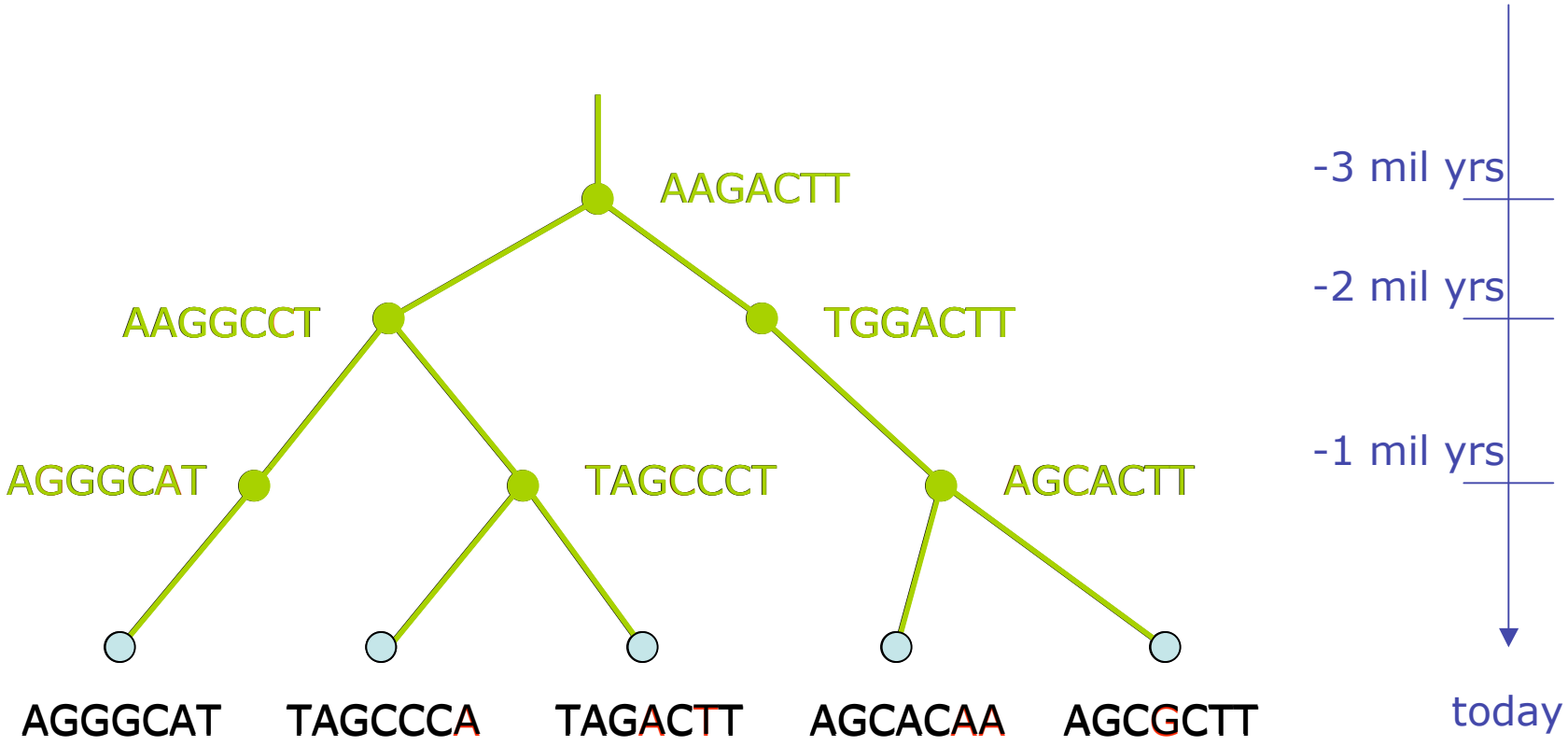
Simultaneous estimation of alignments and trees

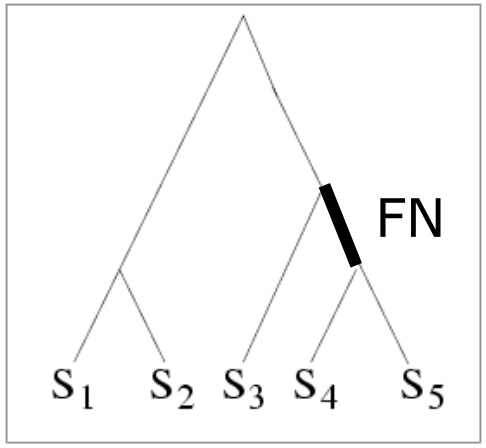
Tandy Warnow

The University of Texas at Austin

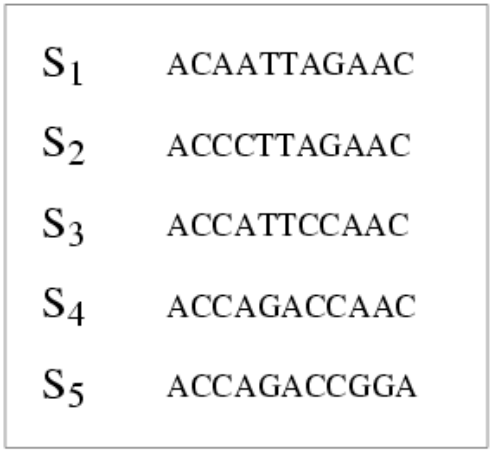
(joint work with Randy Linder, Kevin Liu, Serita Nelesen, and Sindhu Raghavan)

DNA Sequence Evolution





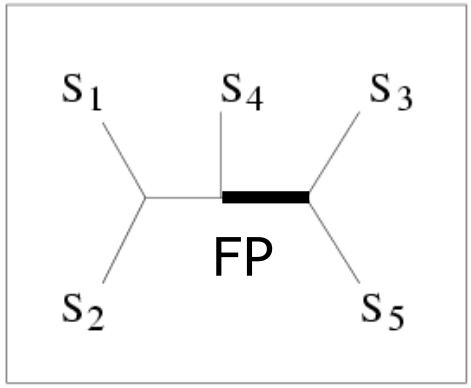
TRUE TREE



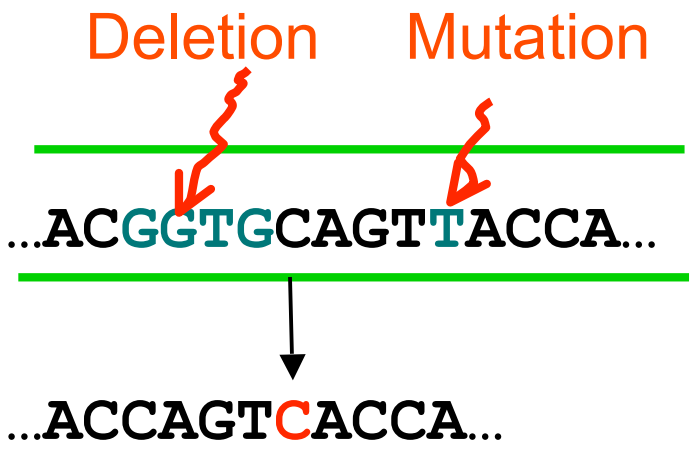
DNA SEQUENCES

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

50% error rate



INFERRED TREE



indels (insertions and deletions) also occur!

Input: unaligned sequences

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

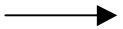
Phase 1: Multiple Sequence Alignment

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

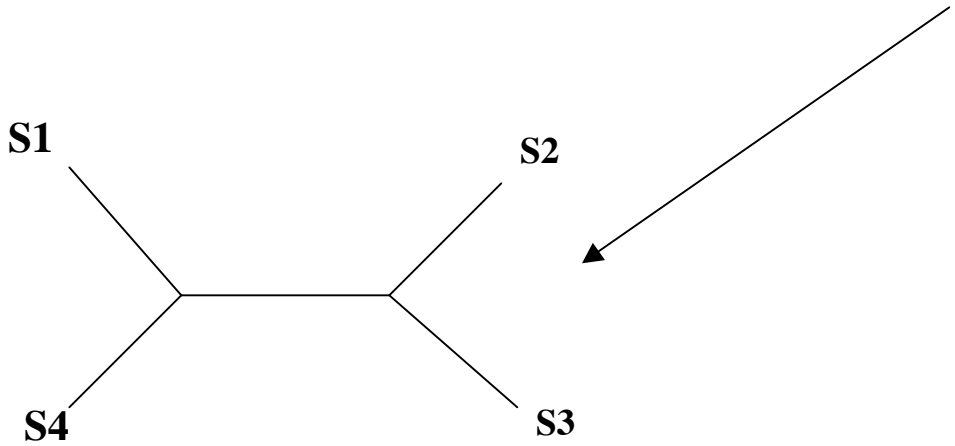
```
S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAC--GACCGC--
S3 = TAG-CT-------GACCGC--
S4 = -------TCAC--GACCGACA
```

Phase 2: Construct tree

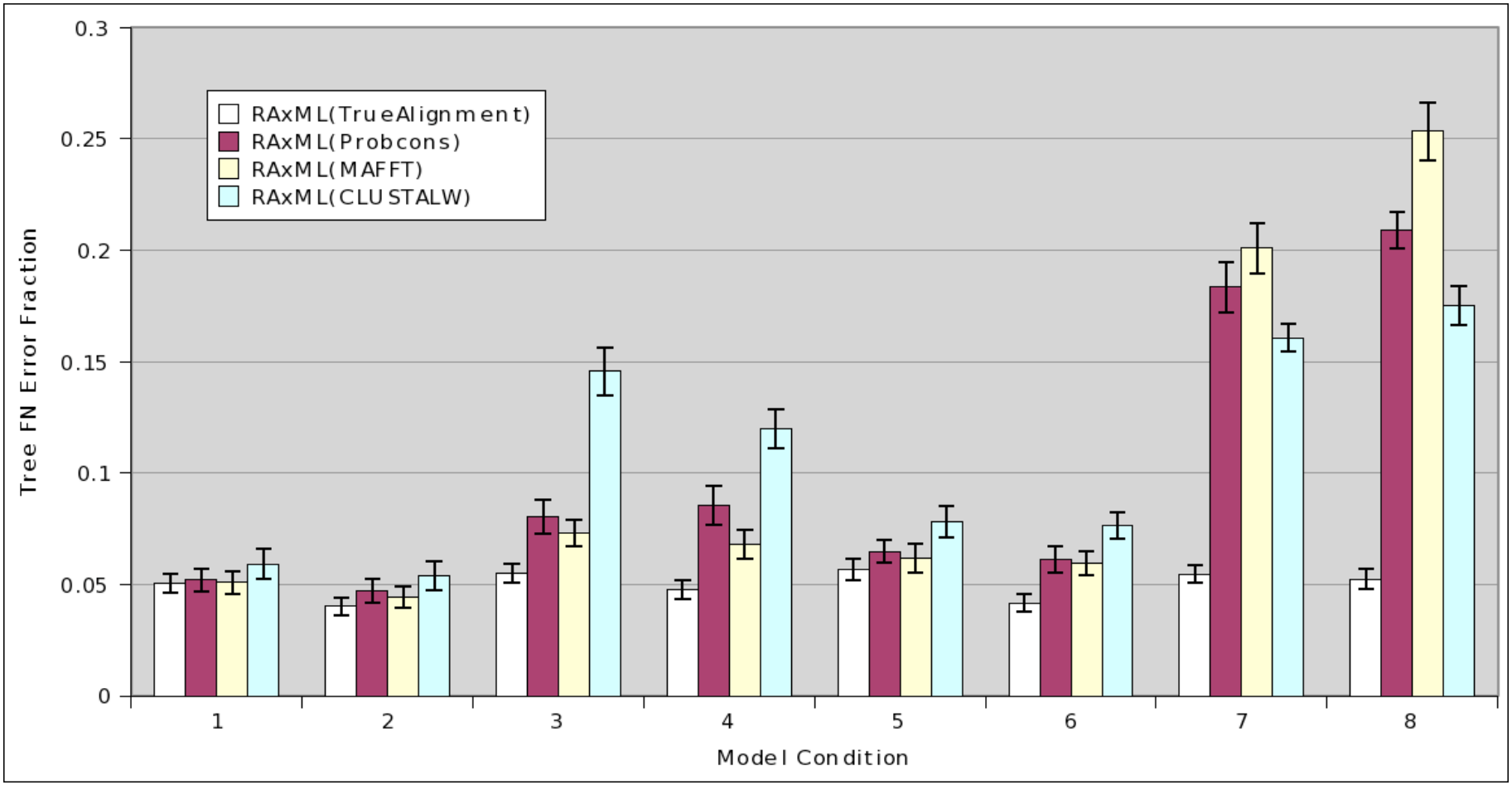
```
S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC
S3 = TAGCTGACCGC
S4 = TCACGACCGACA
```



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



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

Simultaneous estimation?

- Statistical methods (e.g., AliFritz and BaliPhy) cannot be applied to datasets above ~20 sequences.
- POY attempts to solve the NP-hard “minimum treelength” problem, and can be applied to larger datasets.

POY vs. Clustal

- Ogden and Rosenberg did a simulation study showing POY 3.0 alignments (using simple gap penalties) were less accurate than Clustal alignments on over 99% of the datasets they generated.
- Simple gap penalties are of the form gapcost(L)=cL for some constant c

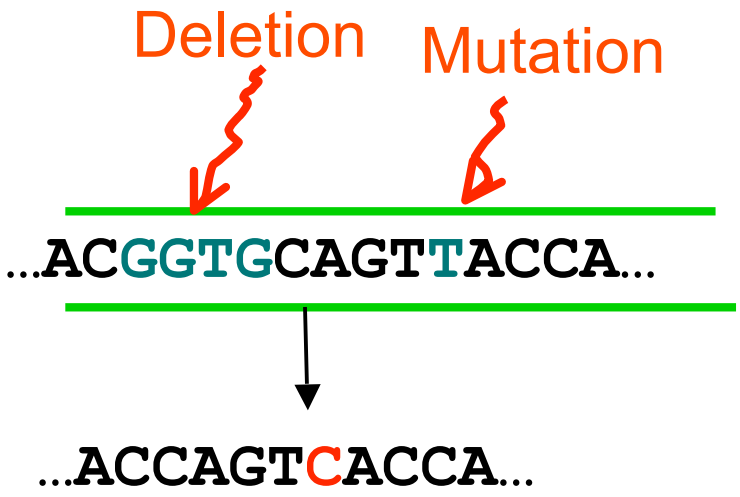
This talk

- **POY vs. Clustal**, and our response to Ogden and Rosenberg (to appear, IEEE Transactions on Computational Biology and Bioinformatics, Liu et al.)
- **SATé**: our work (in progress, unpublished) on statistical co-estimation of trees and alignments.

POY's optimization problem

- Given set S of sequences (not in an alignment) and an edit distance function
- Find tree T with leaves labelled by the sequences of S, and internal nodes labelled by other sequences, of minimum total edit distance.

NP-hard. (Even finding the best sequences for a fixed tree is NP-hard)



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.

Alignment Error (SP)

- A C A T - - - G C — True alignment
- C A A - G A T G C

- A C A T G - - - C — Est. alignment
- - C A A G A T G C

Alignment Error (SP)

- A C A T - - - G C True alignment
- C A A - G A T G C

- A C A T G - - - C Est. alignment
- - C A A G A T G C

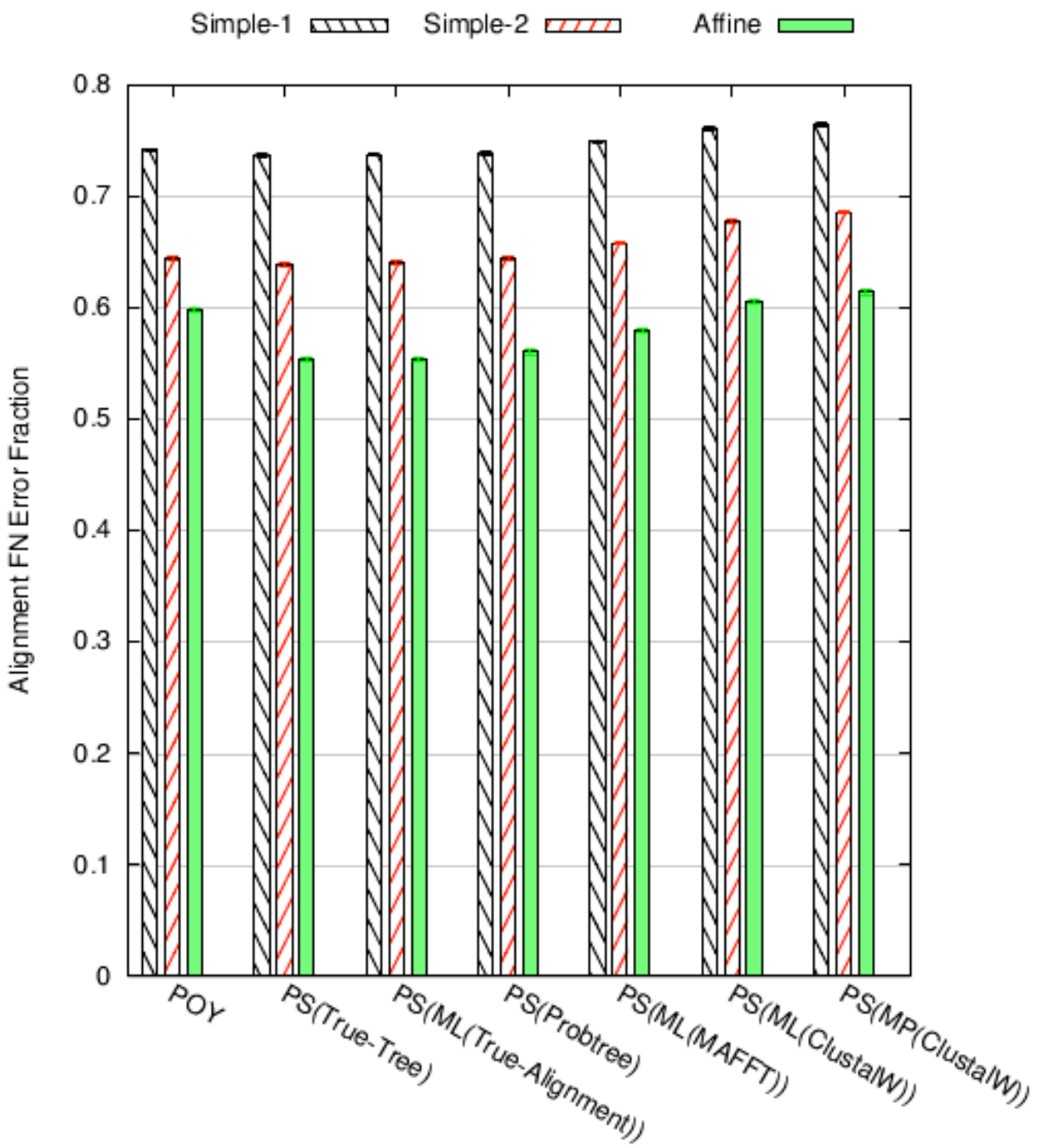
- Four of the five true homologies are missing! So the SP-error rate is 80%.

Gap penalty functions

- Simple 1: all indels and substitutions have the same cost
- Simple2: indels have cost 1, transitions cost 0.5, transversions cost 1
- Affine: gapcost(L)=2+L/2, transitions cost 0.5, transversions cost 1.

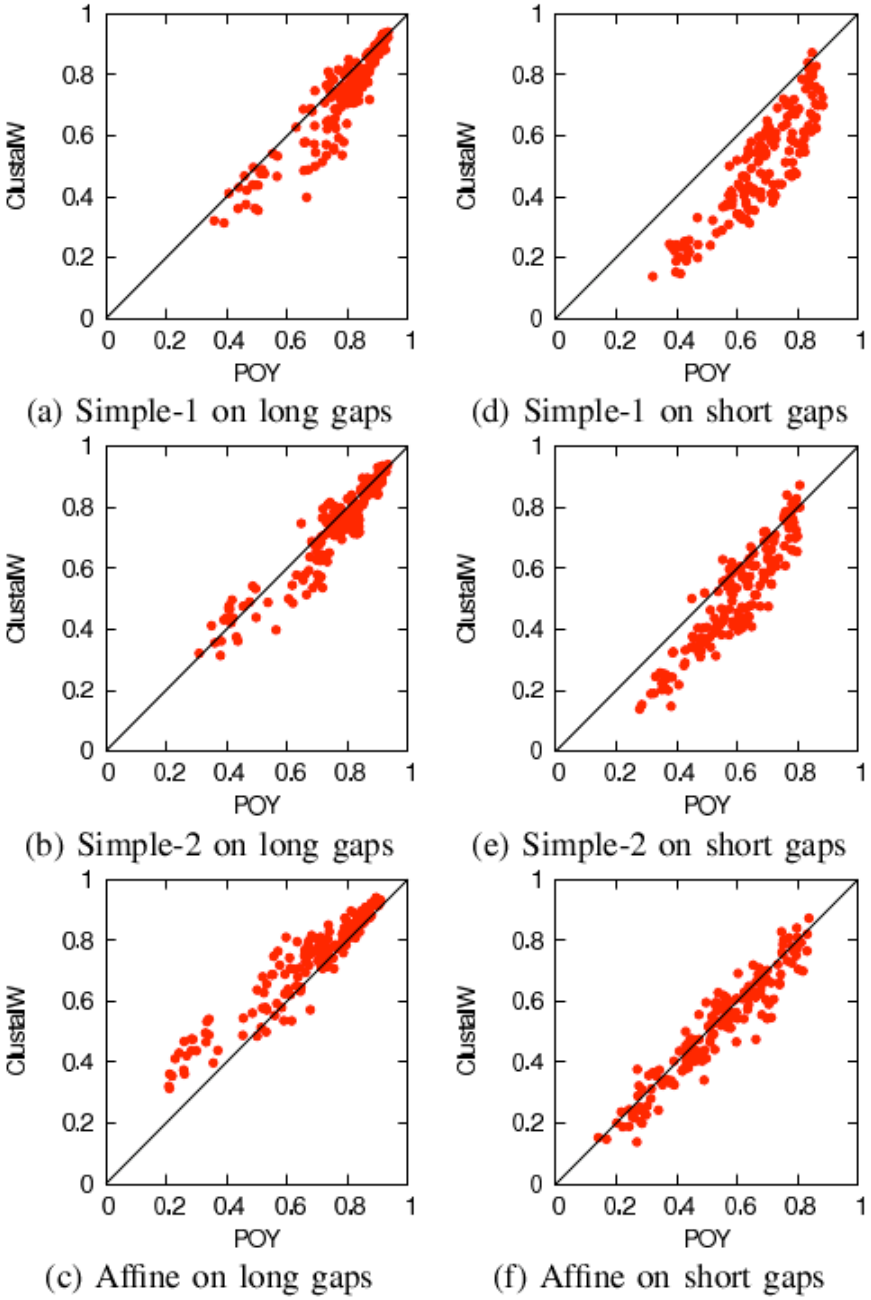
Results – Alignment Errors

- PS is POY-score (used to estimate alignments on various trees)



POY4.0 competitive with ClustalW when using affine gap penalties

- Points below the diagonal are for datasets on which POY4.0 is worse than ClustalW.
- Points above the diagonal are for datasets on which POY4.0 is better than ClustalW.

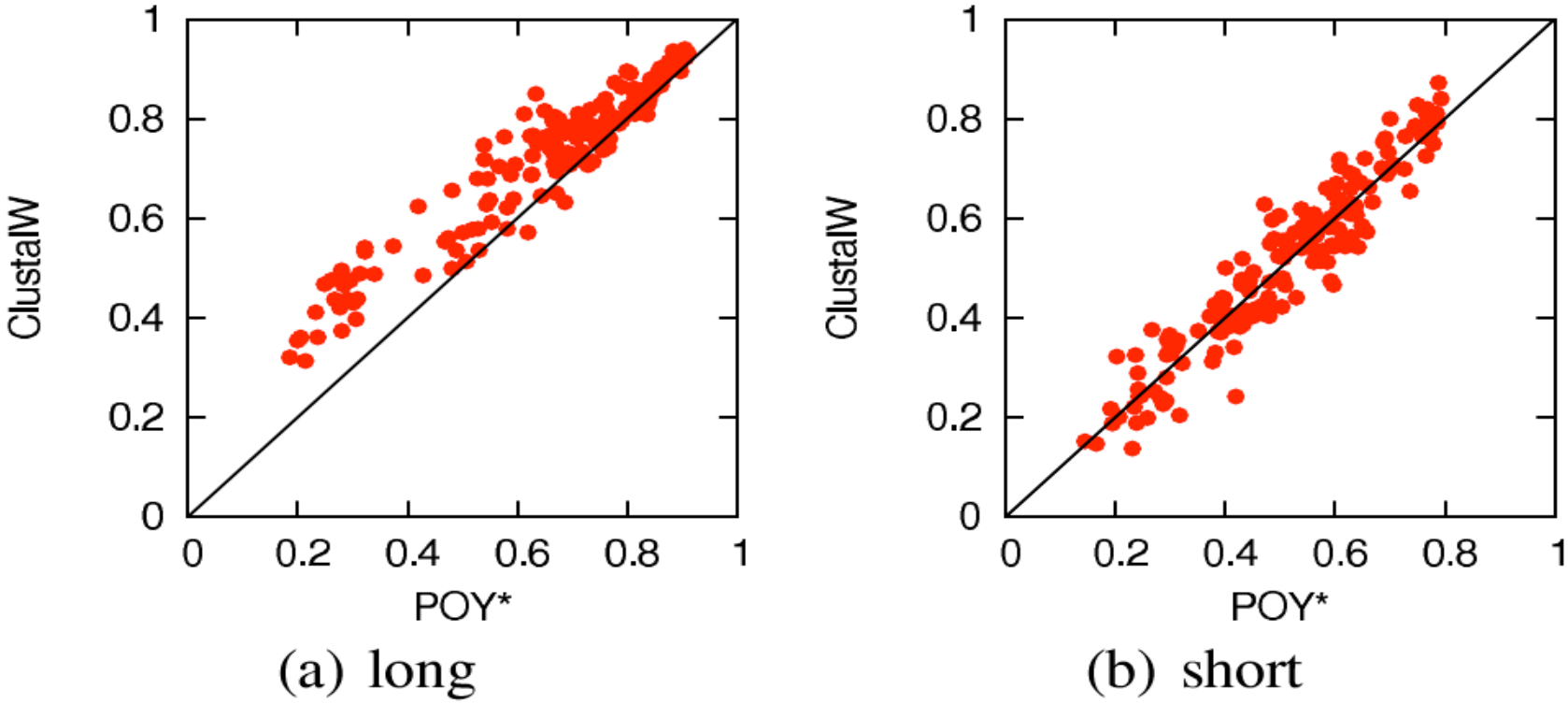


(a) Simple-1 on long gaps (d) Simple-1 on short gaps

(b) Simple-2 on long gaps (e) Simple-2 on short gaps

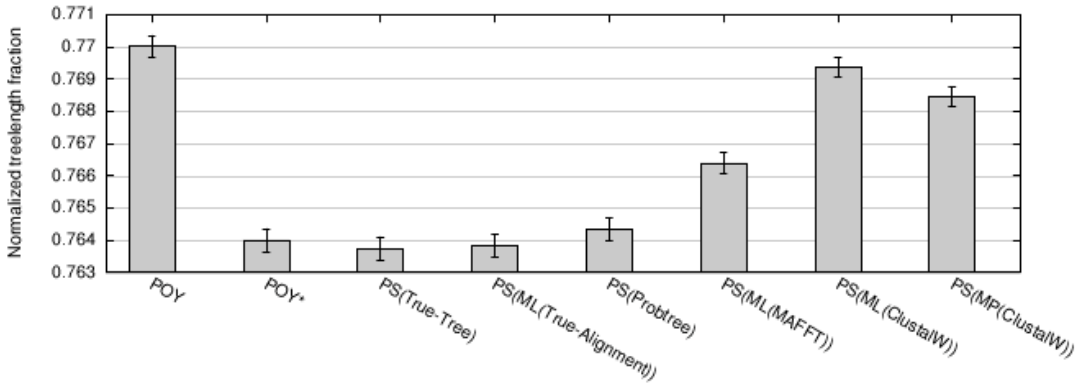
(c) Affine on long gaps (f) Affine on short gaps

Results – ClustalW vs. POY*

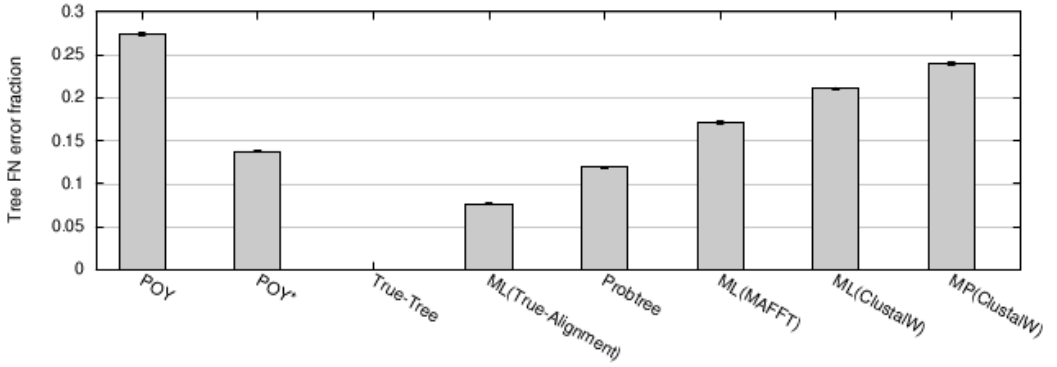


- POY* (our improvement to POY) is better than ClustalW on 90% of the datasets with short gaps (a), and over 50% of the datasets with long gaps (b)

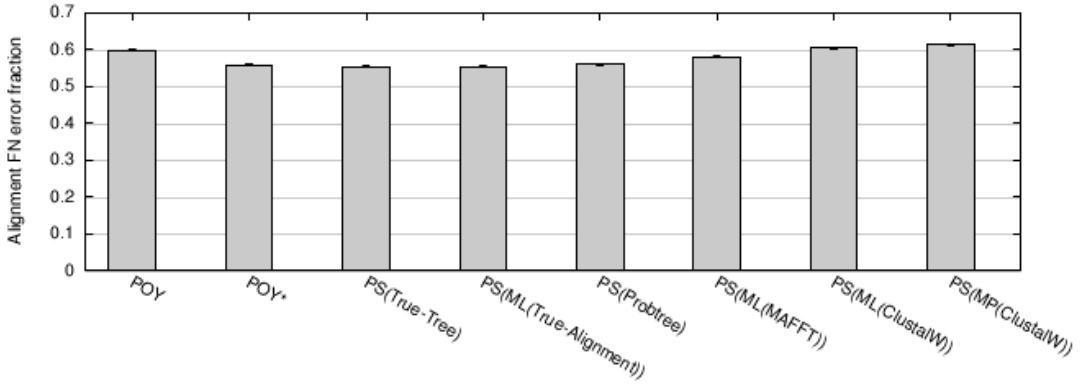
Results – Affine Treelength Criterion



(a)



(b)



(c)

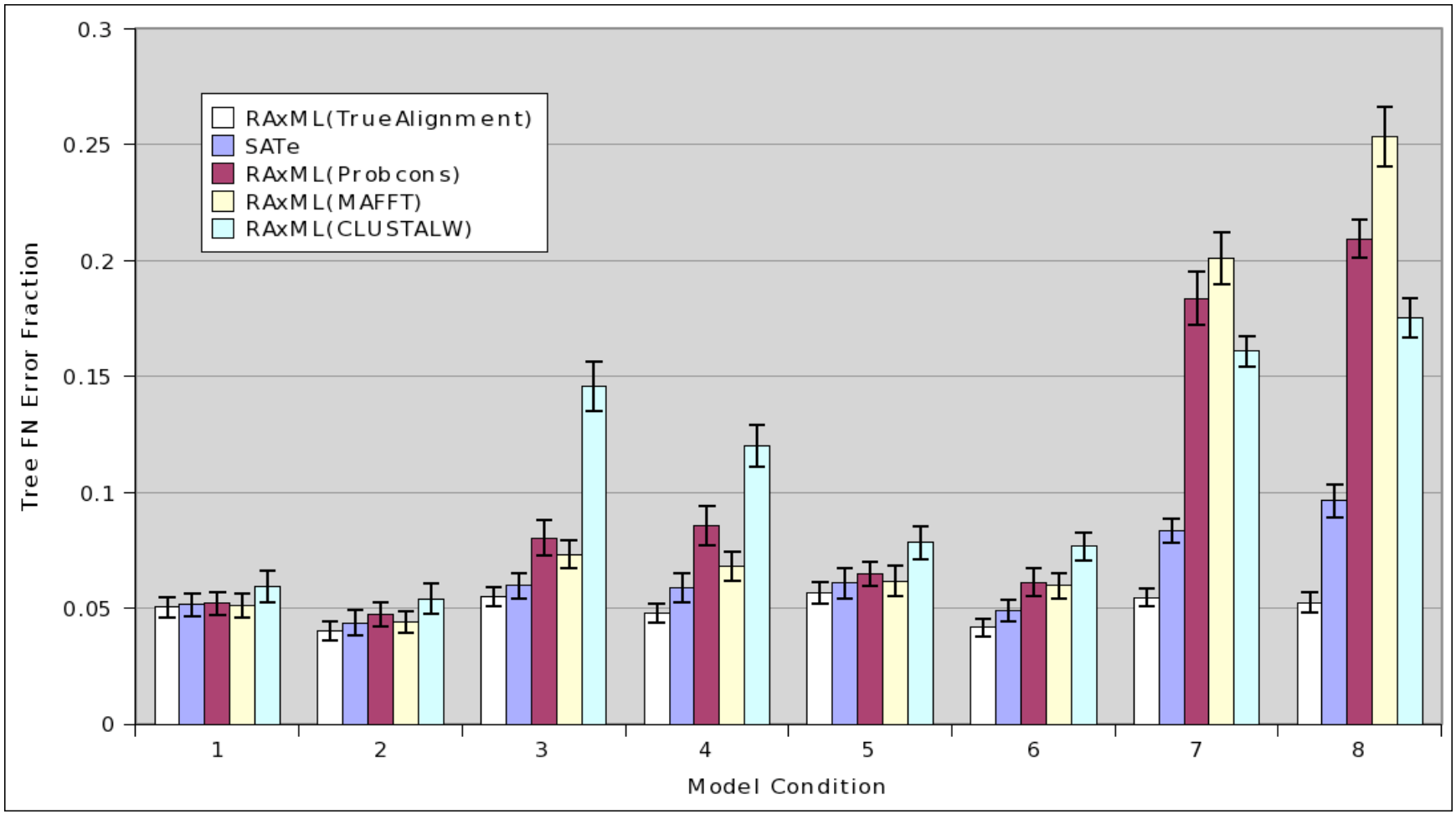
Summary (so far)

- Optimizing treelength can produce very alignments that are better than Clustal, provided that affine gap penalties are used instead of simple (contrary to Ogden and Rosenberg).
- Trees producing through optimizing treelength can be competitive with the best two-phase methods (even with Probtree and ML(MAFFT)).
- However, continued improvement using such techniques seems unlikely.

Part II: SATé:
(Simultaneous Alignment and Tree Estimation)

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

Our method (SATé) vs. other methods



- 100 taxon model trees, GTR+Gamma+gap,
- Long gap models 1-4, short gap models 5-8

Observations, Conclusions, and Conjectures

- Alignment accuracy is probably not best measured using standard criteria, at least if phylogeny estimation is the objective.
- Improved two-phase methods are possible, but simultaneous estimation of alignments and trees is likely to yield better results.
- Statistical co-estimation using gaps is probably essential (but we need good models!).
- Scalability is important.

Acknowledgments

- Collaborators: Randy Linder (Integrative Biology, UT-Austin), and students Kevin Liu, Serita Nelesen, and Sindhu Raghavan
- Funding: the US National Science Foundation, the Newton Institute at Cambridge University, the Program for Evolutionary Dynamics at Harvard, and the Radcliffe Institute.