# Simultaneous alignment and tree reconstruction

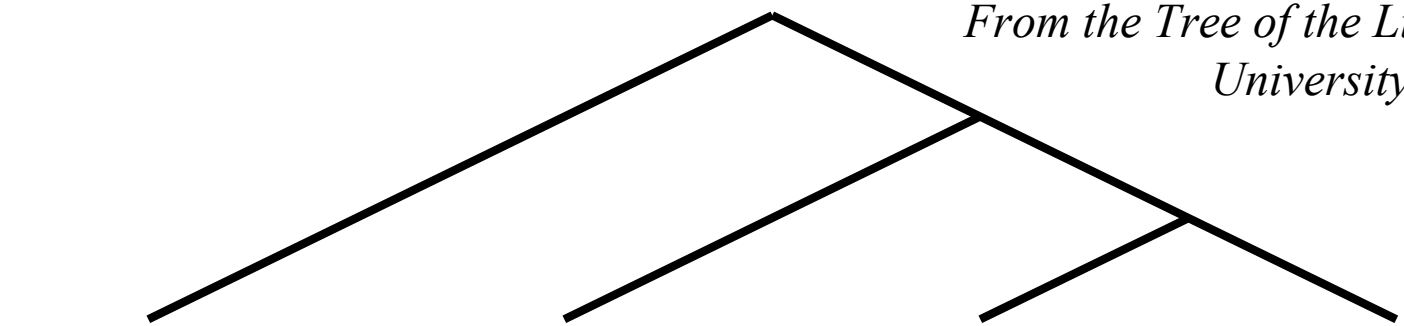
Collaborative grant:

Texas, Nebraska, Georgia, Kansas

Penn State University, Huston-Tillotson, NJIT, and the Smithsonian Institution

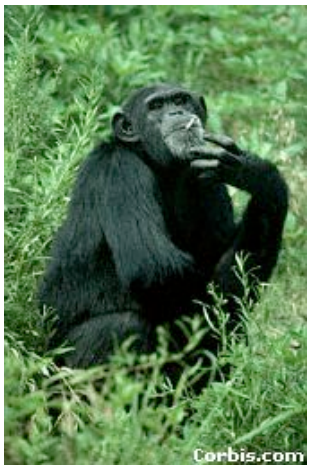
# Species phylogeny

*From the Tree of the Life Website,*
*University of Arizona*



Orangutan Gorilla Chimpanzee Human






# Project Components

- Algorithms and Software
- Simulations
- Outreach to ATOL and the scientific community
- Undergraduate training

# Personnel

- Tandy Warnow (UT-Austin)
- Mark Holder (Kansas)
- Jim Leebens-Mack (UGA)
- Randy Linder (UT-Austin)
- Etsuko Moriyama (UNL)
- Michael Braun (Smithsonian)
- Webb Miller (PSU)
- Usman Roshan (NJIT)
- Postdocs: Derrick Zwickl (NESCENT)
- PhD Students: Cory Strope (UNL), Serita Nelesen (UT-Austin), Kevin Liu (UT-Austin), Sindhu Raghavan (UT-Austin), Michael McKain (UGA)
- Undergraduates: from Huston-Tillotson and the University of Georgia

# 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
```



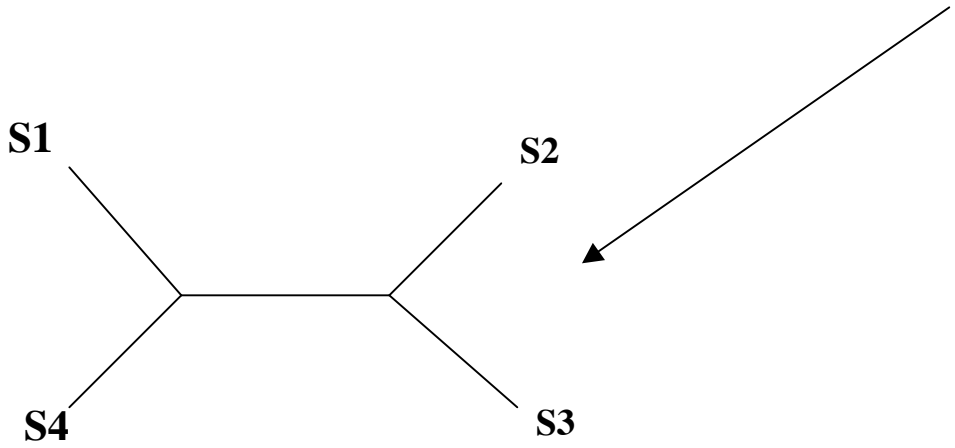
```
S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAC--GACCGC--
S3 = TAG-CT-------GACCGC--
S4 = -------TCAC--GACCGACA
```

# 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
```



# 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”).

# Alignment methods

- The standard alignment method for phylogeny is **Clustal** (or one of its derivatives)
- Others: ProbCons, MAFFT, Muscle, POA, POY, T-Coffee, Di-Align
- On the basis of various tests, **ProbCons, Mafft,** and **Muscle** are generally considered the “best”.

# Basic Questions

- Using simulations: Does improving the alignment lead to an improved phylogeny?
- Using Tree of Life (real) datasets:
  - How much does changing the alignment method change the resultant alignments?
  - How much does changing the alignment method change the estimated tree?
  - What gap patterns do we see on hand-curated alignments, and what biological processes created them?

# Basic Questions

- Using simulations: Does improving the alignment lead to an improved phylogeny?
- Using Tree of Life (real) datasets:
  - How much does changing the alignment method change the resultant alignments?
  - How much does changing the alignment method change the estimated tree?
  - What gap patterns do we see on hand-curated alignments, and what biological processes created them?

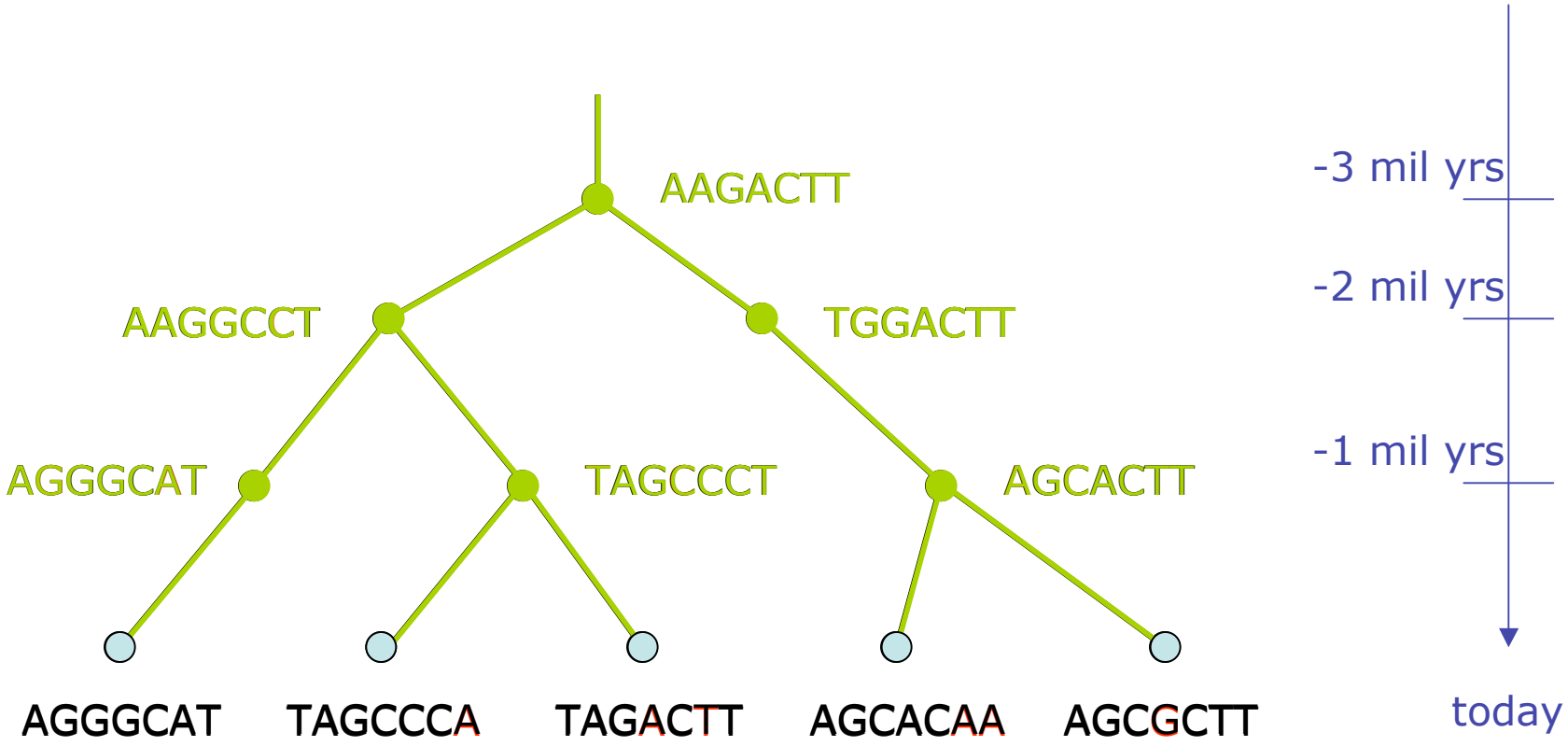
# Simulation study

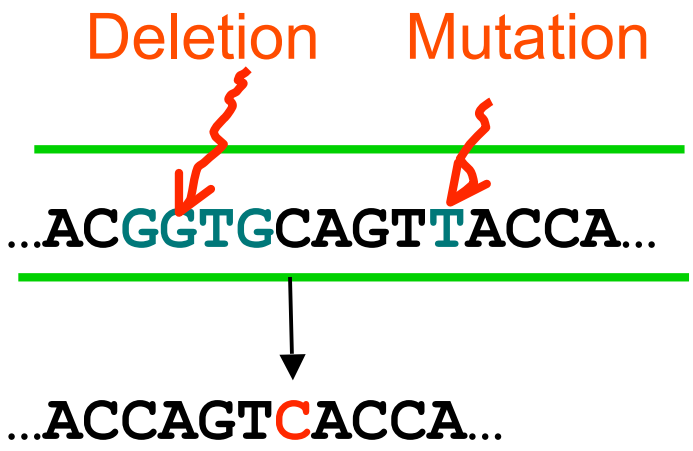
- Simulate sequence evolution down a tree
- Estimate alignments on each set of sequences
- Compare estimated alignments to the true alignment
- Estimate trees on each alignment
- Compare estimated trees to the true tree

# Simulation study

- Simulate sequence evolution down a tree
- Estimate alignments on each set of sequences
- Compare estimated alignments to the true alignment
- Estimate trees on each alignment
- Compare estimated trees to the true tree

# DNA Sequence Evolution



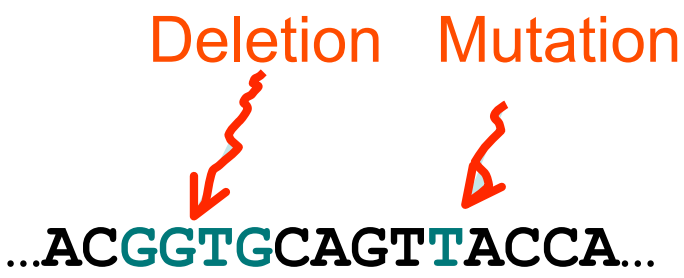



indels (insertions and deletions) also occur!

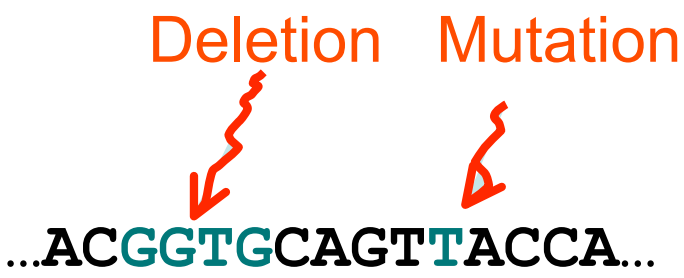
# Indels and substitutions at the DNA level

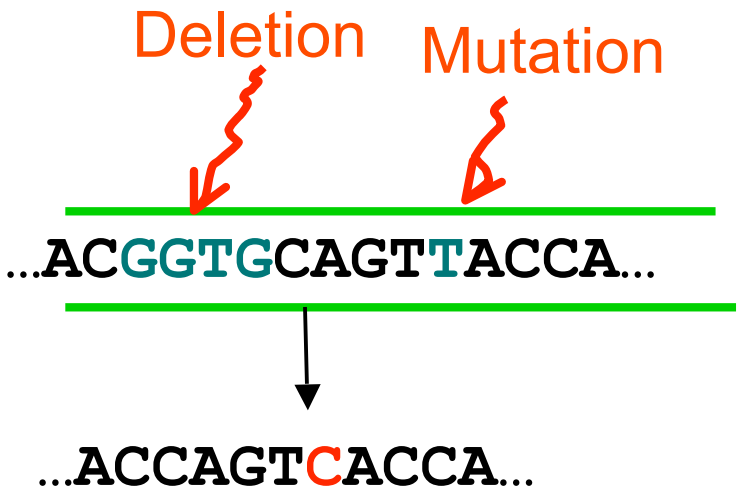
...**ACGGTGCAGTTACCA**...

# Indels and substitutions at the DNA level



# Indels and substitutions at the DNA level






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 Calculation

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

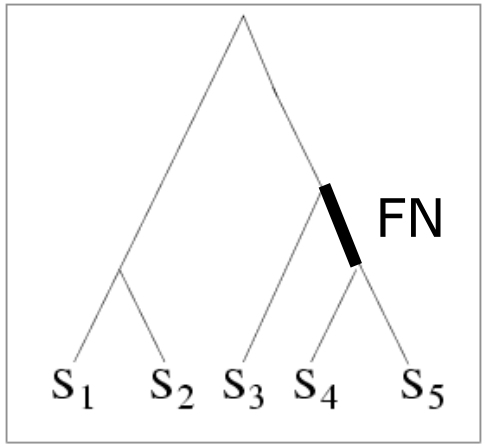
- A C A T G - - -  Est. alignment
- - C A A G A T G

# Alignment Error Calculation

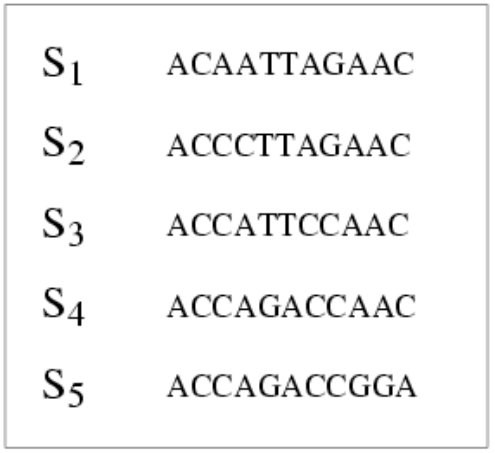
- A C A T - - - G  True alignment
- C A A - G A T G

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

- 75% of the correct pairs are missing!



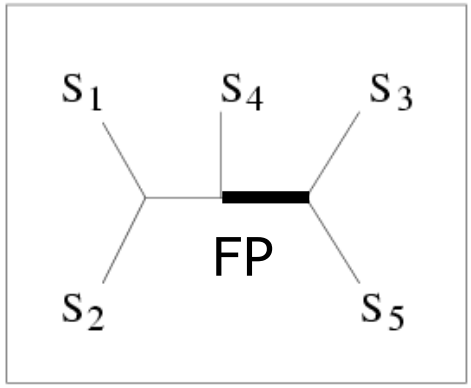

TRUE TREE




DNA SEQUENCES

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

50% error rate

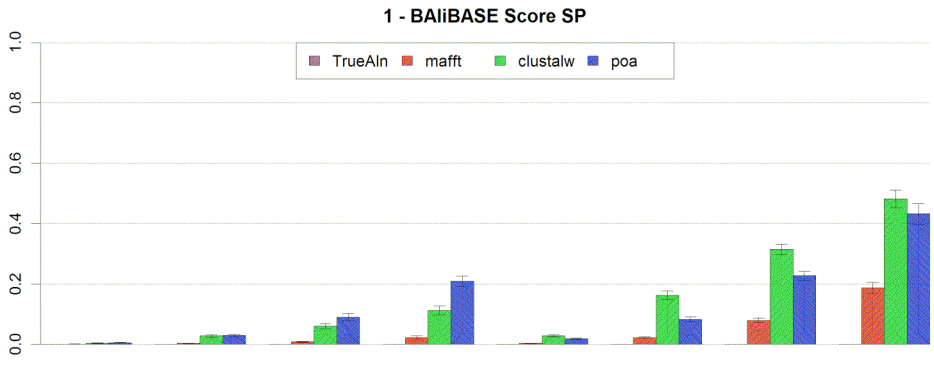


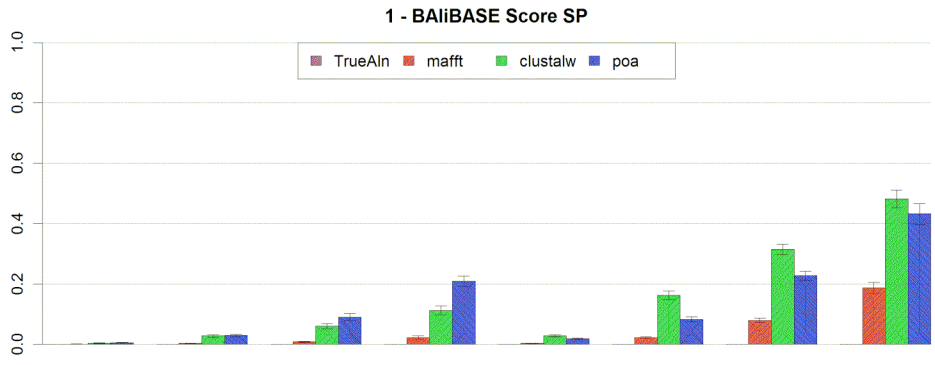

INFERRED TREE

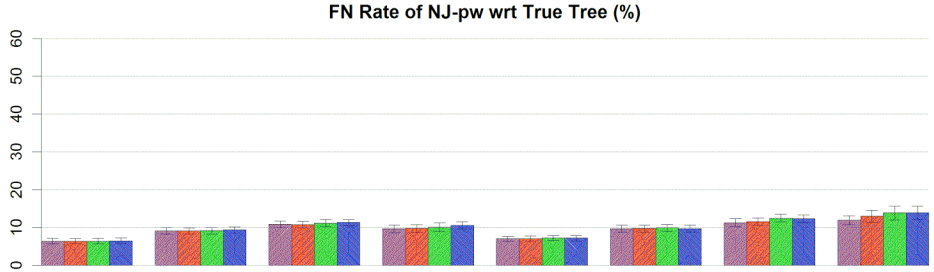
# Our simulation studies (using ROSE*)

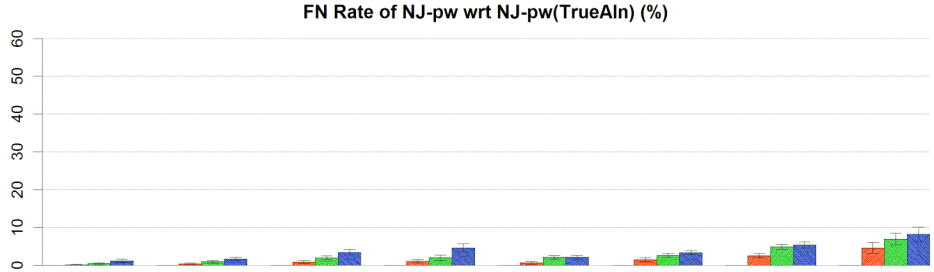
- Amino-acid evolution (Wang et al., unpublished):
  - BaliBase and birth-death model trees, 12 taxa to 100 taxa.
  - Average gap length 3.4.
  - Average identity 23% to 57%.
  - Average gappiness 3% to 60%.
- DNA sequence evolution (Liu et al., unpublished):
  - Birth-death trees, 25 to 500 taxa.
 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

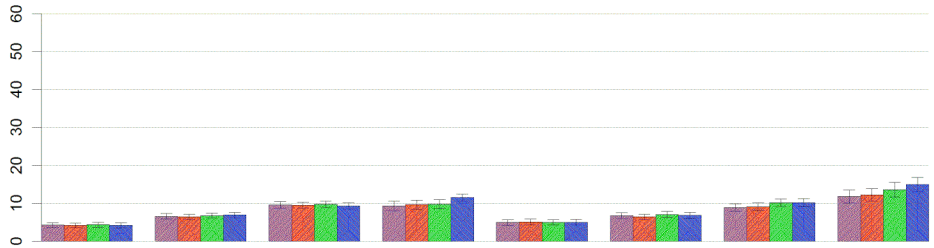
***ROSE has limitations!***

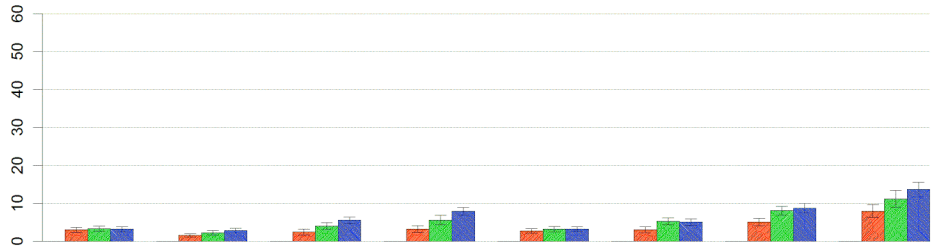

1 - BAliBASE Score SP
TrueAln
mafft
clustalw
poa
0.0
0.2
0.4
0.6
0.8
1.0

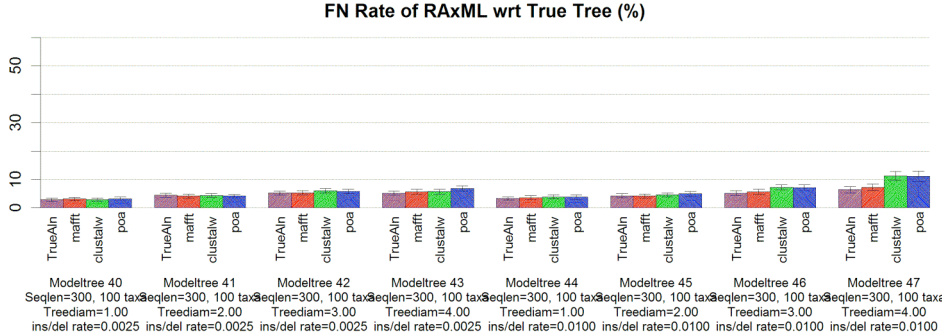

1 - BAliBASE Score SP
TrueAln
mafft
clustalw
poa
0.0
0.2
0.4
0.6
0.8
1.0

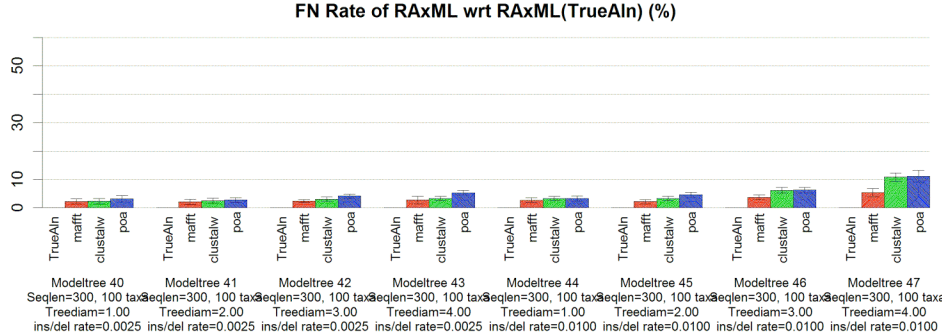

FN Rate of NJ-pw wrt True Tree (%)
0
10
20
30
40
50
60


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

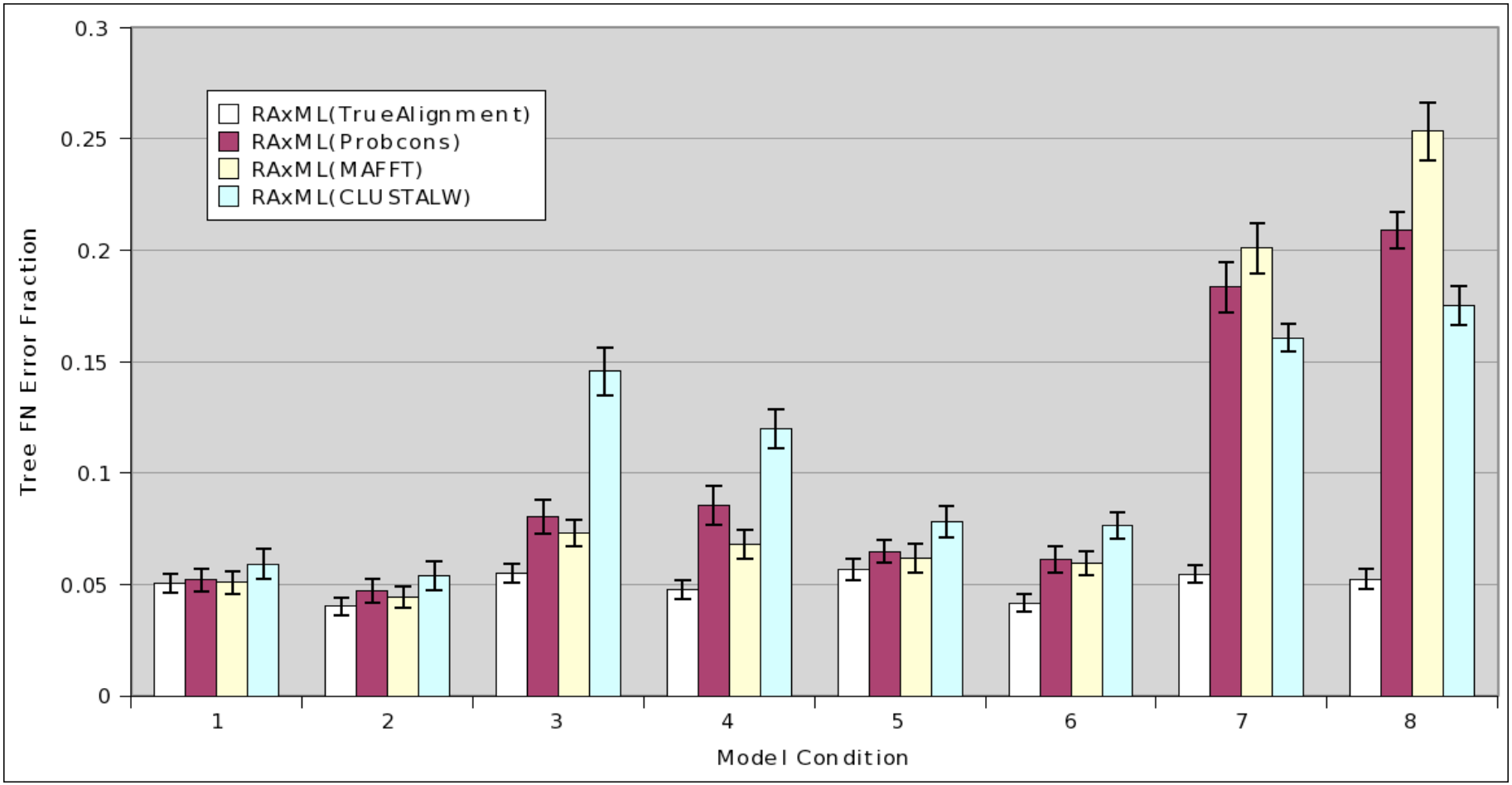

FN Rate of Maj-PAMwRatchet(InfAln) wrt True Tree (%)
0
10
20
30
40
50
60


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.

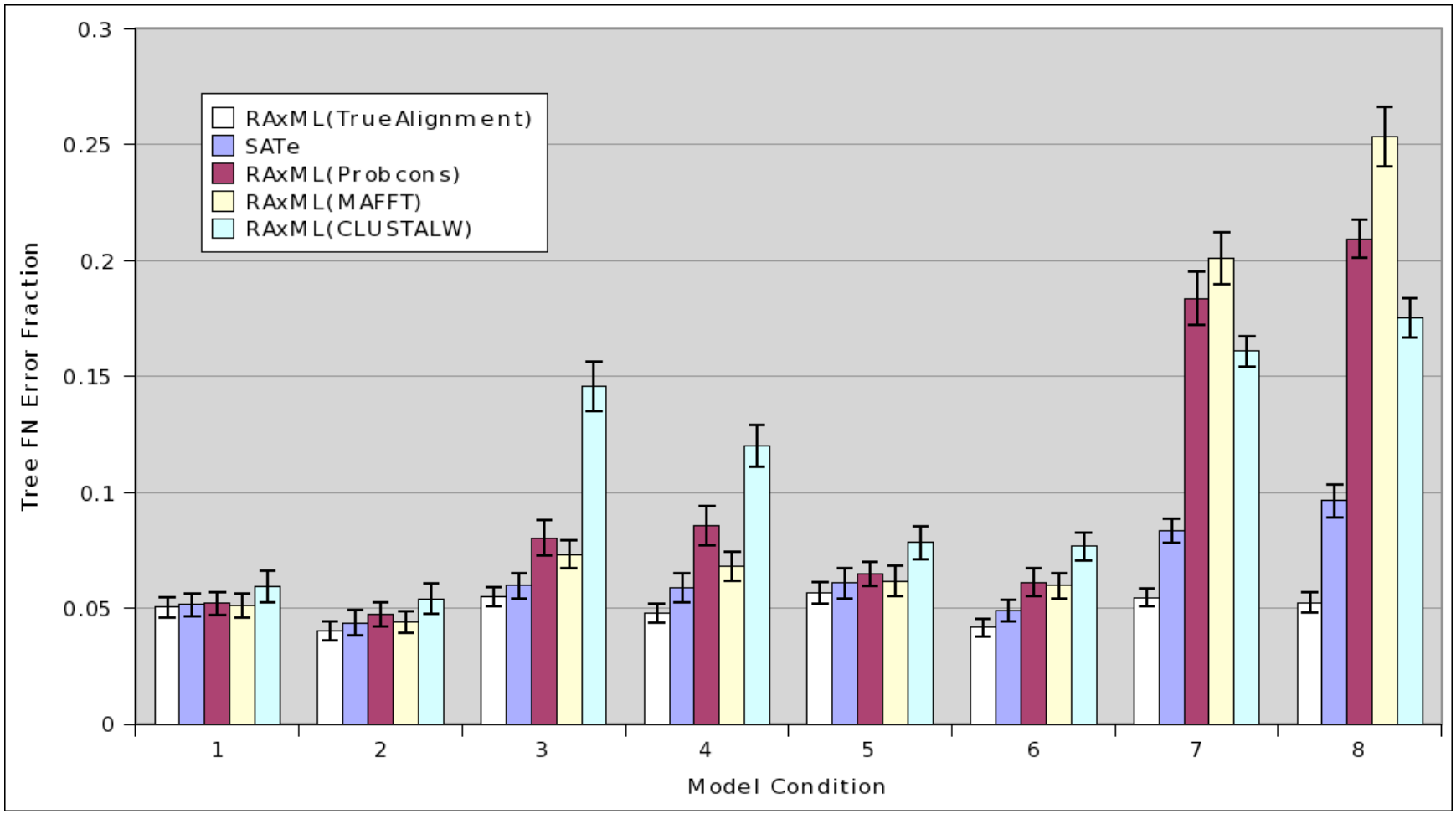
# Progress so far

- Experimental evaluation of existing alignment methods (Wang, Leebens-Mack, de Pamphilis and Warnow) - submitted
- Impact of guide trees (Nelesen, Liu, Linder, and Warnow): Pacific Symp. Biocomputing 2008
- Better ways to run POY: Liu, Nelesen, Raghavan, Linder, and Warnow (submitted)
- SATé: new technique for Simultaneous Alignment and Tree Estimation: Liu, Nelesen, Linder and Warnow (in preparation)

# SATe:
## (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).
- **SATe** returns the alignment/tree pair that optimizes maximum likelihood under GTR+Gamma+I.

# Our method (SATé) vs. other methods




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

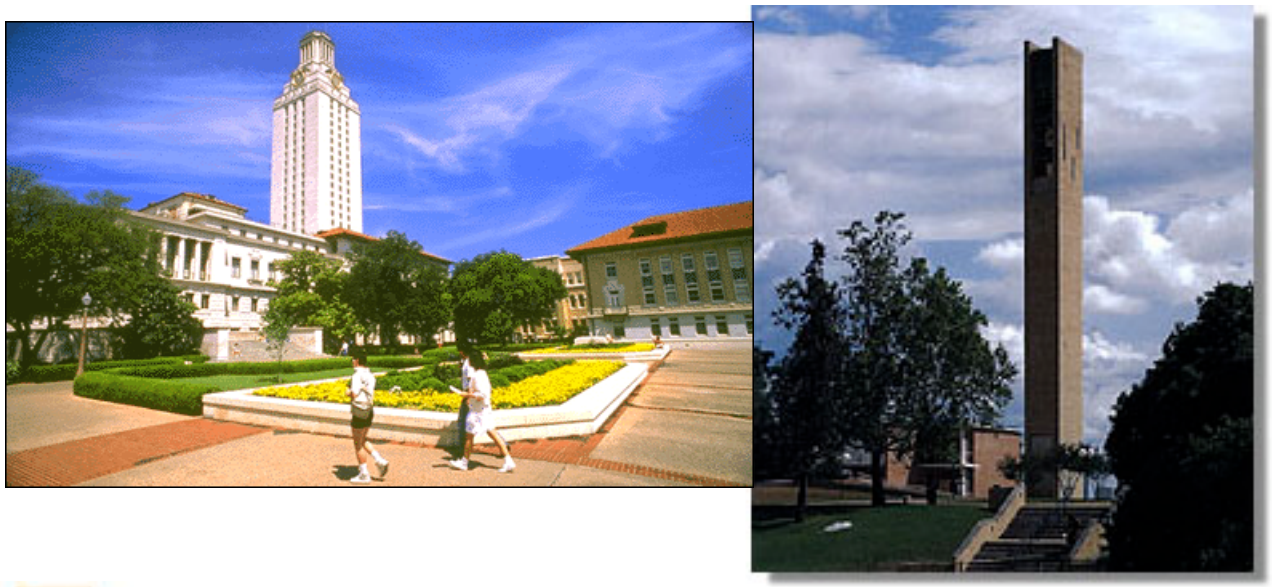
# Undergraduate Training

- Two institutions involved: UT-Austin partnership with Huston-Tillotson, and the University of Georgia
- Training via:
  - Research projects
  - Summer training with the project members
  - Participation in the project meeting
  - Participation at a conference
  - Lectures by project participants at the collaborating institutions
- Focus group leader(s): Jim Leebens-Mack and Randy Linder

# Undergraduate Research Programs at the University of Georgia

# Louis Stokes Alliance for STEM Research

# University of Texas Collaboration with Huston-Tillotson University



# Research projects for undergrads

- Studying the AToL (Assembling the Tree of Life) project datasets:
  - Produce alignments on each dataset, (using existing alignment methods and our new SATe method), and compute trees on each alignment
  - Study differences between alignments and between trees
- Evaluating the simulation software
- Creating a webpage about alignment research
- Others?