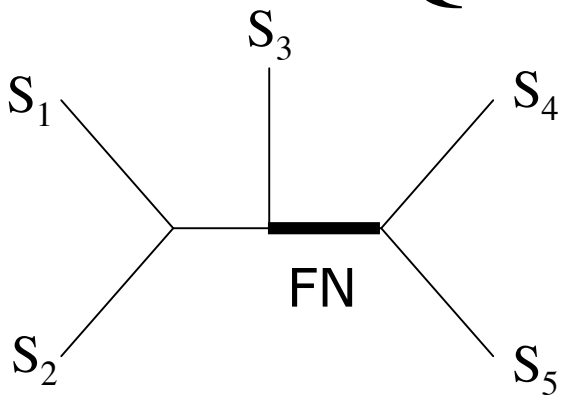
# of Taxa	# of Unrooted Trees
4	3
5	15
6	105
7	945
8	10395
9	135135
10	2027025
20	2.2×10^{20}
100	4.5×10^{190}
1000	2.7×10^{2900}

Fast ML heuristics

- RAxML (Stamatakis) with bootstrapping
- GARLI (Zwickl)
- Rec-I-DCM3 boosting (Roshan et al.) of RAxML to allow analyses of datasets with thousands of sequences

All available on the CIPRES portal (http://www.phylo.org)

Quantifying Error



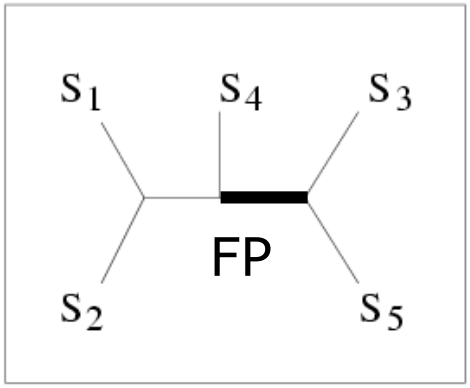
TRUE TREE

S_1	ACAATTAGAAC
S_2	ACCCTTAGAAC
S_3	ACCATTCCAAC
S_4	ACCAGACCAAC
S_5	ACCAGACCGGA

DNA SEQUENCES

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

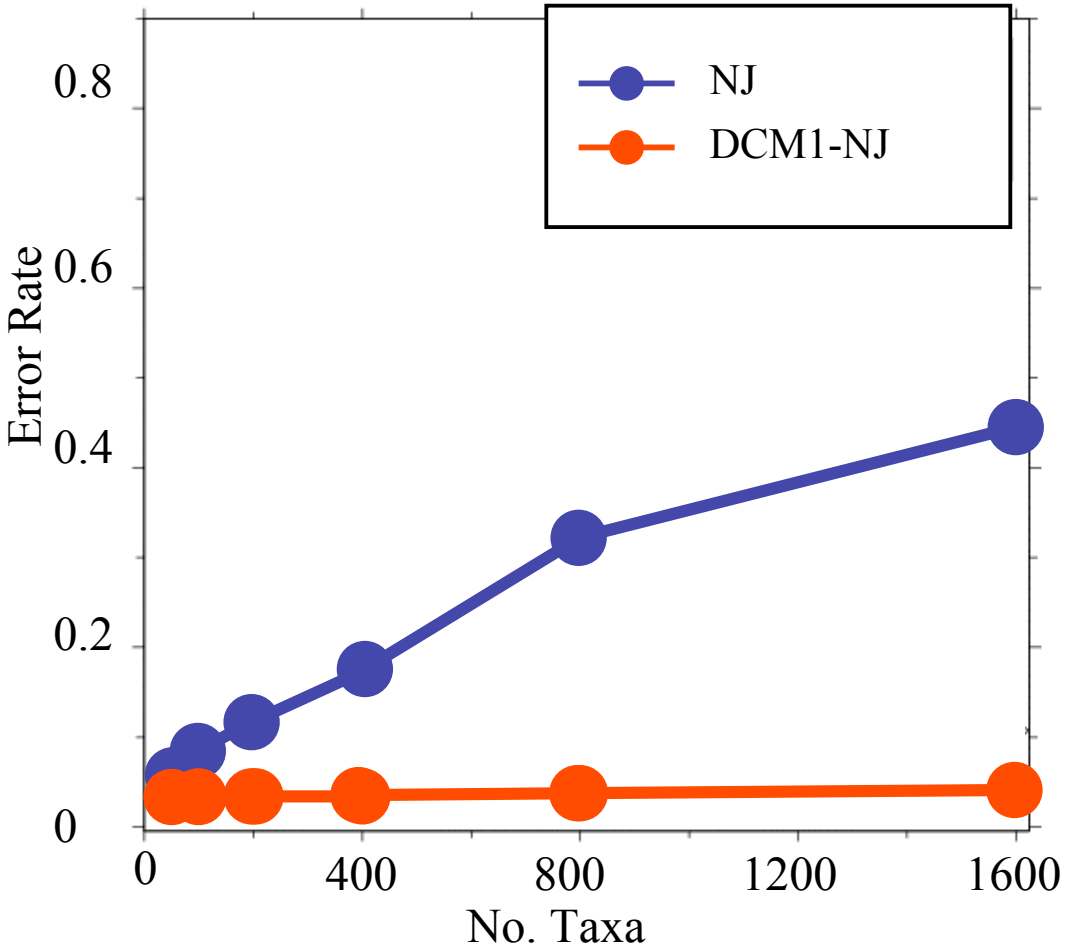
50% error rate



INFERRED TREE

DCM1-boosting distance-based methods

[Nakhleh et al. ISMB 2001]



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

But…

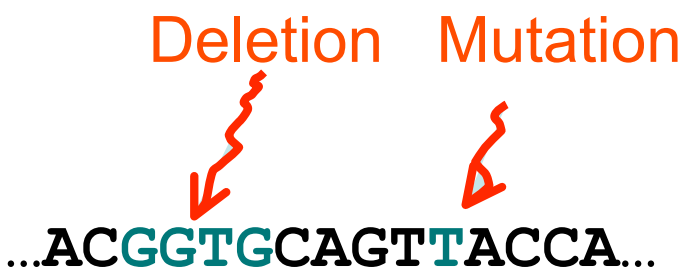
- Evolution is more complicated than these simple models:
  - Insertions and deletions (indels)
  - Duplications, inversions, transpositions (genome rearrangements)
  - Horizontal gene transfer and hybridization (reticulate evolution)
  - Etc.

Indels and substitutions at the DNA level

...**ACGGTGCAGTTACCA**...

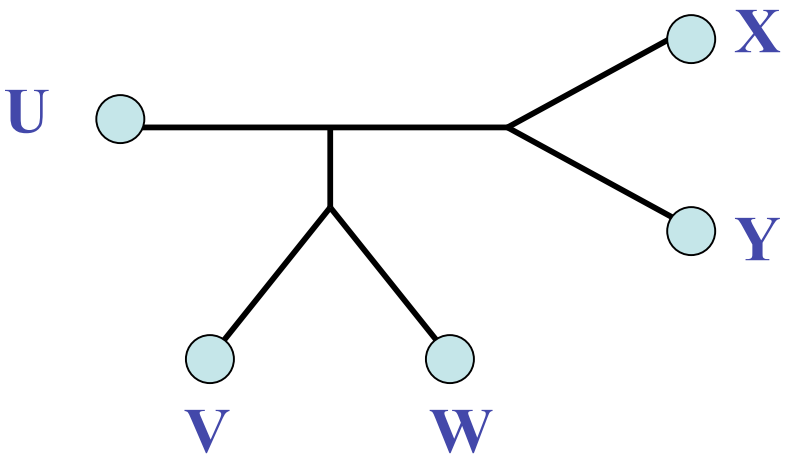
Indels and substitutions at the DNA level

Indels and substitutions at the DNA level



...ACCAGTCACCA...

U	AGTGGAT
V	TATGCCCA
W	TATGACTT
X	AGCCCTA
Y	AGCCCGCTT

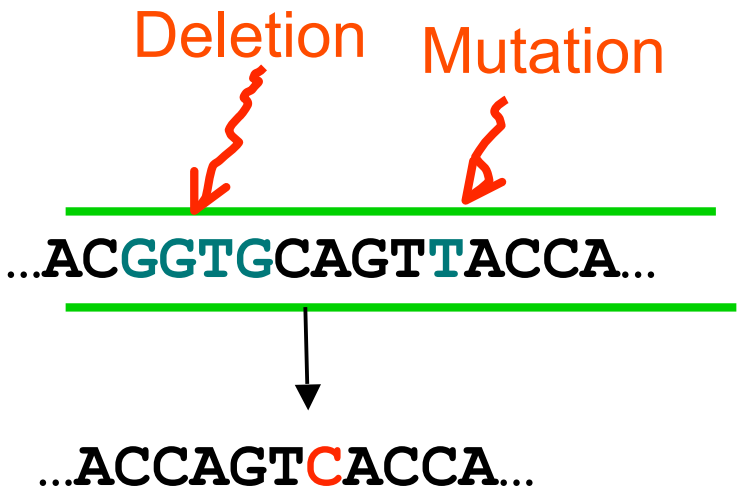


- Phylogenetic reconstruction methods assume the sequences all have the same length.
- Standard models of sequence evolution used in maximum likelihood and Bayesian analyses assume sequences evolve only via substitutions, producing sequences of equal length.
- And yet, almost all nucleotide datasets evolve with insertions and deletions ("indels"), producing datasets that violate these models and methods.

How can we reconstruct phylogenies from sequences of unequal length?

Roadmap for Today

- How it's currently done
- How it might be done
- How we're doing it (and how well)
- Where we're going with it



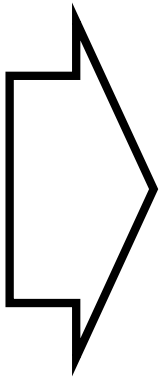
The true pairwise alignment is:

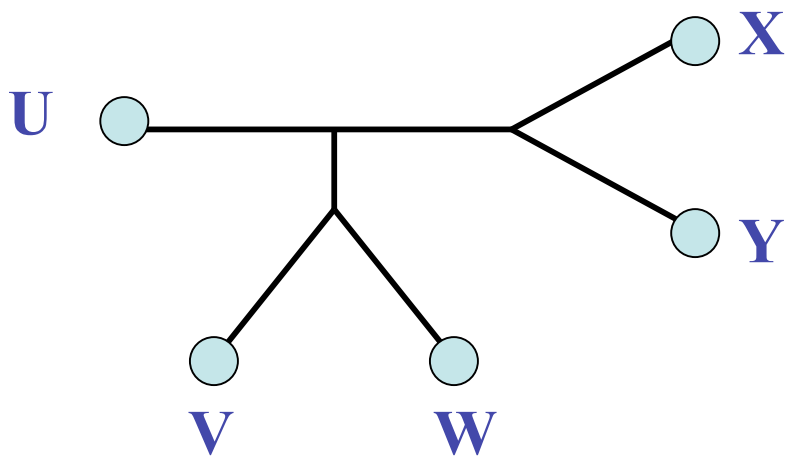
...ACGGTGCAGTTACCA...

...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.

U	AGTGGAT
V	TATGCCCA
W	TATGACTT
X	AGCCCTA
Y	AGCCCGCTT





Input: unaligned sequences

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

Phase 1: Multiple Sequence Alignment

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



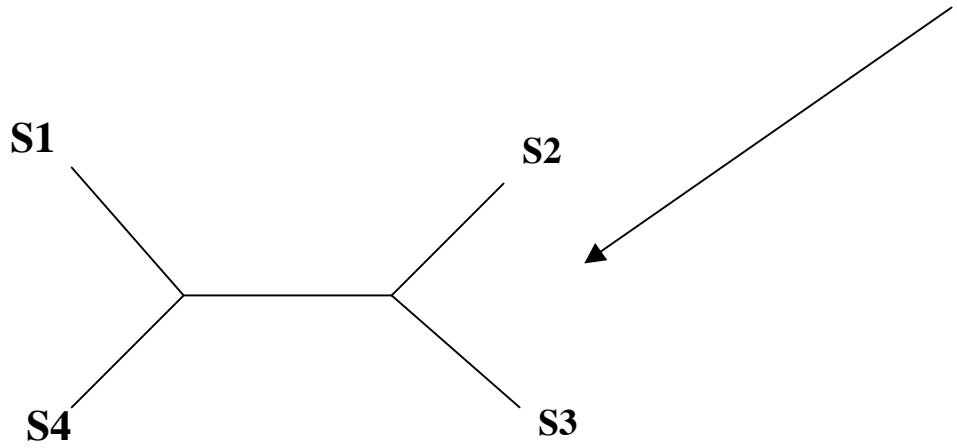
```
S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAC--GACCGC--
S3 = TAG-CT-------GACCGC--
S4 = -------TCAC--GACCGACA
```

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
```



So many methods!!!

Alignment method

- Clustal
- POY (and POY*)
- Probcons (and Probtree)
- MAFFT
- Prank
- Muscle
- Di-align
- T-Coffee
- Satchmo
- Etc.

Phylogeny method

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

So many methods!!!

Alignment method

- Clustal
- POY (and POY*)
- Probcons (and Probtree)
- MAFFT
- Prank
- Muscle
- Di-align
- T-Coffee
- Satchmo
- Etc.

Phylogeny method

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

So many methods!!!

Alignment method

- Clustal
- POY (and POY*)
- Probcons (and Probtree)
- MAFFT
- Prank
- Muscle
- Di-align
- T-Coffee
- Satchmo
- Etc.

Blue = used by systematists

Purple = recommended by Edgar and Batzoglou for protein alignments

Phylogeny method

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

Basic Questions

- Does improving the alignment lead to an improved phylogeny?
- Are we getting good enough alignments from MSA methods? (In particular, is ClustalW - the usual method used by systematists - good enough?)
- Are we getting good enough trees from the phylogeny reconstruction methods?
- Can we improve these estimations, perhaps through **simultaneous estimation** of trees and alignments?

Easy Sequence Alignment

```
B_WEAU160   ATGGAAAACAGATGGCAGGTGATGATTGTGTGGCAAGTAGACAGG 45
A_U455      ............................A....G......... 45
A_IFA86     ..................................G......... 45
A_92UG037   ..................................G......... 45
A_Q23       ..................C...............G......... 45
B_SF2       ....................................... 45
B_LAI       ....................................... 45
B_F12       ....................................... 45
B_HXB2R     ....................................... 45
B_LW123     ....................................... 45
B_NL43      ....................................... 45
B_NY5       ....................................... 45
B_MN        ...........C........................C..... 45
B_JRCSF     ....................................... 45
B_JRFL      ....................................... 45
B_NH52      ........................G.............. 45
B_OYI       ....................................... 45
B_CAM1      ....................................... 45
```

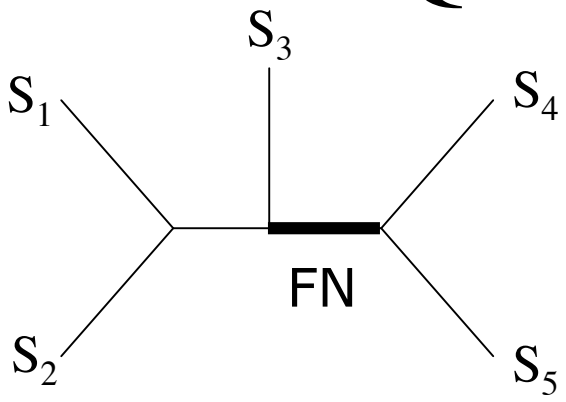
Harder Sequence Alignment

```
B_WEAU160          ATGAGAGTGAAGGGGATCAGGAAGAATTATCAGCACTTG     39
A_U455             .........T......ACA..G........CTTG....     39
A_SF1703           .........T......ACA..T...C.G...AA....A     39
A_92RW020.5            ......G......ACA..C..G..GG..AA.....    35
A_92UG031.7            ......G.A....ACA..G.....GG........A    35
A_92UG037.8            ......T......AGA..G........CTTG..G.    35
A_TZ017            .........G..A...G.A..G............A..A     39
A_UG275A           ....A..C..T.....CACA..T.....G...AA...G.    39
A_UG273A           ................ACA..G.....GG.........     39
A_DJ258A           .........T......ACA...........CA.T...A     39
A_KENYA            .........T.....CACA..G.....G.........A     39
A_CARGAN           .........T......ACA............A......     39
A_CARSAS           ..............CACA.........CTCT.C....     39
A_CAR4054          ............A..CACA..G.....GG..CA.....     39
A_CAR286A          ..............CACA..G.....GG..AA.....     39
A_CAR4023          ............A.---------..A............     30
A_CAR423A          ............A.---------..A............     30
A_VI191A           ..............ACA..T.....GG..A......     39
```

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 varied 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”.
- We estimated maximum likelihood trees (using RAxML) on various alignments (including the true alignment).
- We evaluated estimated trees for topological accuracy using the Missing Edge rate.

Quantifying Error



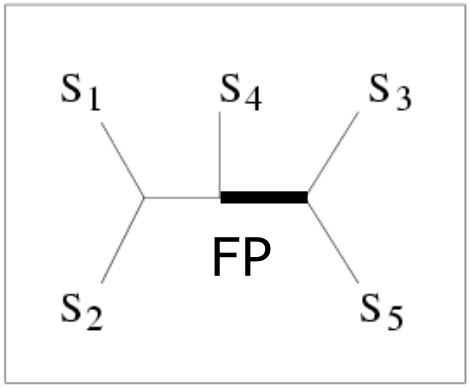
TRUE TREE

S_1	ACAATTAGAAC
S_2	ACCCTTAGAAC
S_3	ACCATTCCAAC
S_4	ACCAGACCAAC
S_5	ACCAGACCGGA

DNA SEQUENCES

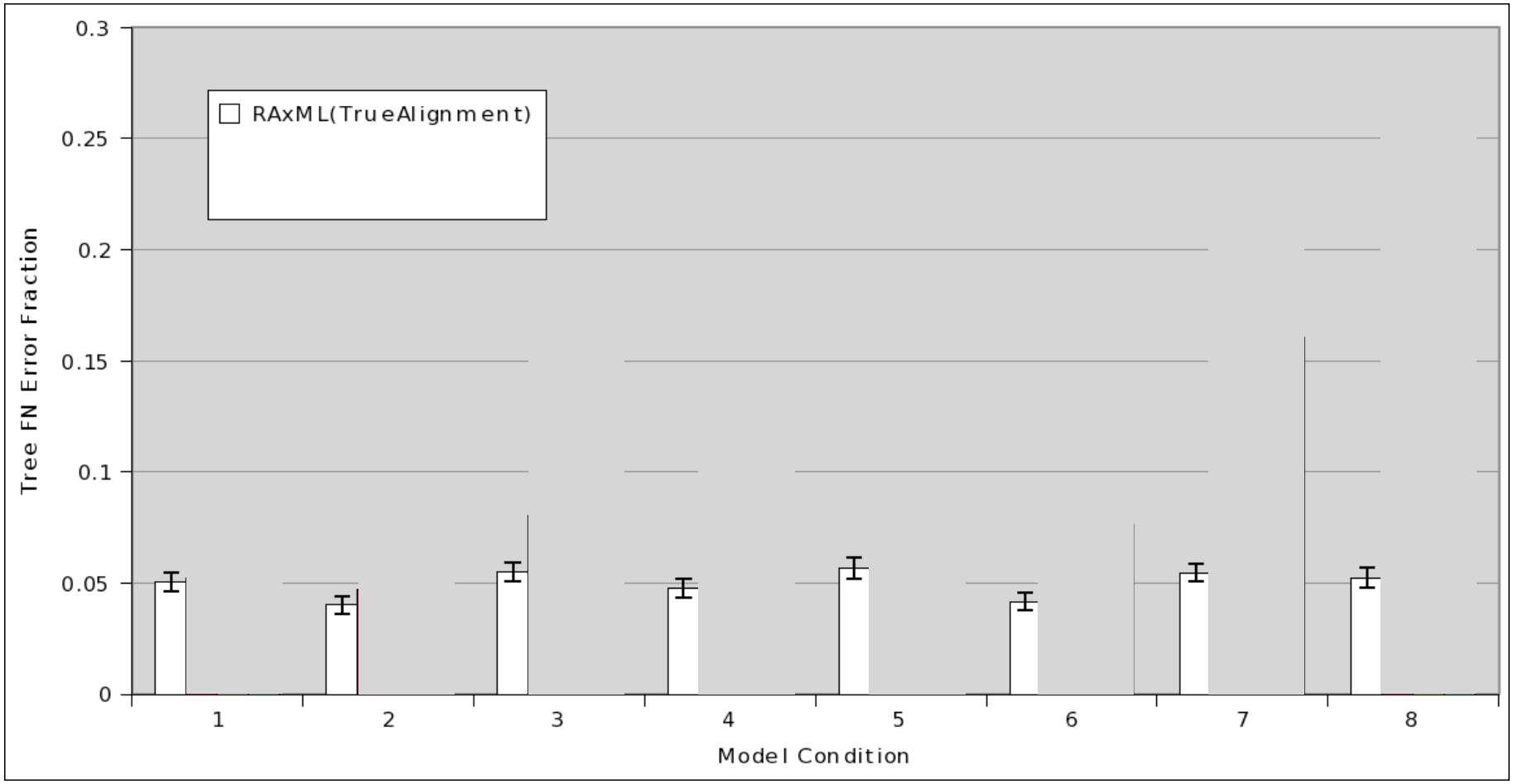
FN: false negative
(missing edge)
FP: false positive
(incorrect edge)

50% error rate



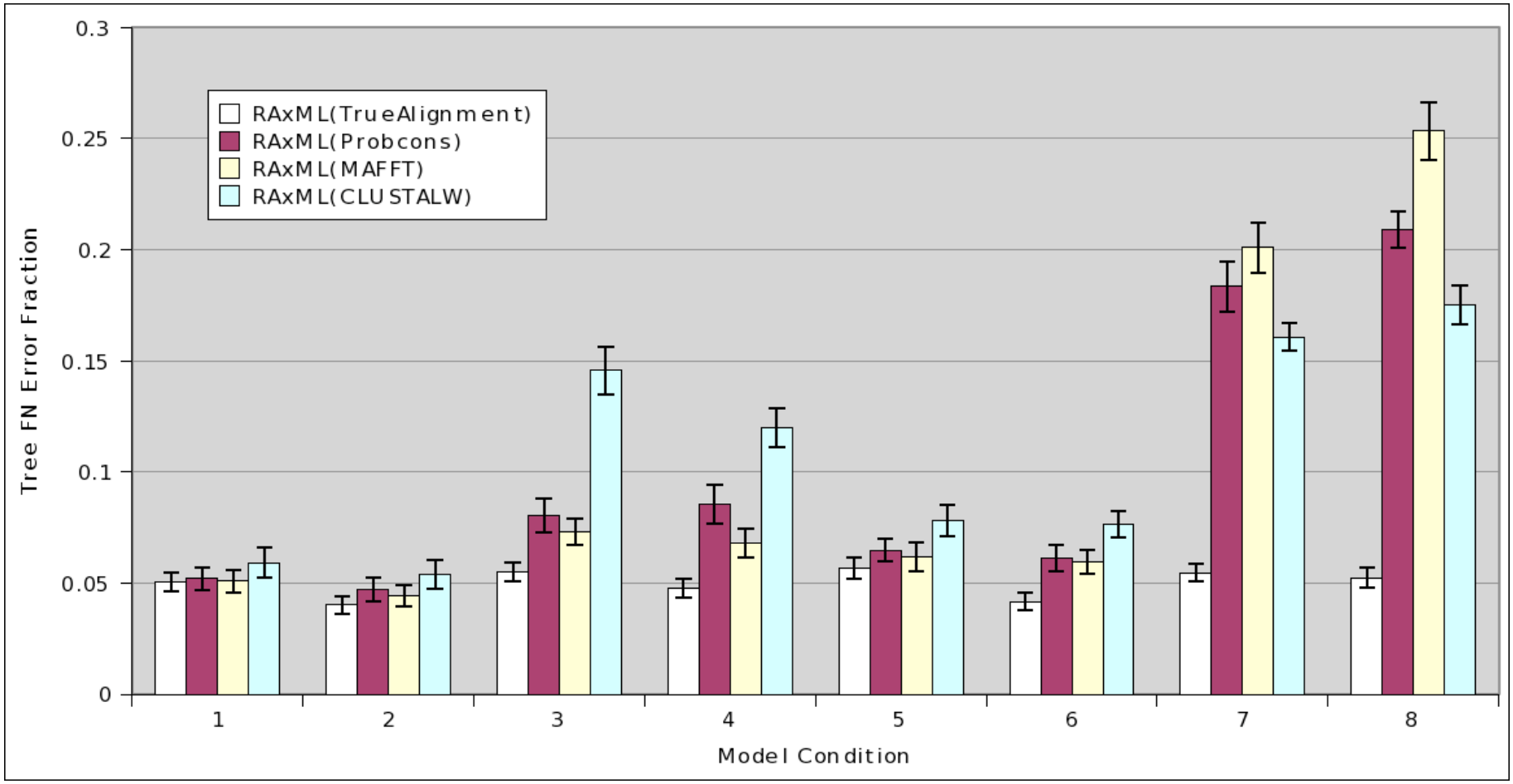
INFERRED TREE

DNA sequence evolution



Simulation using ROSE: 100 taxon model trees, models 1-4 have “long gaps”, and 5-8 have “short gaps”, site substitution is HKY+Gamma

DNA sequence evolution

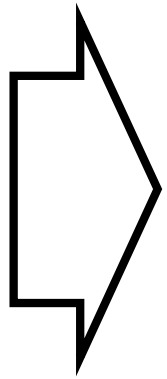


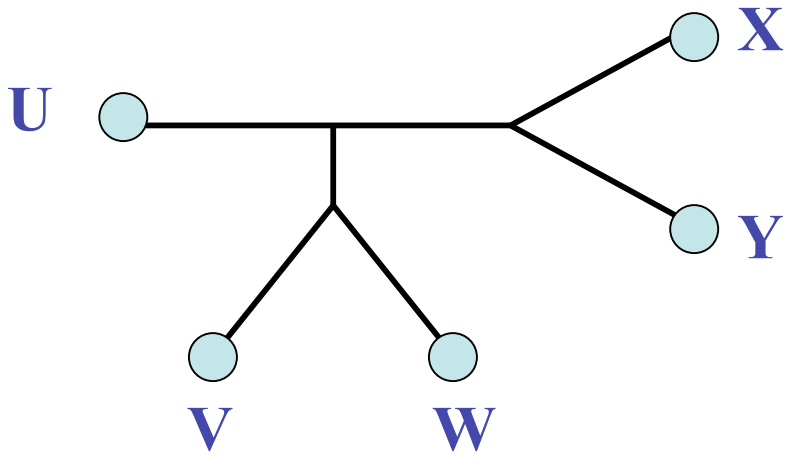
Simulation using ROSE: 100 taxon model trees, models 1-4 have “long gaps”, and 5-8 have “short gaps”, site substitution is HKY+Gamma

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).

U AGTGGAT
V TATGCCCA
W TATGACTT
X AGCCCTA
Y AGCCCGCTT





What about "simultaneous estimation"?

Simultaneous Estimation

- Statistical methods (e.g., AliFritz and BaliPhy) take a long time to converge (limited possibly to small datasets?)
- POY attempts to solve the NP-hard “minimum treelength” problem, and can be applied to larger datasets.
  - Somewhat equivalent to maximum parsimony
  - Sensitive to gap treatment, but even with very good gap treatments is only comparable to good two-phase methods in accuracy (while not as accurate as the better ones), and takes a long time to reach local optima

What we'd like (ideally)

- An automated means of practically inferring alignments and very large phylogenetic trees using sequence (DNA, protein) data
  - Very large means at least thousands, but as many as tens of thousands of taxa
  - Preferably able to run on a desktop computer
  - Validated on both real and simulated data

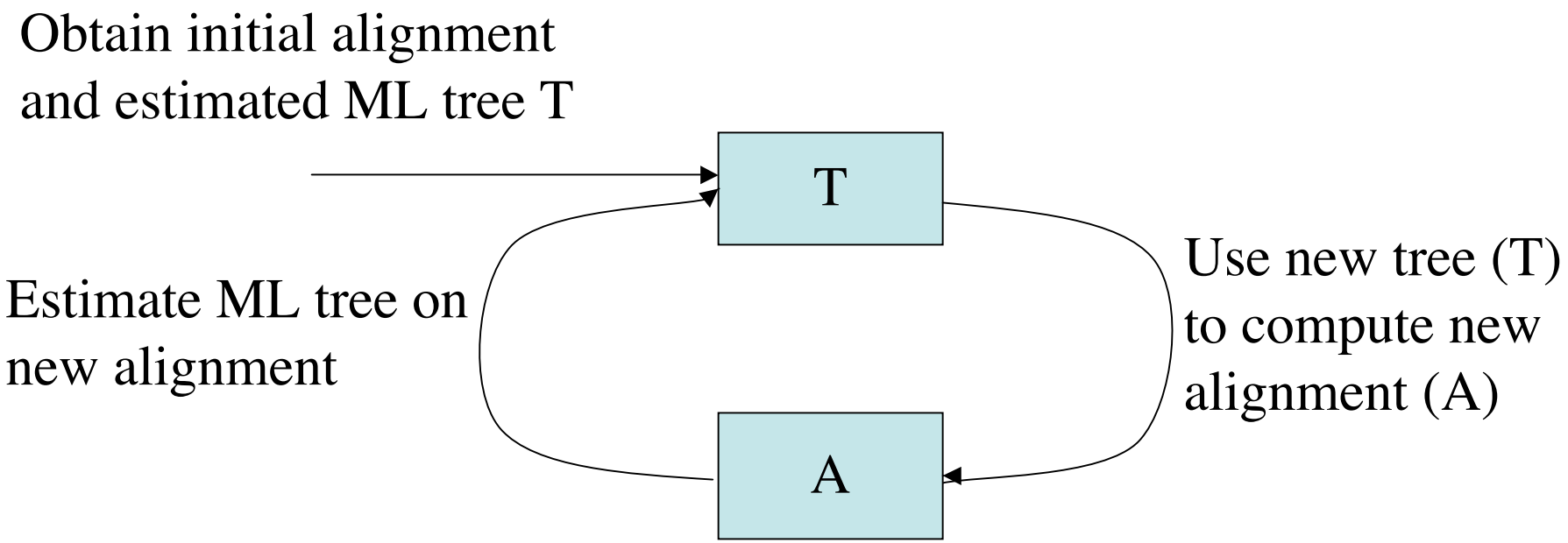
SATé:
(Simultaneous Alignment and Tree Estimation)

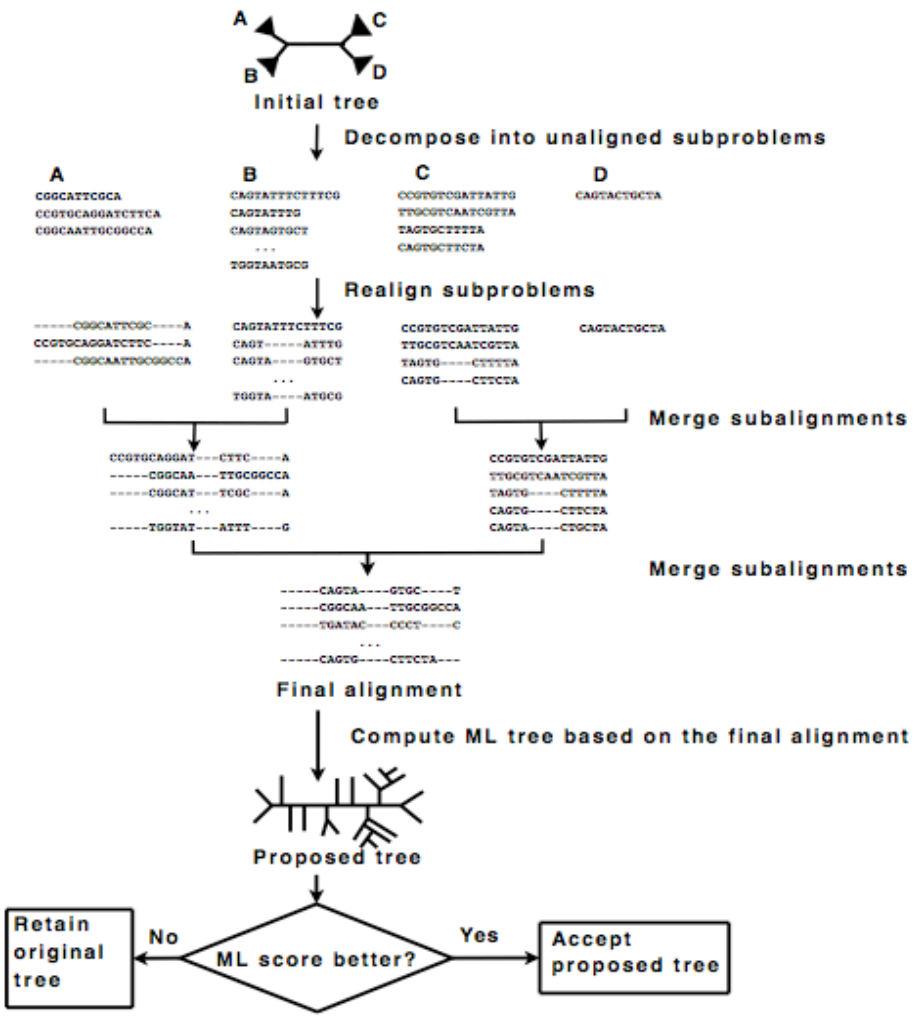
- Developers: Liu, Nelesen, Raghavan, Linder, and Warnow
- Search strategy: search through tree space, and ***realigns sequences on each tree using a novel divide-and-conquer approach***.
- Optimization criterion: alignment/tree pair that optimizes maximum likelihood under GTR+Gamma (RAxML GTRMIX).

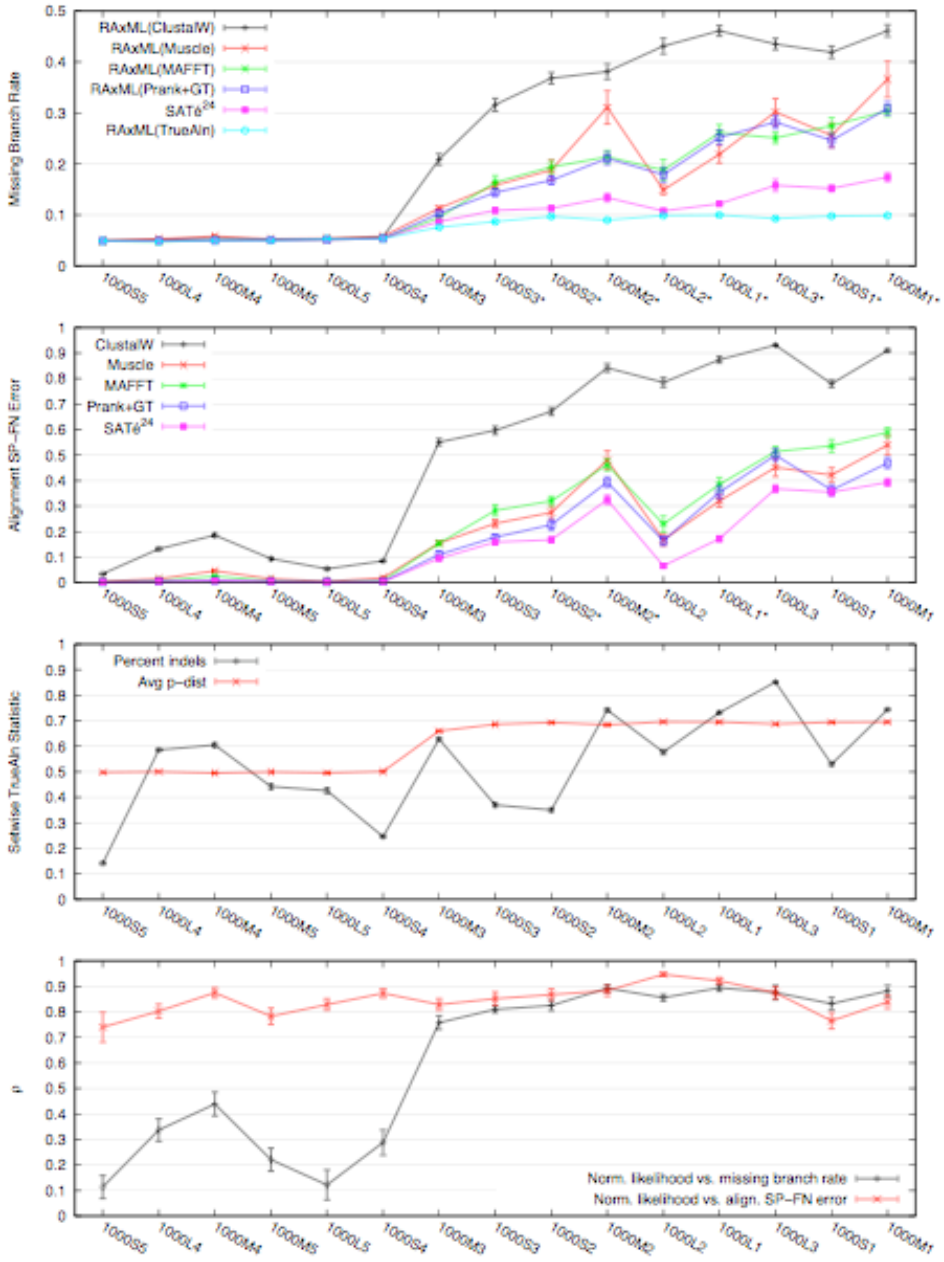
- Submitted

SATé Algorithm (unpublished)

SATé keeps track of the maximum likelihood scores of the tree/alignment pairs it generates, and returns the best pair it finds




A
C
B
D
Initial tree
Decompose into unaligned subproblems
A
B
C
D
Realign subproblems
Merge subalignments
Merge subalignments
Final alignment
Compute ML tree based on the final alignment
Proposed tree
Retain original tree
No
ML score better?
Yes
Accept proposed tree

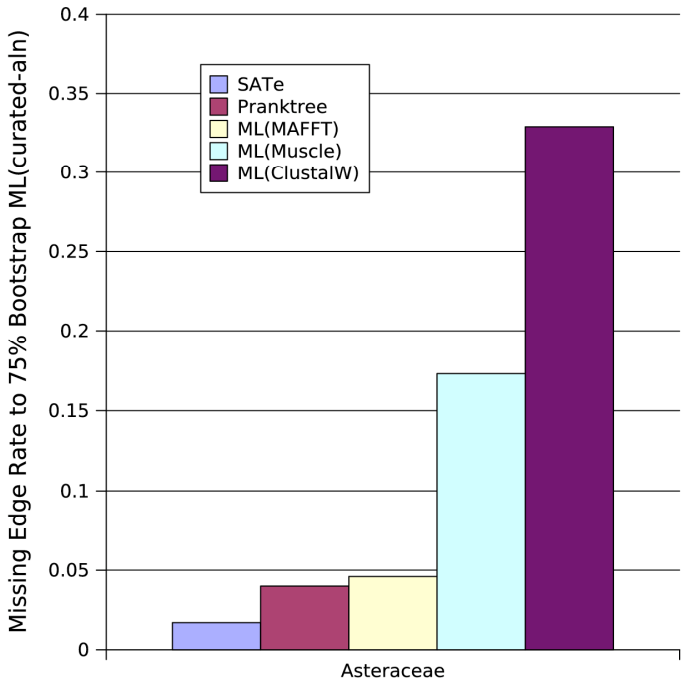

Missing Branch Rate
RAxML(ClustalW)
RAxML(Muscle)
RAxML(MAFFT)
RAxML(Prank+GT)
SATé24
RAxML(TrueAln)
Alignment SP−FN Error
ClustalW
Muscle
MAFFT
Prank+GT
SATé24
Selwise TrueAln Statistic
Percent indels
Avg p−dist
ρ
Norm. likelihood vs. missing branch rate
Norm. likelihood vs. align. SP−FN error
1000S5
1000L4
1000M4
1000M5
1000L5
1000S4
1000M3
1000S3*
1000S2*
1000M2*
1000L2*
1000L1*
1000L3*
1000S1*
1000M1*
1000S5
1000L4
1000M4
1000M5
1000L5
1000S4
1000M3
1000S3
1000S2*
1000M2*
1000L2
1000L1*
1000L3
1000S1
1000M1
1000S5
1000L4
1000M4
1000M5
1000L5
1000S4
1000M3
1000S3
1000S2
1000M2
1000L2
1000L1
1000L3
1000S1
1000M1
1000S5
1000L4
1000M4
1000M5
1000L5
1000S4
1000M3
1000S3
1000S2
1000M2
1000L2
1000L1
1000L3
1000S1
1000M1

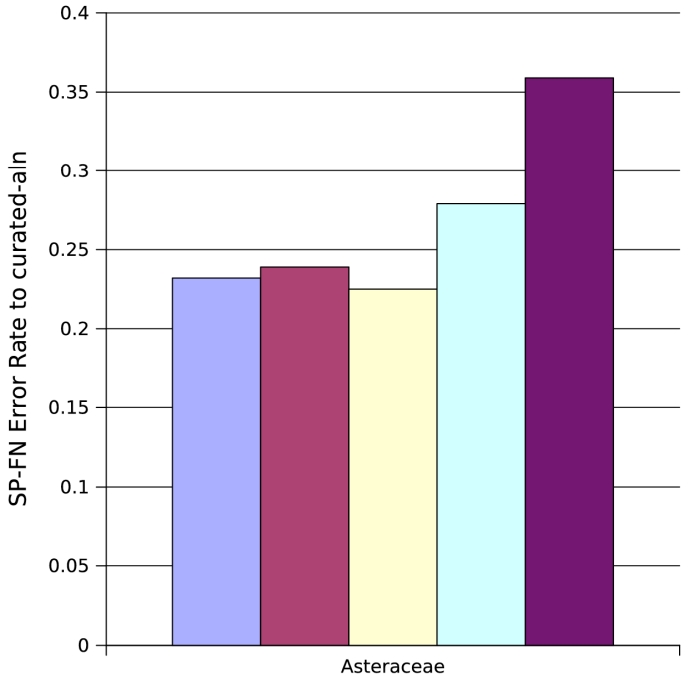
Biological datasets

- Used ML analyses of curated alignments (8 produced by Robin Gutell, others from the Early Bird ATOL project, and some from UT faculty)
- Computed several alignments and maximum likelihood trees on each alignment, and SATe trees and alignments.
- Compared alignments and trees to the curated alignment and to the reference tree (75% bootstrap ML tree on the curated alignment)

Asteraceae ITS

- The curated alignment consists of 328 ITS sequences drawn from the Asteraceae family (Goertzen et al 2003).
- Empirical statistics:
  - 36% ANHD
  - 79% MNHD
  - 23% gapped





Conclusions

- SATé produces trees and alignments that improve upon the best two-phase methods for “hard to align” datasets, and can do so in reasonable time frames (24 hours) on desktop computers
- Further improvement is likely with longer analyses
- Better results would likely be obtained by ML under models that include indel processes (ongoing work)

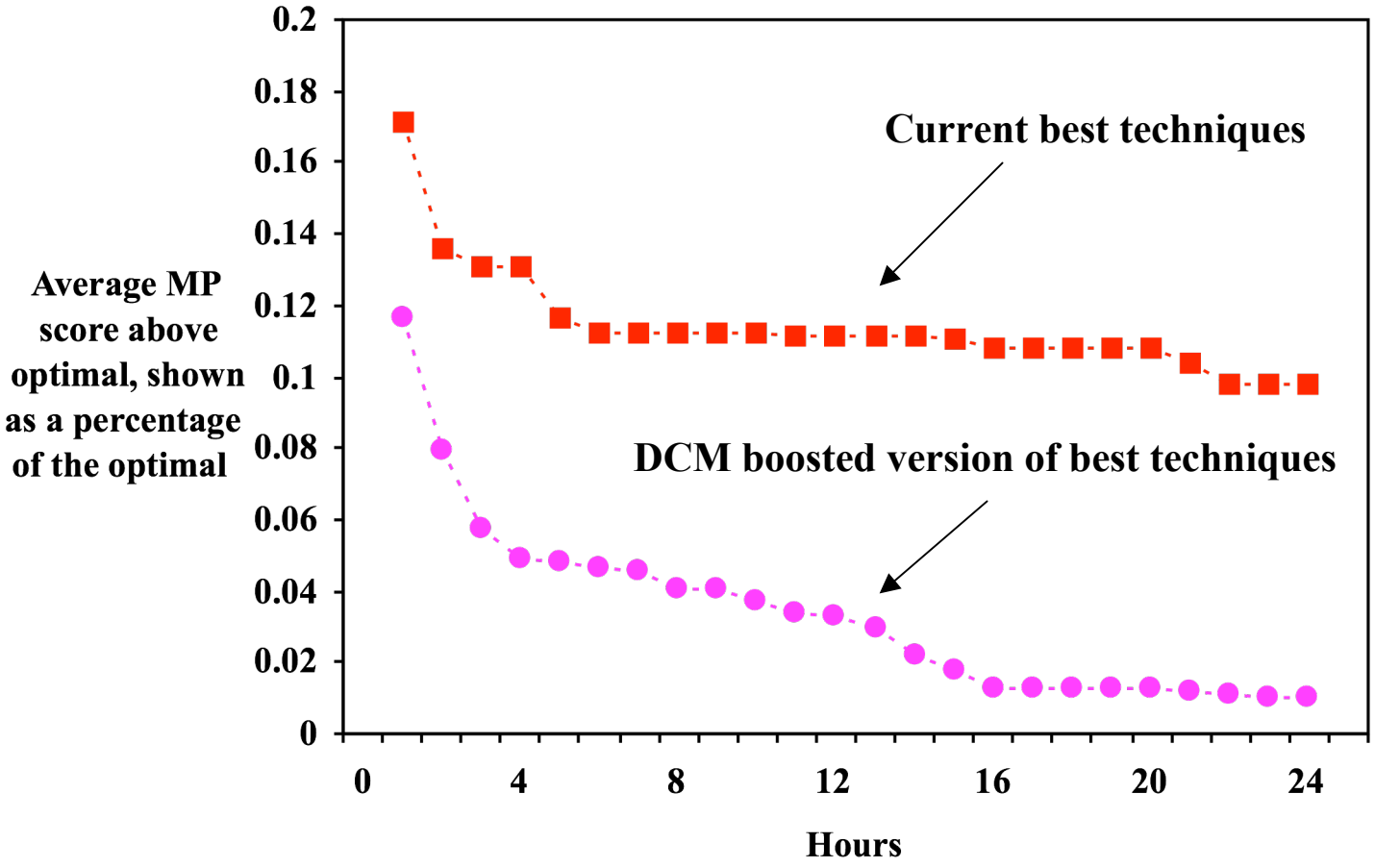
But…

- Evolution is more complicated than these simple models:
  - Insertions and deletions (indels)
  - Duplications, inversions, transpositions (genome rearrangements)
  - Horizontal gene transfer and hybridization (reticulate evolution)
  - Etc.

Acknowledgements

- Funding: NSF, The Program in Evolutionary Dynamics at Harvard, and The Institute for Cellular and Molecular Biology at UT-Austin.
- Collaborators:
  - Randy Linder (Integrative Biology, UT-Austin)
  - Students Kevin Liu, Serita Nelesen, and Sindhu Raghavan

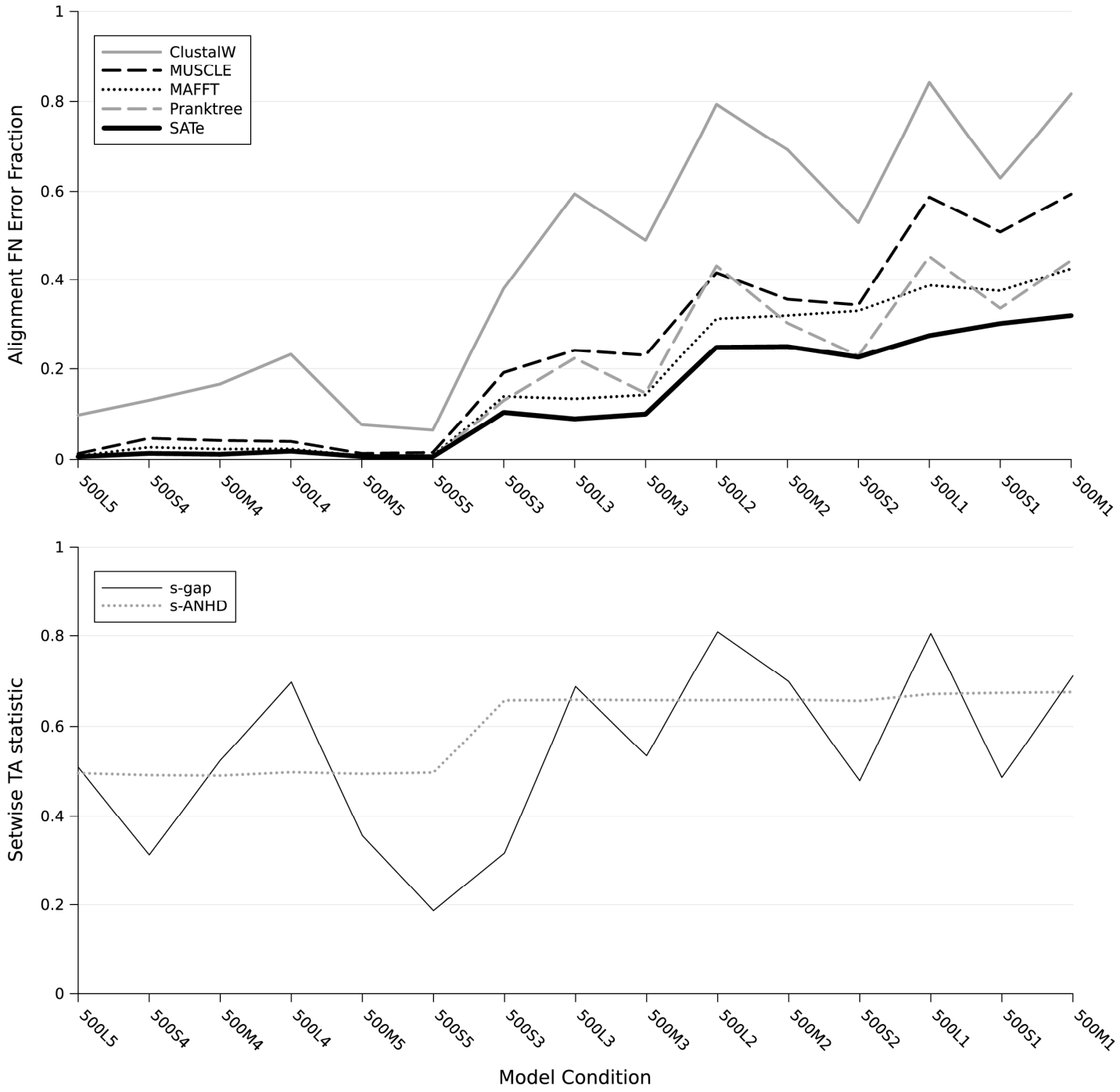
Rec-I-DCM3 significantly improves performance (Roshan et al. CSB 2004)



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

Similar improvements obtained for RAxML (maximum likelihood).

- Alignment (SP-FN) error rates on 500 taxon simulated datasets.
- Empirical statistics for the simulated data



- Missing edge rates on 500 taxon simulated datasets.

- Empirical statistics