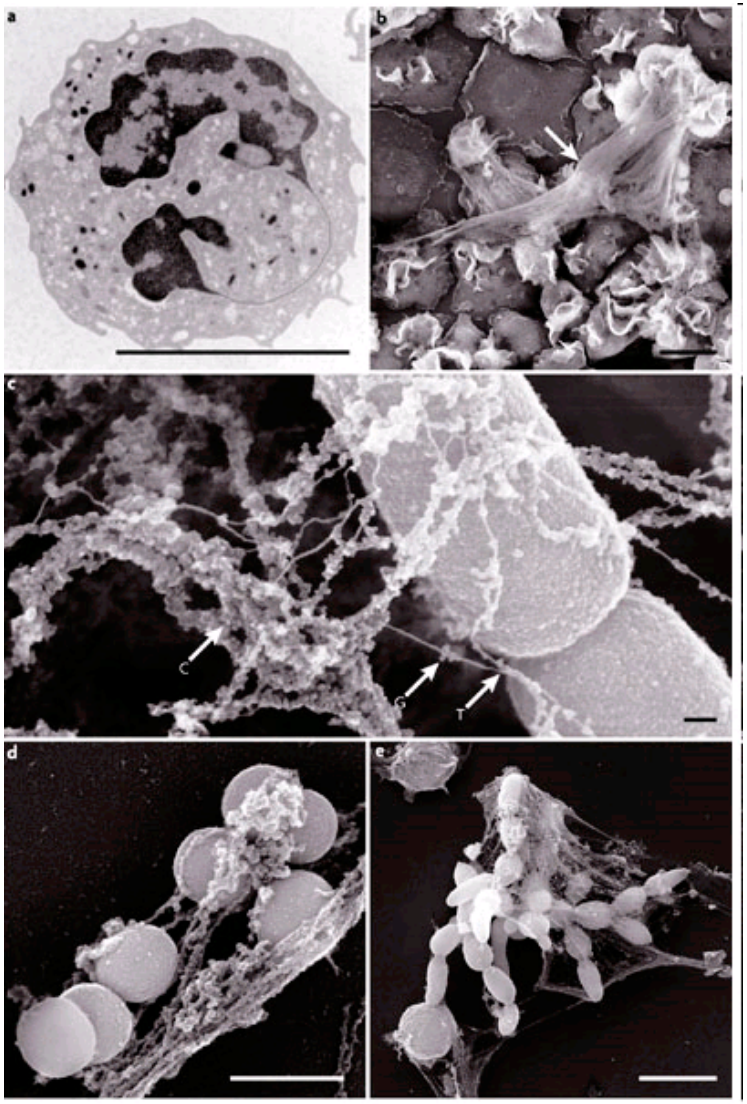
Assembling the Tree of Life: Simultaneous Sequence Alignment and Tree Reconstruction

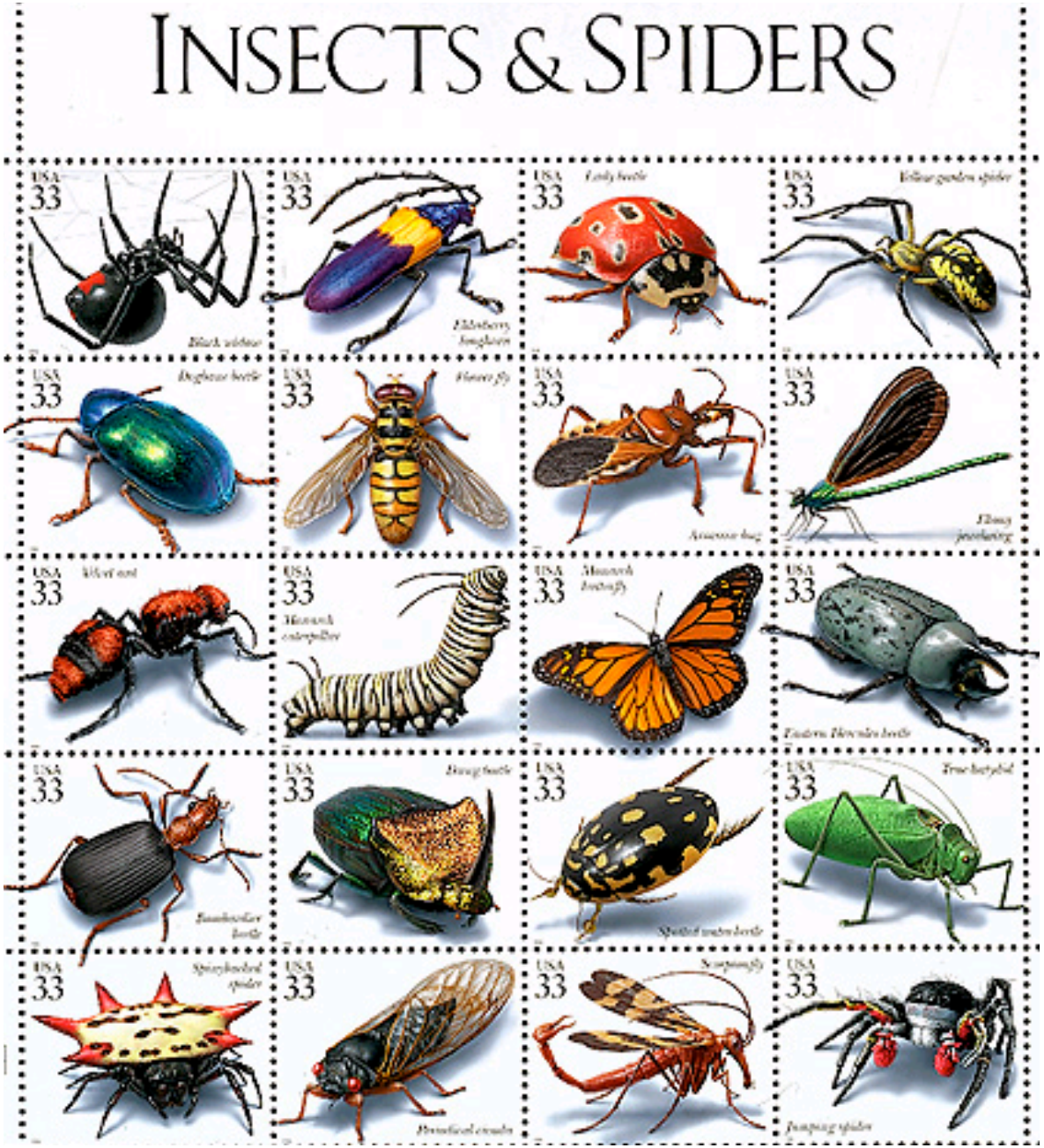
Collaborative grant:

Texas, Nebraska, Georgia, Kansas

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



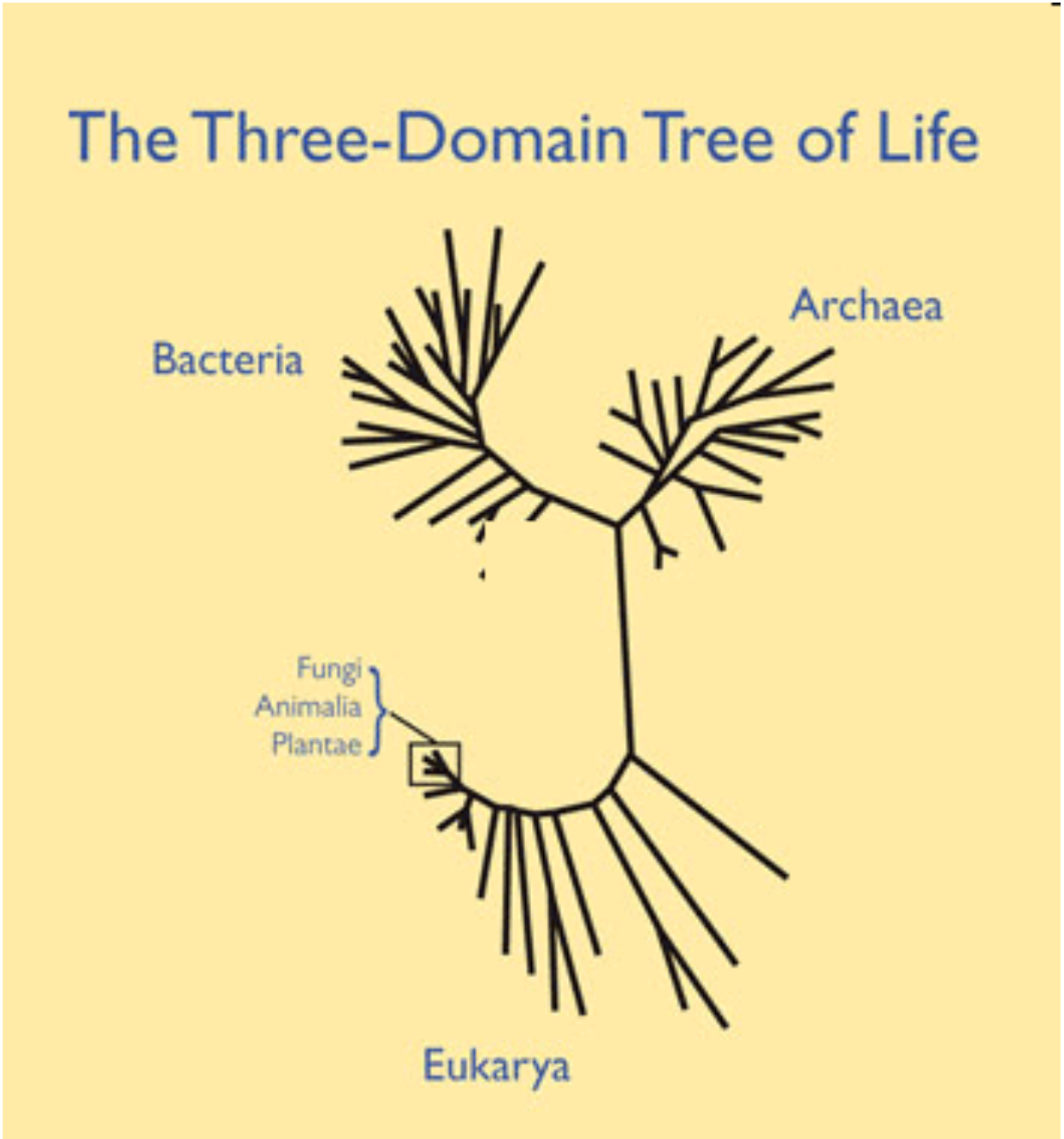
Nature Reviews | Microbiology



Nobody Knows …
How Many Species There Are

- Probably around 10 million

- Evolutionary biology and molecular biology have both strongly supported the idea that all of life has arisen from a single common ancestor, ~3.6 billion years ago


The Three-Domain Tree of Life
Bacteria
Archaea
Fungi
Animalia
Plantae
Eukarya

But how can we figure out the speciation pattern of life?

- The process of speciation has played out over billions of years
- We weren't around to witness most species
- Instead we have a detective story
  - Life has left us clues about its evolution
  - We have to figure out how to best collect and use those clues!
- Our project is working to develop methods that do a better job of using the data and allowing researchers to work with much larger datasets.

Project Components

- Algorithms and Software
- Simulations
- Outreach to ATOL and the scientific community
- Undergraduate training and research
  - (This is where you come in.)

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)

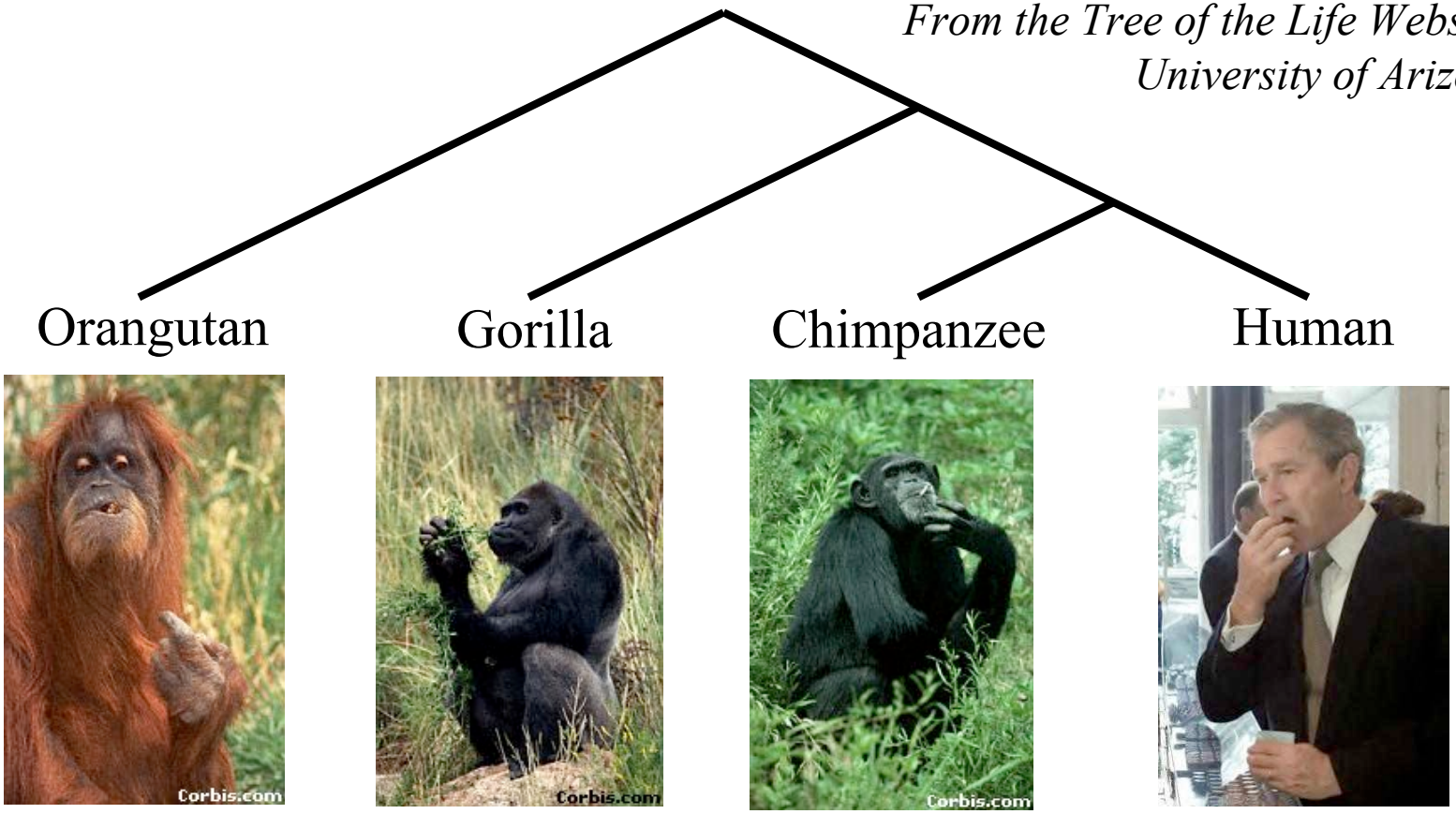



- Postdocs: Derrick Zwickl (NESCENT), Cory Strope (UNL)
- UT PhD Students: Serita Nelesen, Kevin Liu, Sindhu Raghavan, Shel Swenson
- UGA PhD Student: Michael McKain
- Undergraduates: from Huston-Tillotson and the University of Georgia

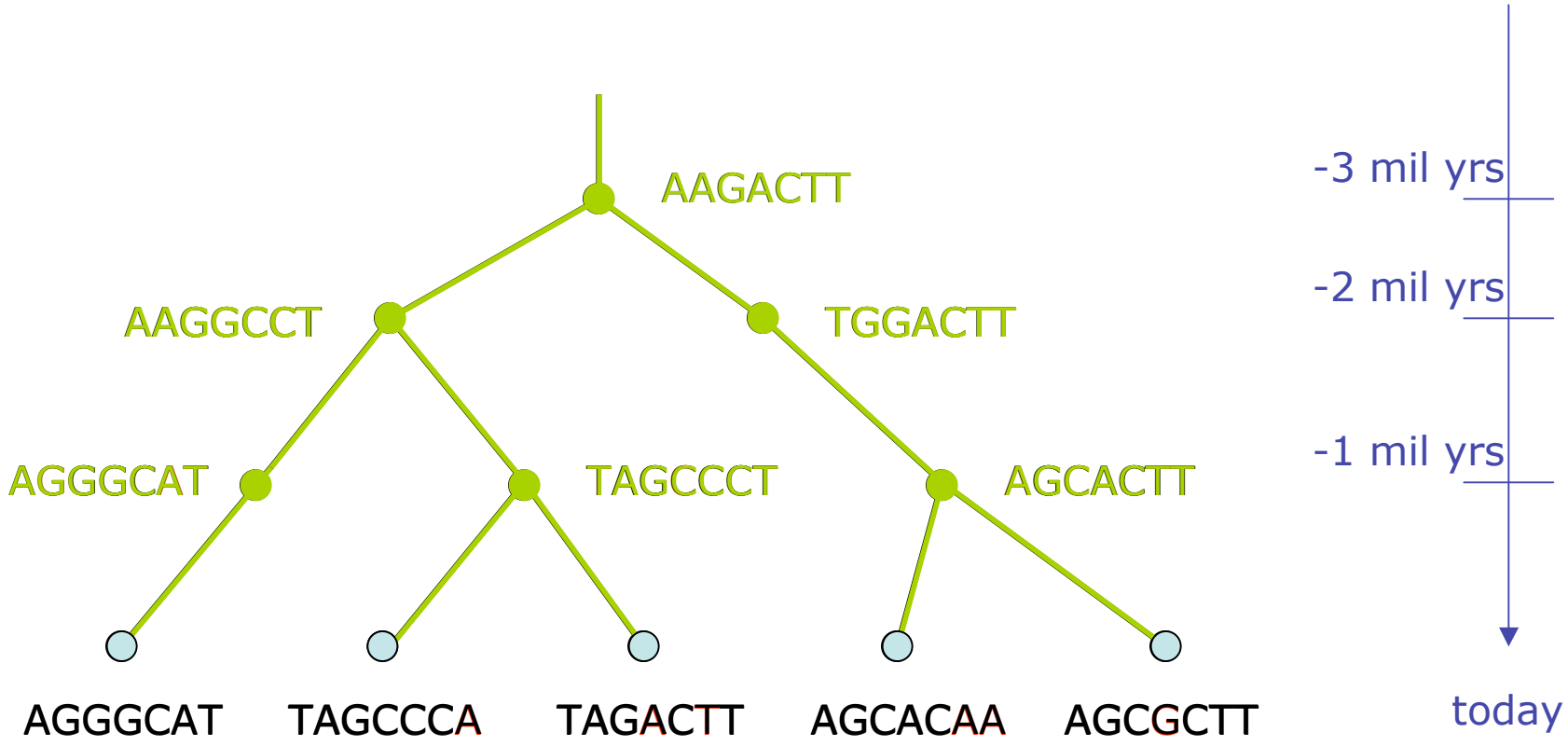



Species phylogeny

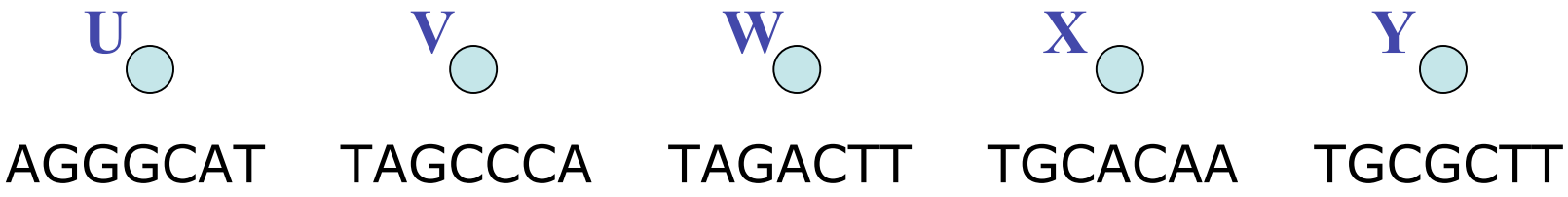
From the Tree of the Life Website,
University of Arizona

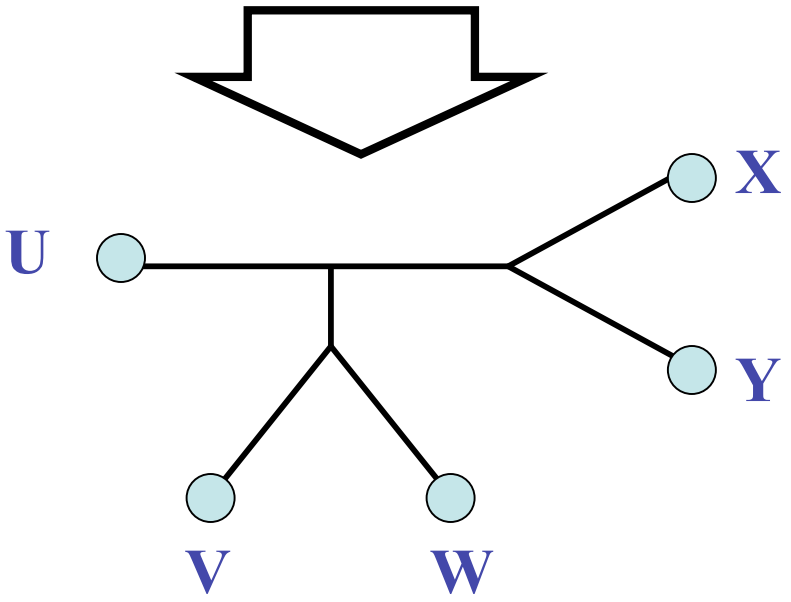


DNA Sequence Evolution



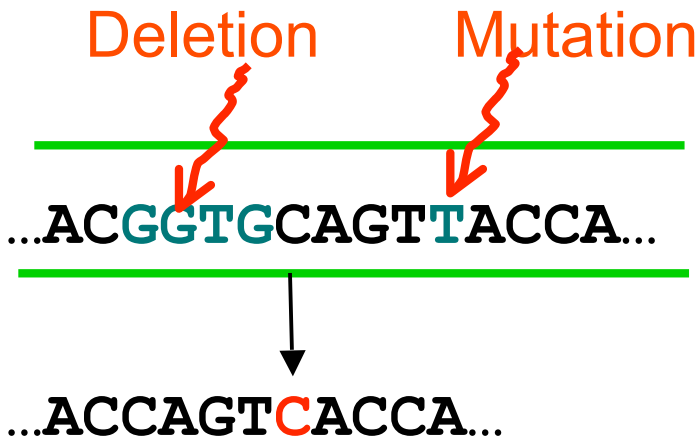
Phylogeny Problem



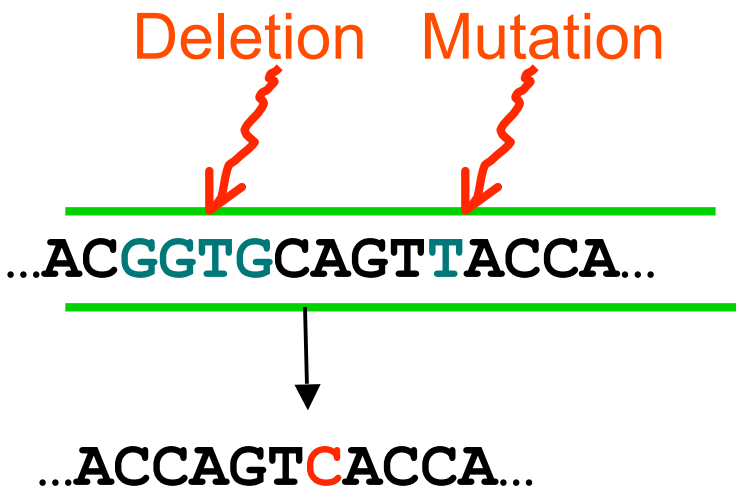


But solving this problem exactly is … unlikely

# 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}



But indels (insertions and deletions) also occur!



The true pairwise alignment is:

...ACGGTGCAGTTACCA...

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

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”).

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
```



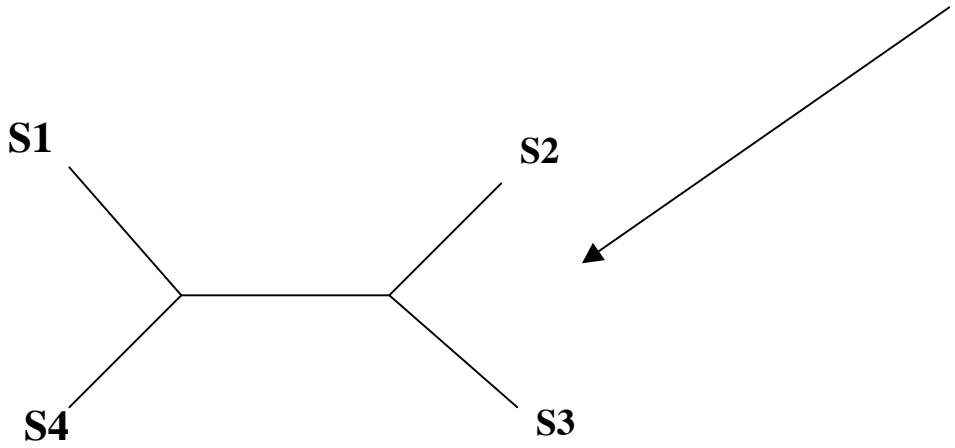
```
S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAC--GACCGC--
S3 = TAG-CT-------GACCGC--
S4 = -------TCAC--GACCGACA
```

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
```



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

- 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?

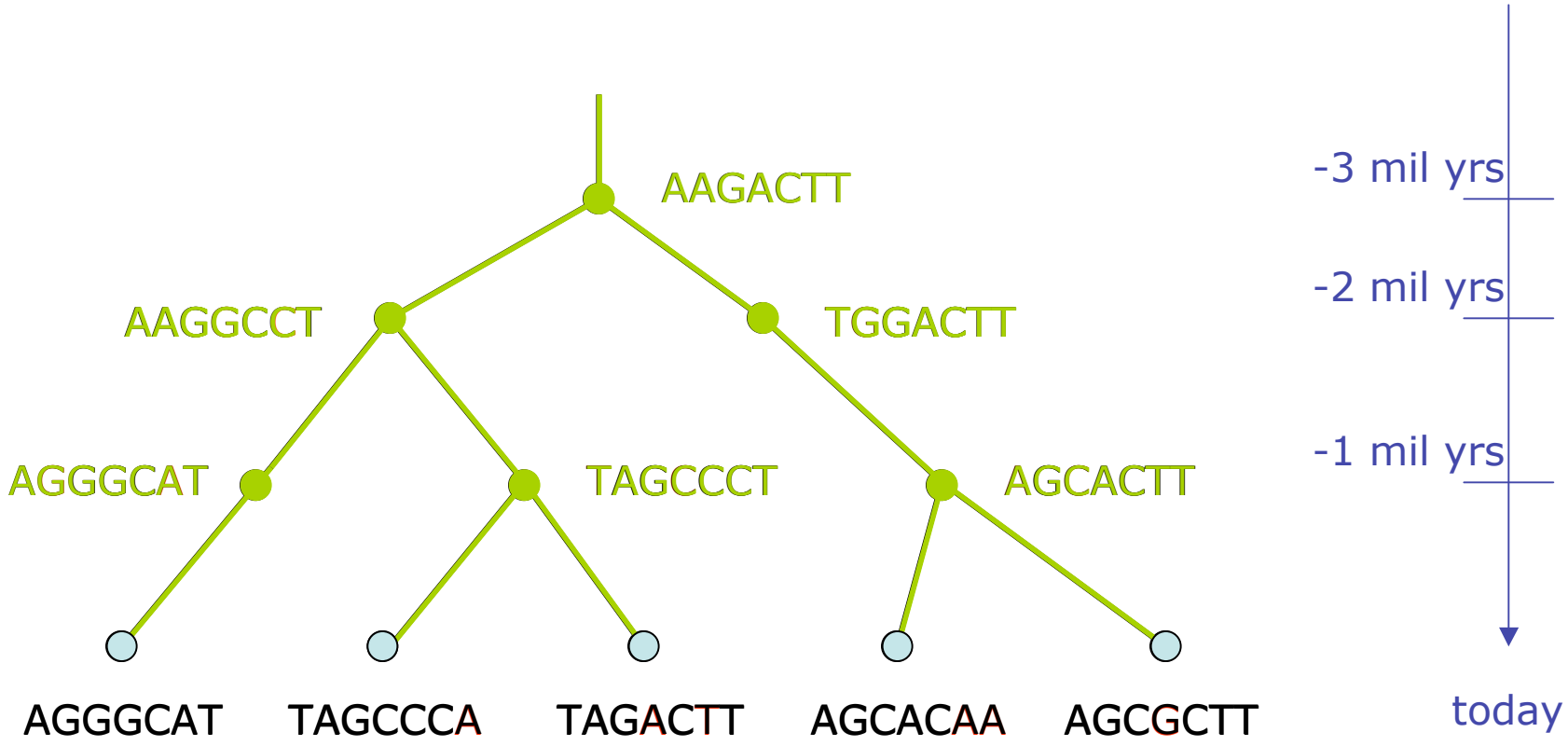
Our progress (so far)

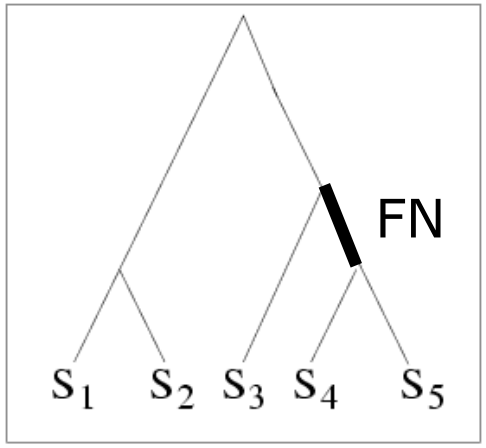
- Experimental evaluation of existing alignment methods - submitted
- Impact of guide trees: Pacific Symp. Biocomputing 2008
- “Barking up the wrong treelength” (Better ways to run POY): Transactions on Computational Biology and Bioinformatics 2009
- SATé: new technique for Simultaneous Alignment and Tree Estimation: submitted

Simulation study

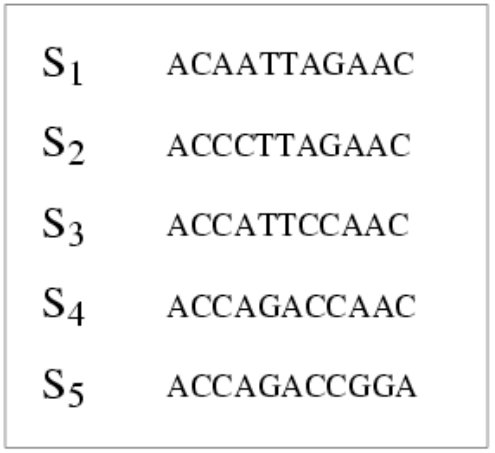
- Simulate sequence evolution down a tree
- Estimate alignments on each set of sequences
- Compare estimated alignments to the true alignment
- Estimate trees on each alignment
- Compare estimated trees to the true tree

DNA Sequence Evolution





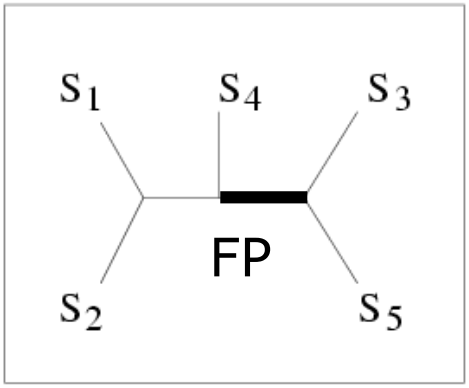
TRUE TREE



DNA SEQUENCES

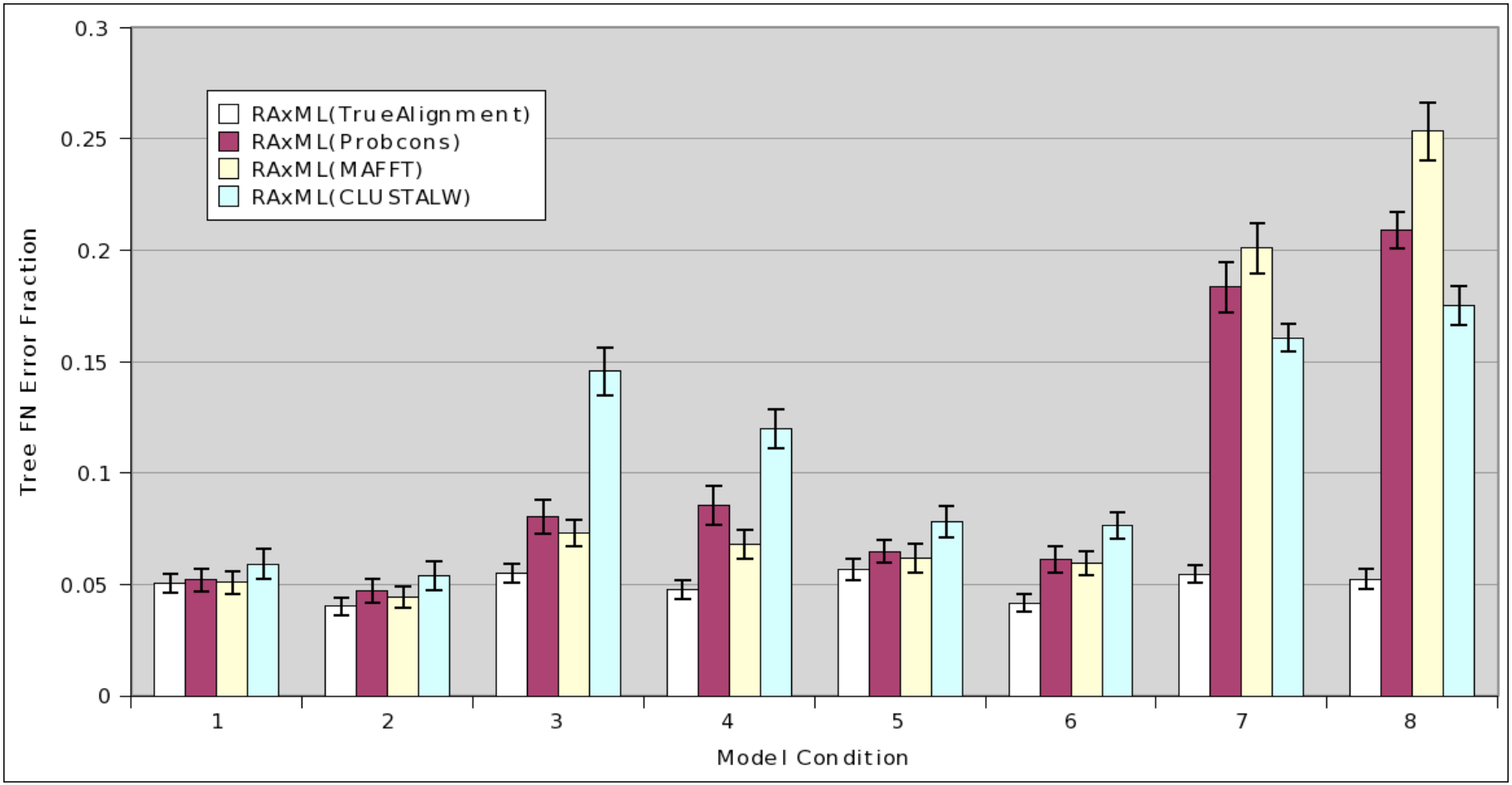
FN: false negative
(missing edge)
FP: false positive
(incorrect edge)

50% error rate



INFERRED TREE

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, but the degree of improvement in tree accuracy is ***much smaller*** (data not shown).
- 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.

What we'd like (ideally)

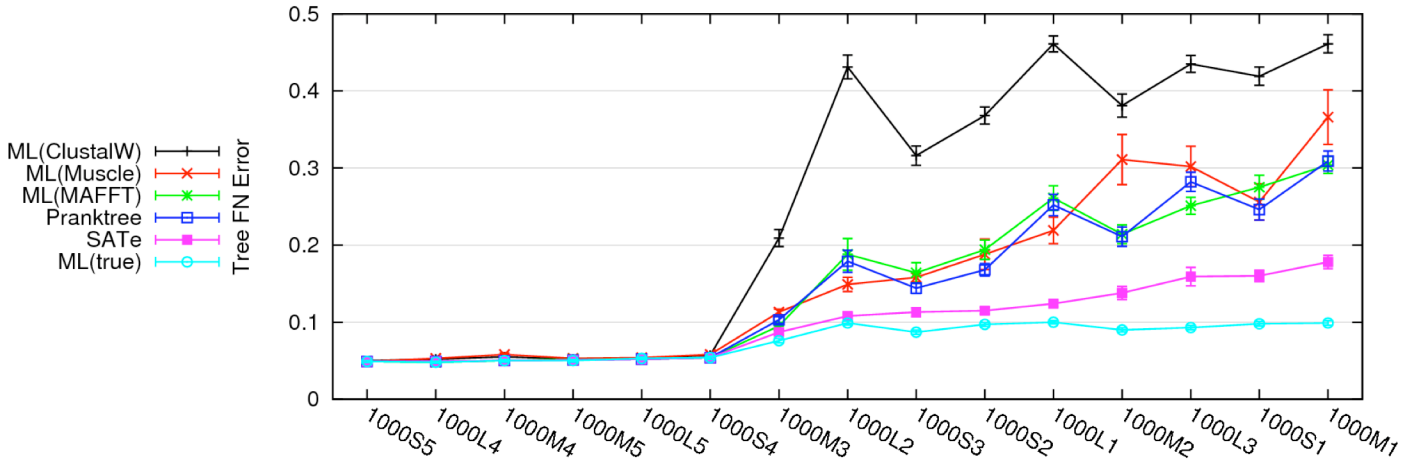
- An automated means of practically inferring alignments and very large phylogenetic trees using sequence (DNA, protein) data
  - Very large means at least thousands, but as many as tens of thousands of taxa
  - Preferably able to run on a desktop computer
- Doing this with a minimum of human (subjective) input on the alignment in particular

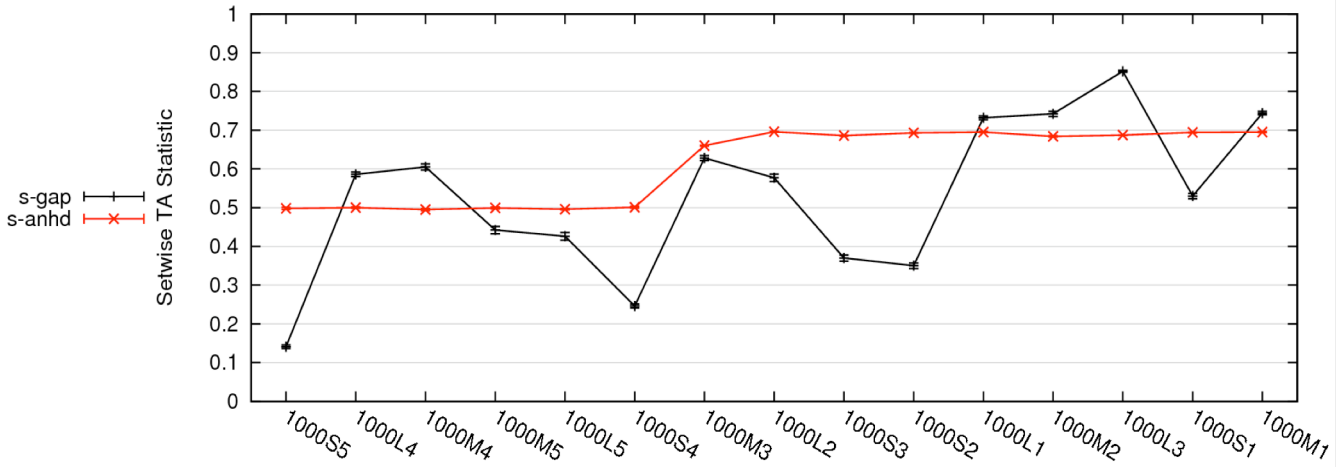
SATe:
(Simultaneous Alignment and Tree Estimation)

- Developers: Liu, Nelesen, Raghavan, Linder, and Warnow.
- Technique: search through tree/alignment space (re-align sequences on each tree using a novel divide-and-conquer strategy, and then compute ML trees on the resultant multiple alignments).
- **SATe** returns the alignment/tree pair that optimizes maximum likelihood under GTR+Gamma+I.

1000 taxon simulation study

- Missing edge rates
- Empirical statistics





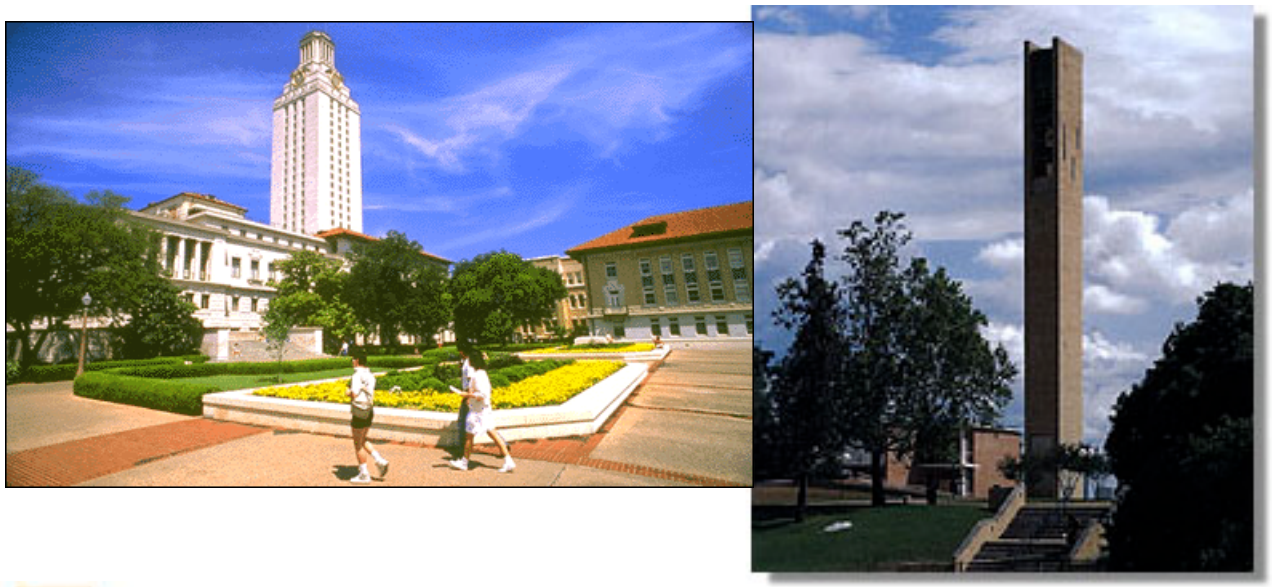
Undergraduate Training

- 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?