

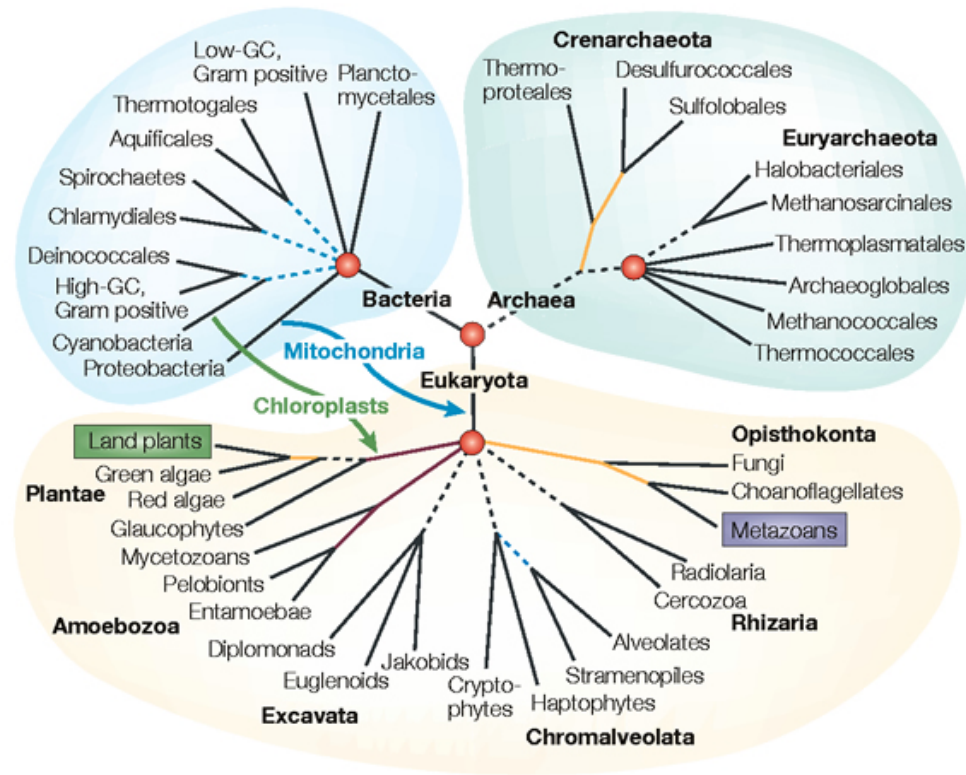
# **Four techniques for large-scale multiple sequence alignment and phylogenetic estimation**

Tandy Warnow

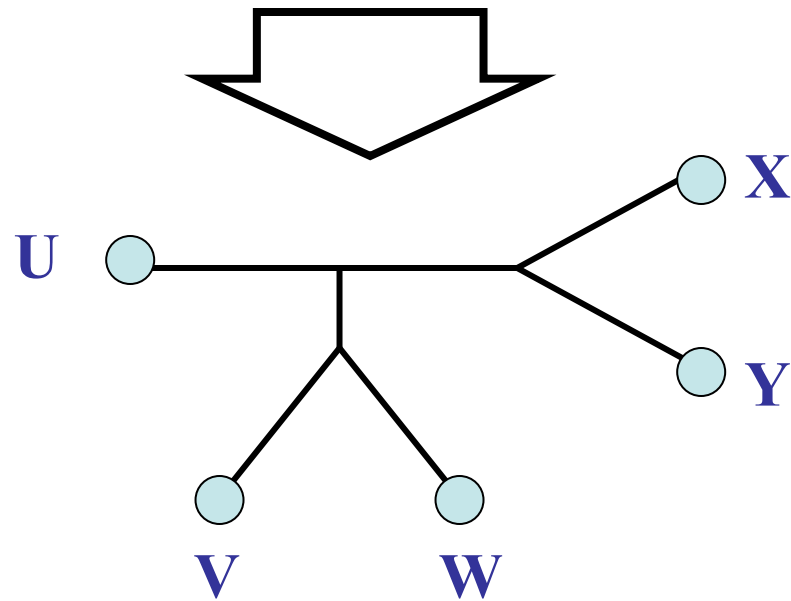
Department of Computer Science

The University of Texas at Austin

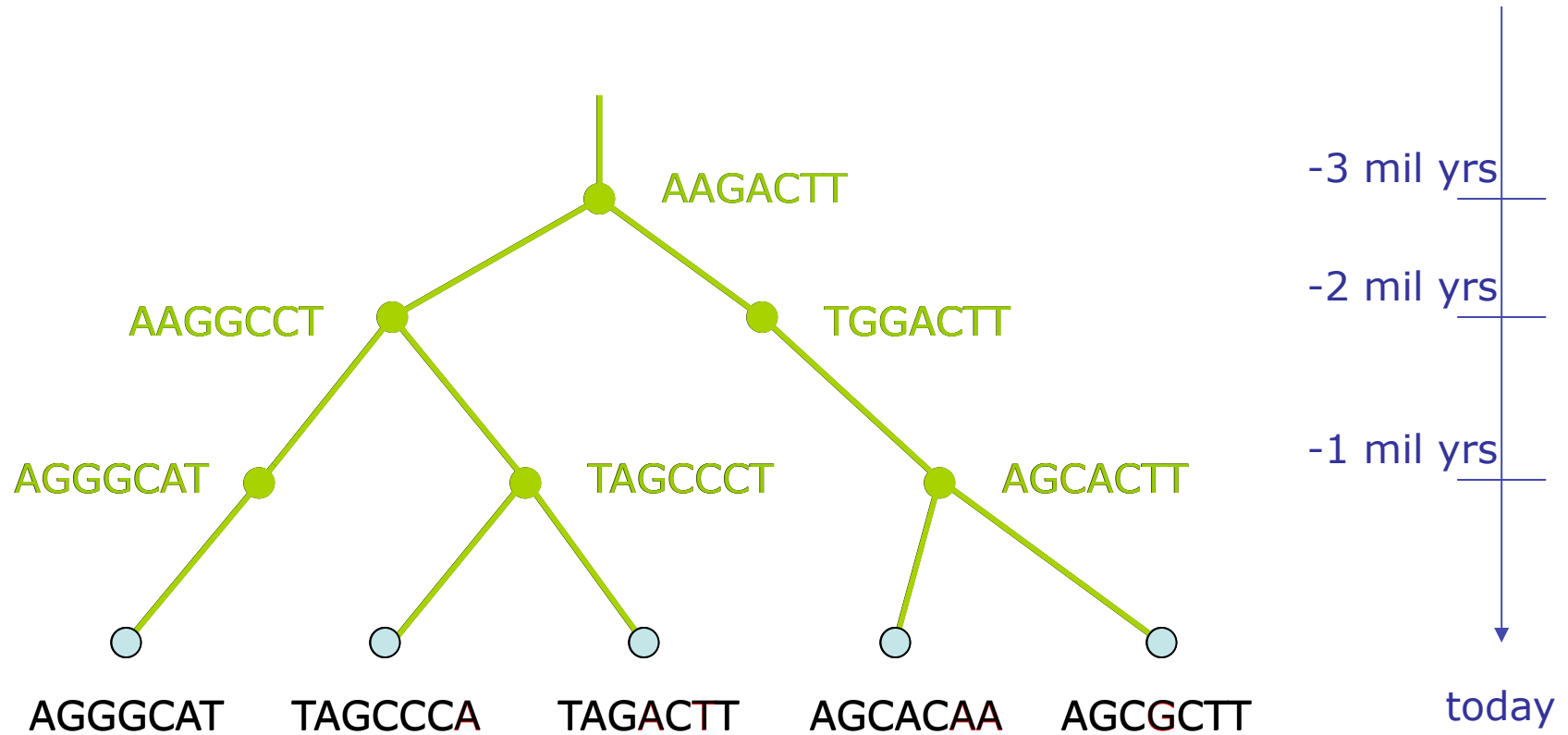
# Assembling the Tree of Life



U AGGGCATGA      V AGAT      W TAGACTT      X TGCACAA      Y TGCGCTT



# DNA Sequence Evolution





### The true multiple alignment

- Reflects historical substitution, insertion, and deletion events
- Defined using transitive closure of pairwise alignments computed on edges of the true tree

# Input: unaligned sequences

S1 = AGGCTATCACCTGACCTCCA

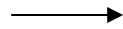
S2 = TAGCTATCACGACCGC

S3 = TAGCTGACCGC

S4 = TCACGACCGACA

# Phase 1: Alignment

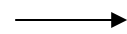
S1 = AGGCTATCACCTGACCTCCA  
S2 = TAGCTATCACGACCGC  
S3 = TAGCTGACCGC  
S4 = TCACGACCGACA



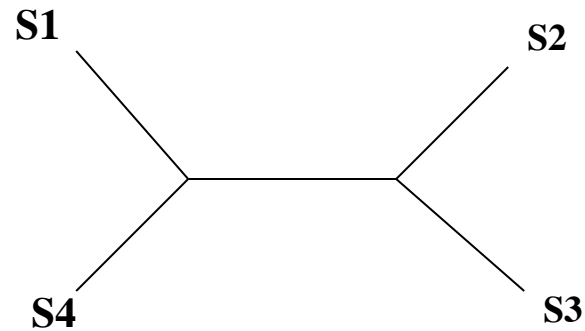
S1 = -AGGCTATCACCTGACCTCCA  
S2 = TAG-CTATCAC--GACCGC--  
S3 = TAG-CT-----GACCGC--  
S4 = -----TCAC--GACCGACA

# 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





# Two-phase estimation

## Alignment methods

- Clustal
- POY (and POY\*)
- Probcons (and Probtree)
- Probalign
- MAFFT
- Muscle
- Di-align
- T-Coffee
- Prank (PNAS 2005, Science 2008)
- Opal (ISMB and Bioinf. 2007)
- *FSA (PLoS Comp. Bio. 2009)*
- *Infernal (Bioinf. 2009)*
- Etc.

## Phylogeny methods

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

***RAxML***: heuristic for large-scale ML optimization

# “Big” phylogenetic datasets

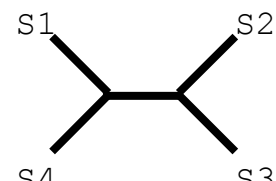
- Large numbers of taxa

# Simulation Studies

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

Unaligned  
Sequences

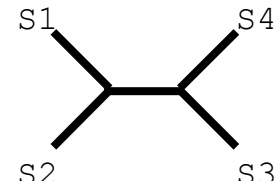
```
S1 = -AGGCTATCACCTGACCTCCA  
S2 = TAG-CTATCAC--GACCGC--  
S3 = TAG-CT-----GACCGC--  
S4 = -----TCAC--GACCGACA
```



A phylogenetic tree diagram showing the relationships between four sequences. The root is at the top, with two main branches. The left branch leads to a node that splits into S1 (top) and S4 (bottom). The right branch leads to a node that splits into S2 (top) and S3 (bottom).

True tree and  
alignment

```
S1 = -AGGCTATCACCTGACCTCCA  
S2 = TAG-CTATCAC--GACCGC--  
S3 = TAG-C--T-----GACCGC--  
S4 = T---C-A-CGACCGA-----CA
```

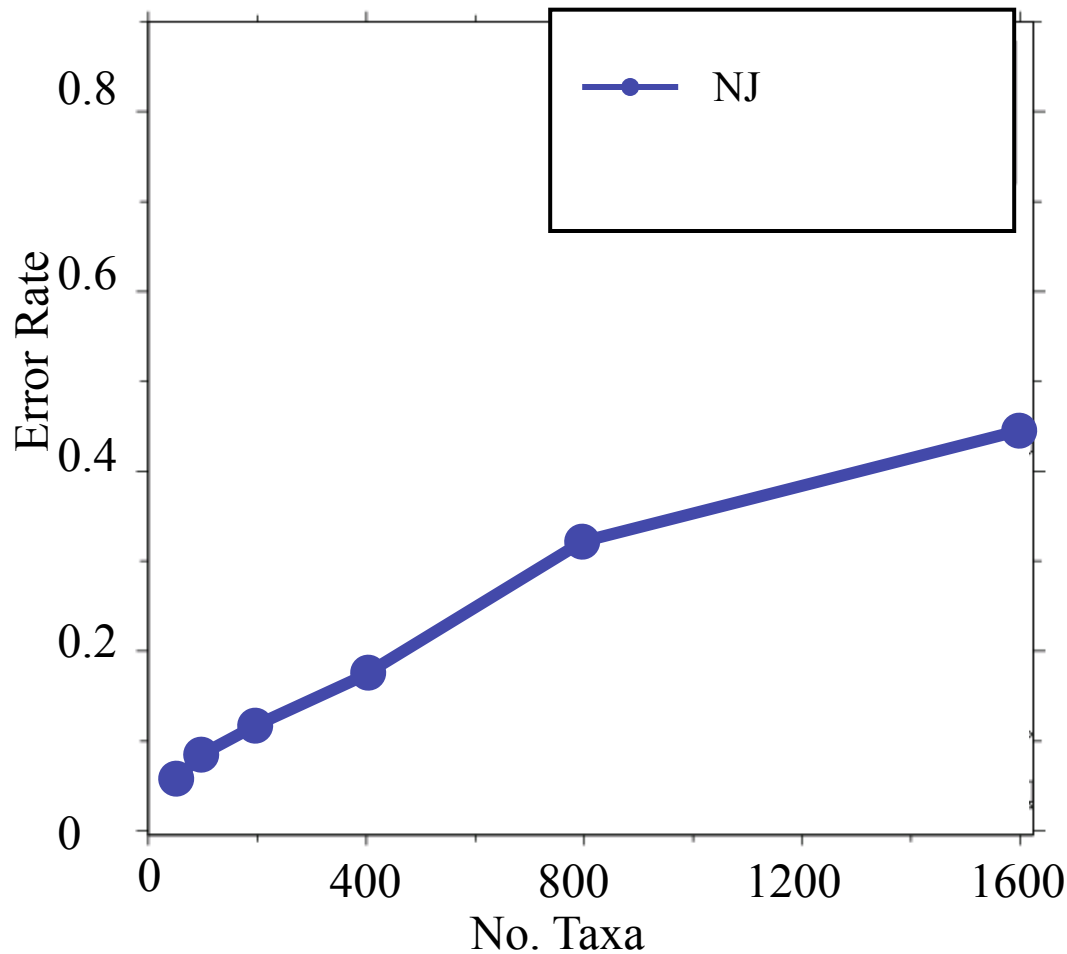


A phylogenetic tree diagram showing the estimated relationships between four sequences. The root is at the top, with two main branches. The left branch leads to a node that splits into S1 (top) and S2 (bottom). The right branch leads to a node that splits into S4 (top) and S3 (bottom).

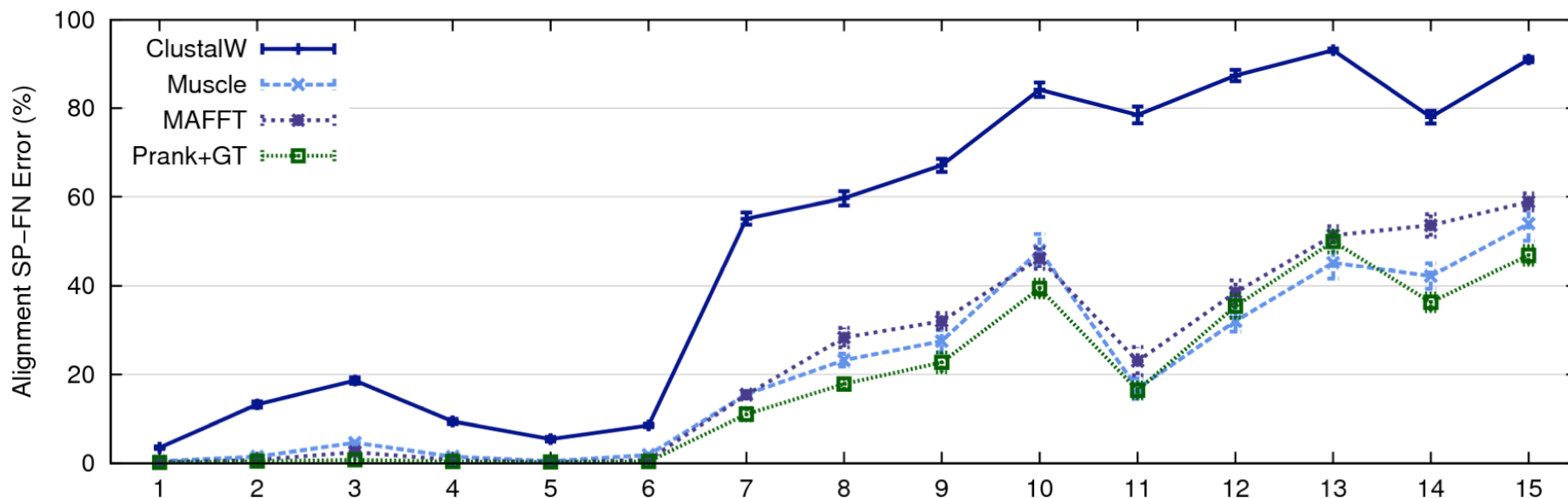
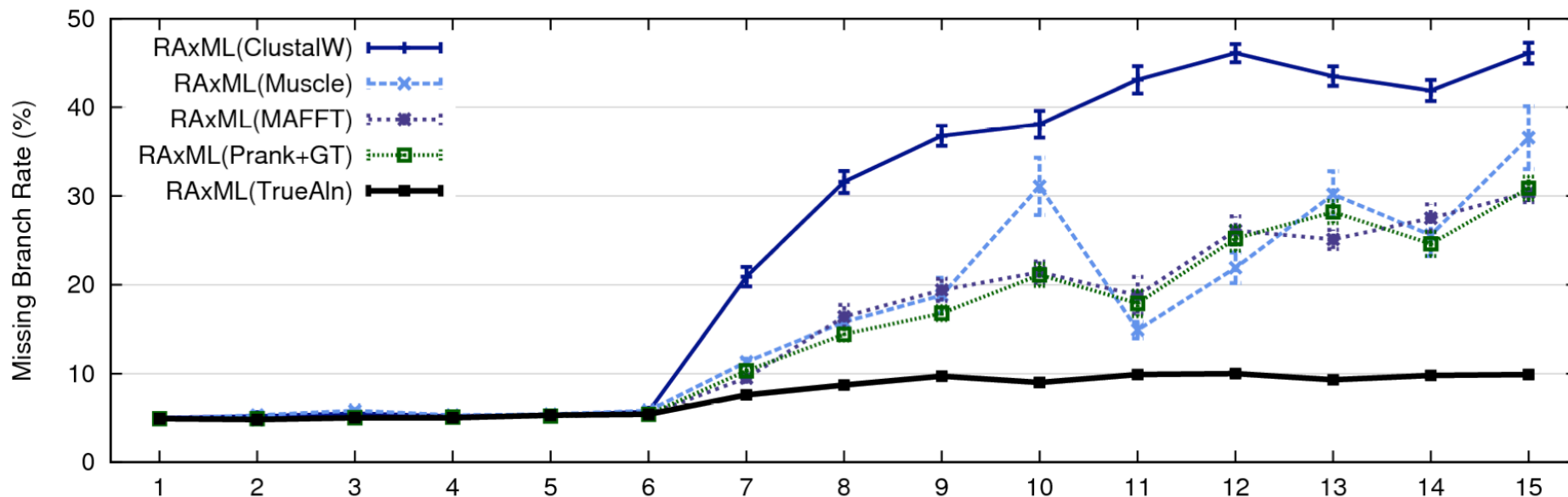
Estimated tree and  
alignment

Compare

# Neighbor joining has poor performance on large diameter trees [Nakhleh et al. ISMB 2001]



Theorem (Atteson):  
**Exponential**  
sequence length  
requirement for  
Neighbor Joining!



1000 taxon models, ordered by difficulty (Liu et al., 2009)

# “Big” phylogenetic datasets

- Large numbers of taxa
  - Accurate multiple sequence alignment is challenging, and has a large impact on phylogeny estimation
  - The best phylogeny estimation methods are heuristics for NP-hard problems (standard polynomial time methods can have poor accuracy, even on the true alignment)

# “Big” phylogenetic datasets

- Large numbers of genes

# “Big” phylogenetic datasets

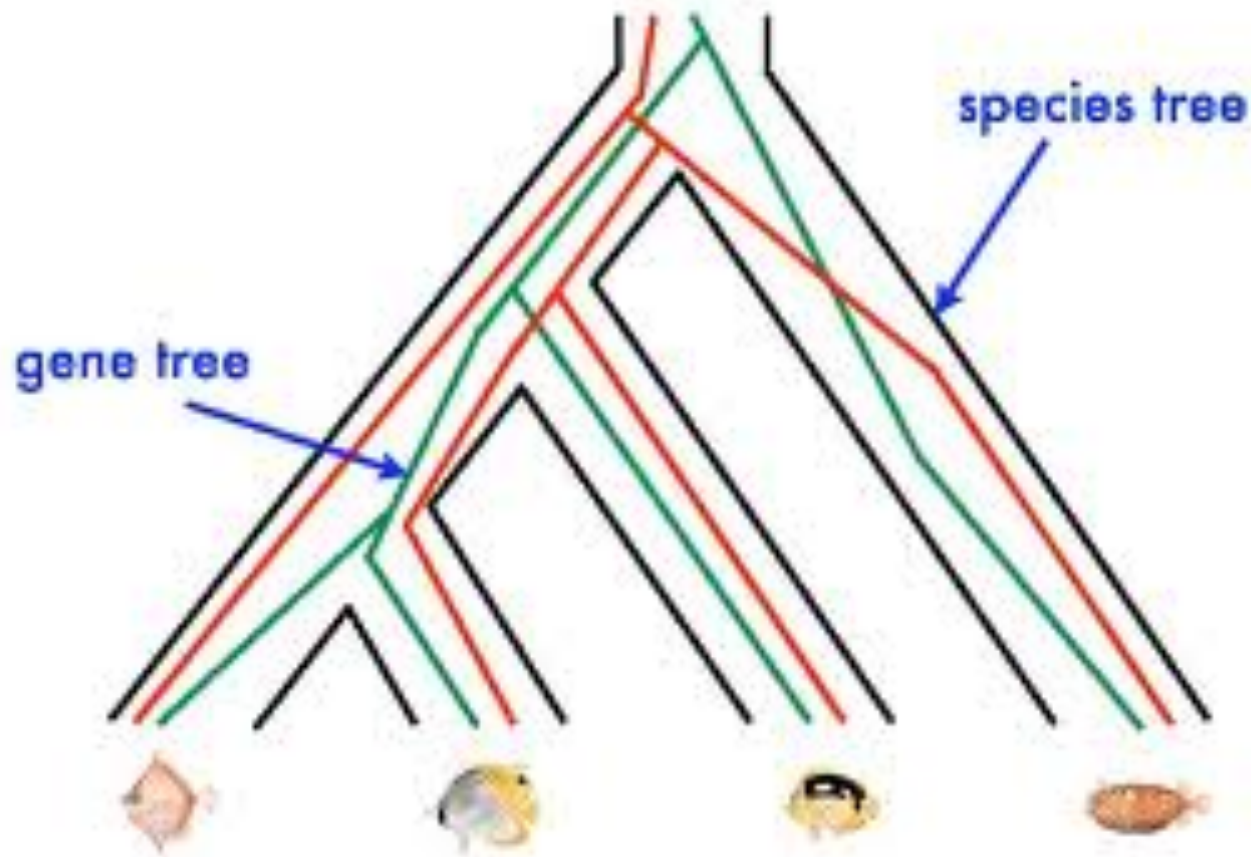
- Large numbers of genes
  - “Concatenation” can become computationally infeasible



# “Big” phylogenetic datasets

- Large numbers of genes
  - “Concatenation” can become computationally infeasible
  - Gene tree incongruence can make accurate species tree estimation challenging

**Red gene tree  $\neq$  species tree  
(green gene tree okay)**



# 1kp (<http://www.onekp.com/>)



Gane Ka-Shu  
Wong  
U Alberta



Jim  
Leebens-Mack  
U Georgia



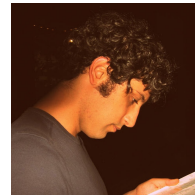
Norm  
Wickett  
Northwestern



Naim Matasci  
iPlant – U Arizona



Tandy Warnow,



Siavash Mirarab,  
UT-Austin



Nam Nguyen, and



Md. S. Bayzid

- Transcriptomes of approx. 1200 species
- More than 13,000 gene families (most not single copy)
- Multi-institutional project (10+ universities)

## Challenges:

Estimating very large gene alignments and trees (100,000+ sequences)

Estimating species trees from incongruent gene trees

# Avian Phylogenomics Project

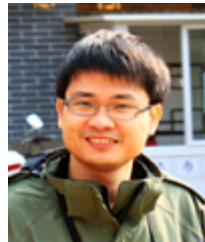
E. Jarvis,  
HHMI



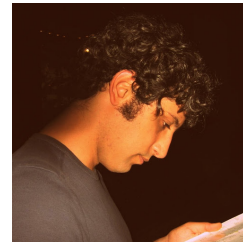
MTP Gilbert,  
Copenhagen



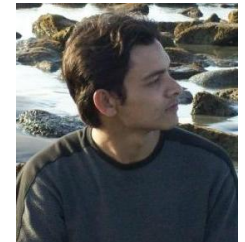
G. Zhang,  
BGI



S. Mirarab,



T. Warnow, and Md. S. Bayzid,  
UT-Austin



- Approx. 50 species, whole genomes
- 8000+ genes, UCEs
- Gene trees and sequence alignments computed using SATé
- Species tree estimated using maximum likelihood (RAxML)
- Multi-national team (20+ investigators)

Biggest challenges:

Estimating species tree from incongruent gene trees,  
Poor phylogenetic signal in most genes

# Major Challenges:

## large datasets, fragmentary sequences

- **Multiple sequence alignment:** Few methods can run on large datasets, and alignment accuracy is generally poor for large datasets with high rates of evolution.
- **Gene Tree Estimation:** standard methods have *poor accuracy* on even moderately large datasets, and the most accurate methods are enormously *computationally intensive* (weeks or months, high memory requirements).
- **Species Tree Estimation:** gene tree incongruence makes accurate estimation of species tree challenging.

Both phylogenetic estimation and multiple sequence alignment are also impacted by *fragmentary data*.

# This Talk

**SATé** - co-estimating trees and alignments  
(Science, 2009)

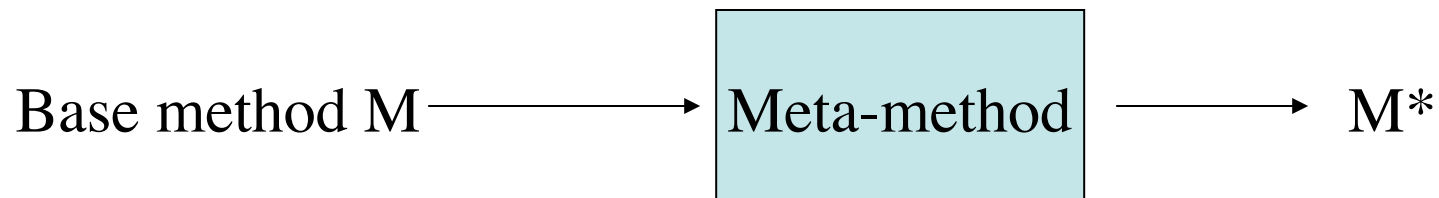
**DACTAL** – estimating trees (almost) without  
alignments (ISMB 2012)

**SEPP** - phylogenetic placement of **fragmentary**  
sequence data (e.g., short reads) (PSB 2012)

**UPP** - Ultra-large alignment using SEPP  
(unpublished)

# Meta-Methods

- Meta-methods “boost” the performance of base methods (phylogeny reconstruction, alignment estimation, etc).



# Part I: SATé

Simultaneous Alignment and Tree Estimation

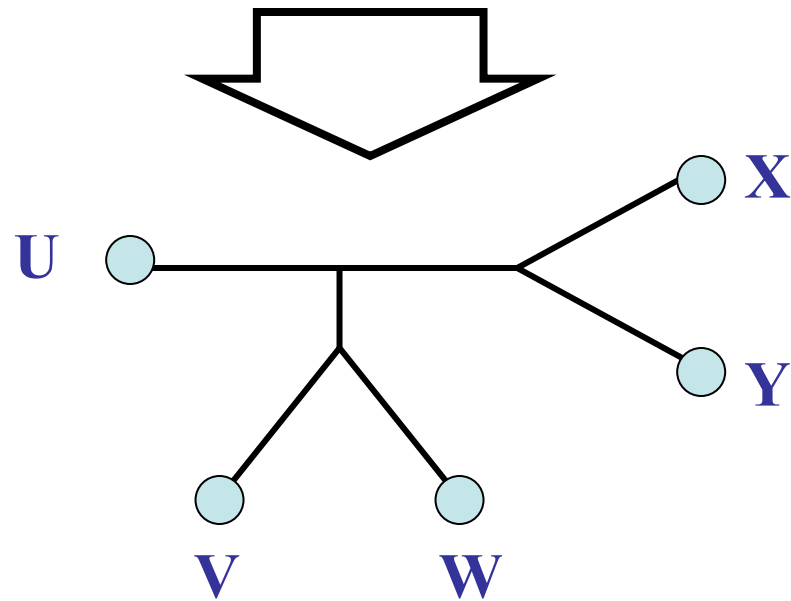
Liu, Nelesen, Raghavan, Linder, and Warnow,  
*Science*, 19 June 2009, pp. 1561-1564.

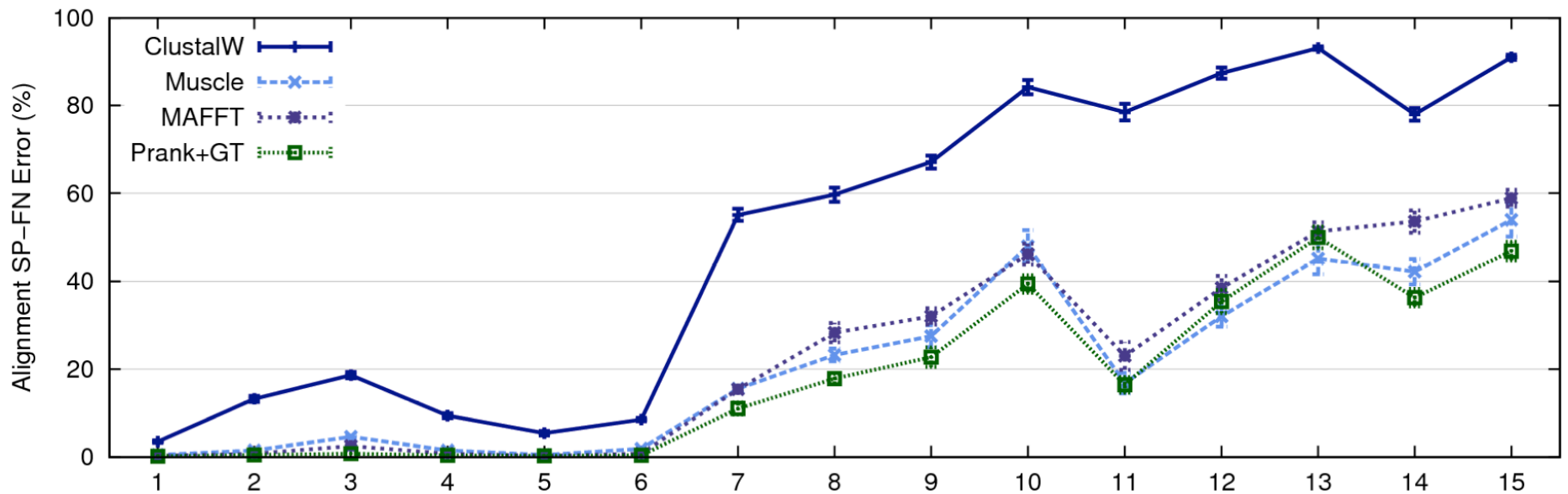
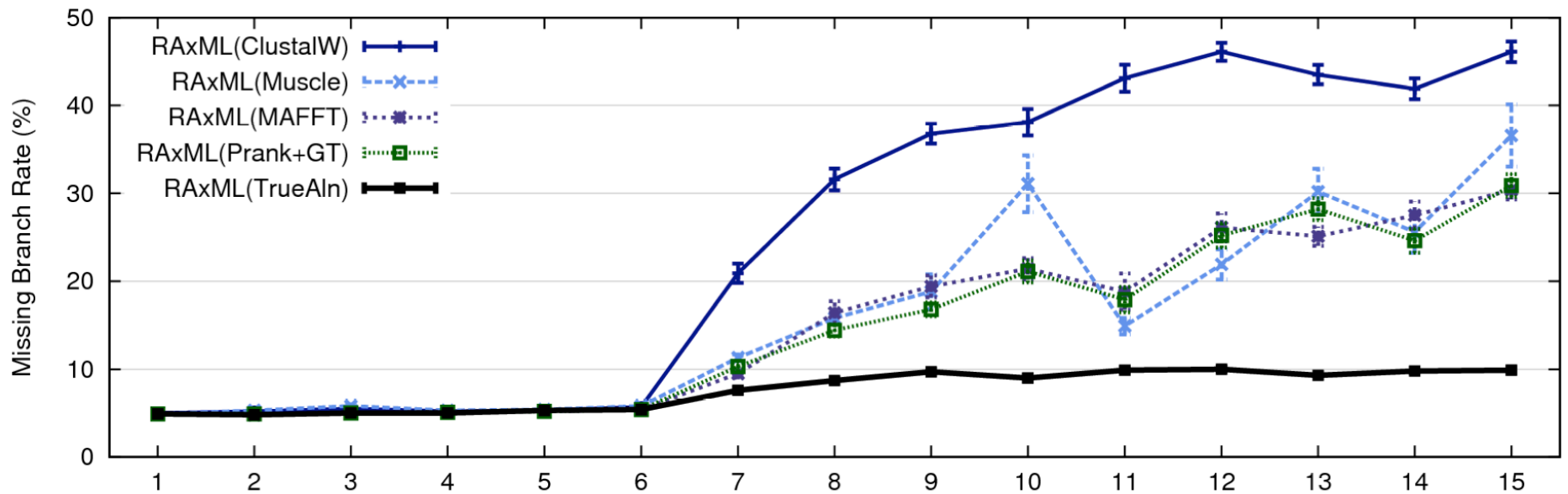
Liu et al., *Systematic Biology* 2012

Public software distribution (open source)  
through Mark Holder's group at the University  
of Kansas



U AGGGCATGA      V AGAT      W TAGACTT      X TGCACAA      Y TGCGCTT

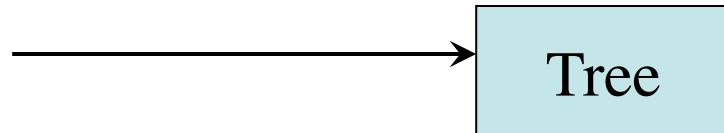




1000 taxon models, ordered by difficulty (Liu et al., 2009)

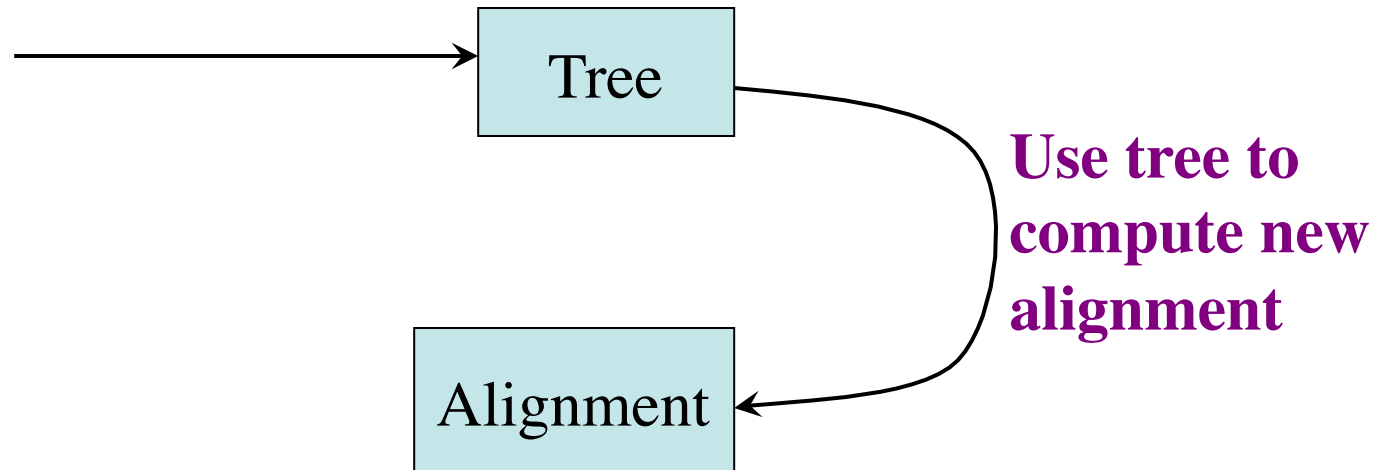
# SATé Algorithm

Obtain initial alignment  
and estimated ML tree



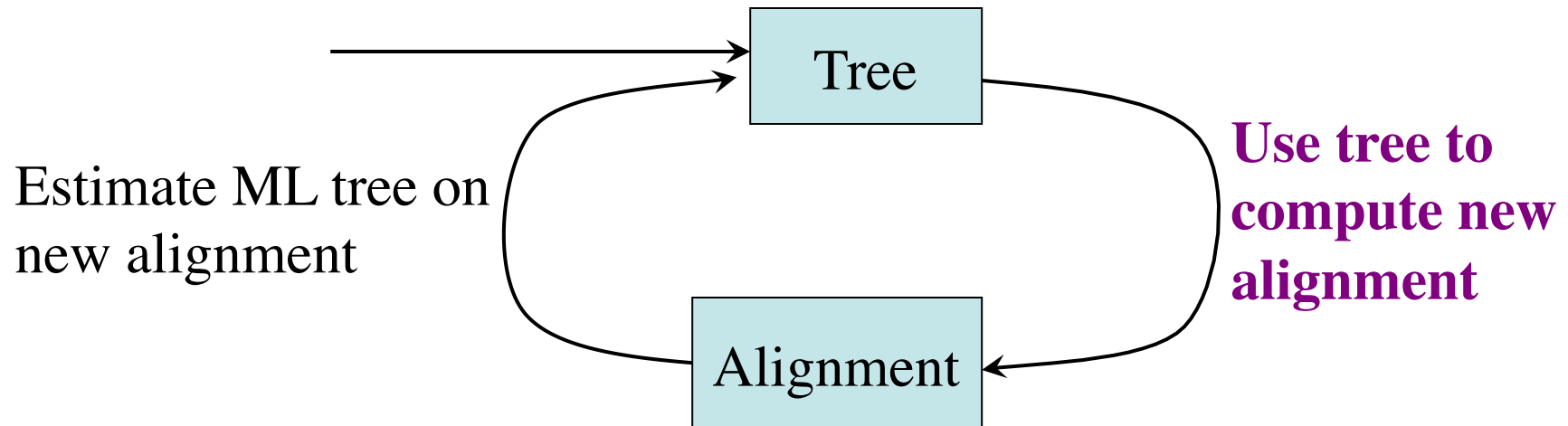
# SATé Algorithm

Obtain initial alignment  
and estimated ML tree

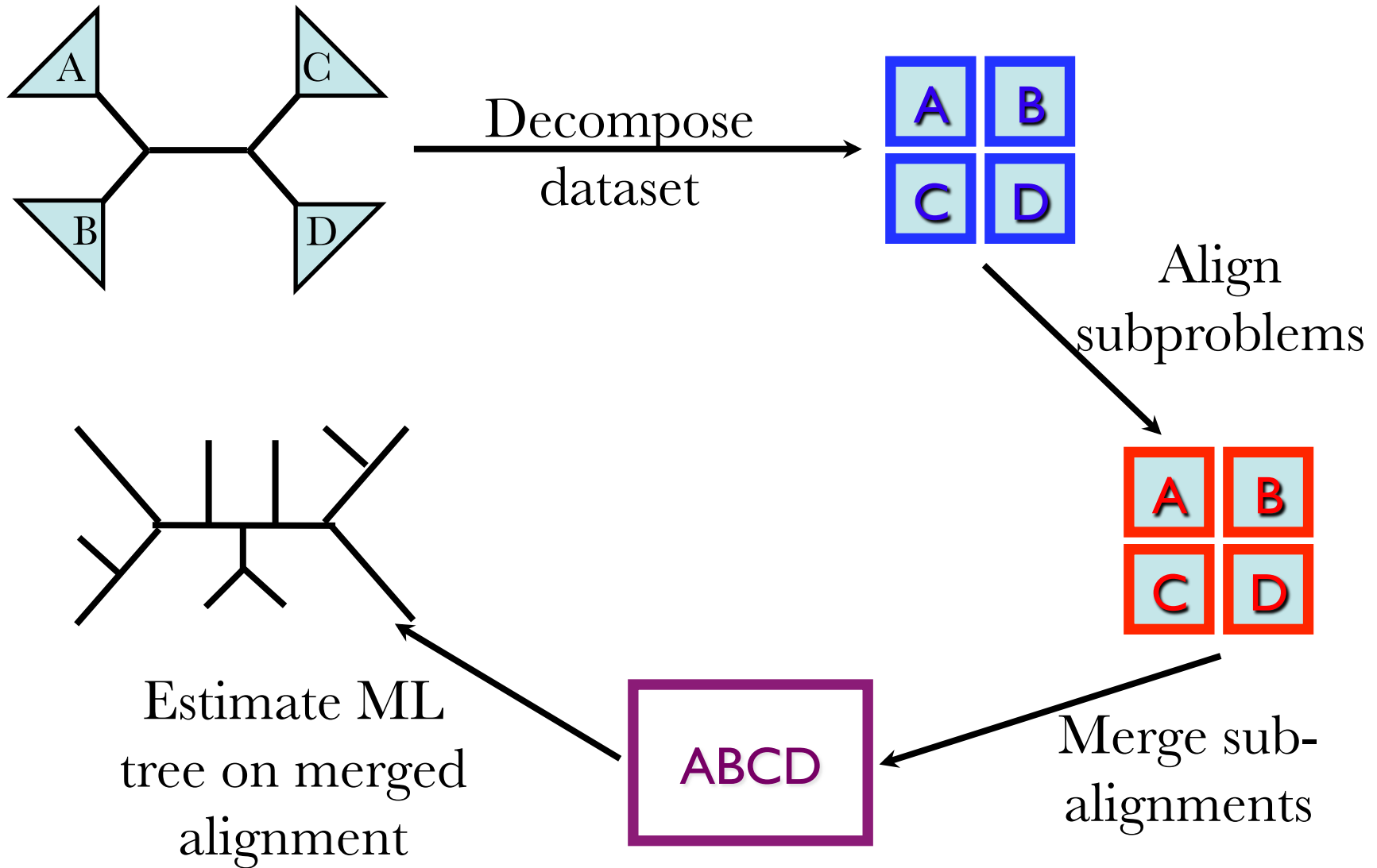


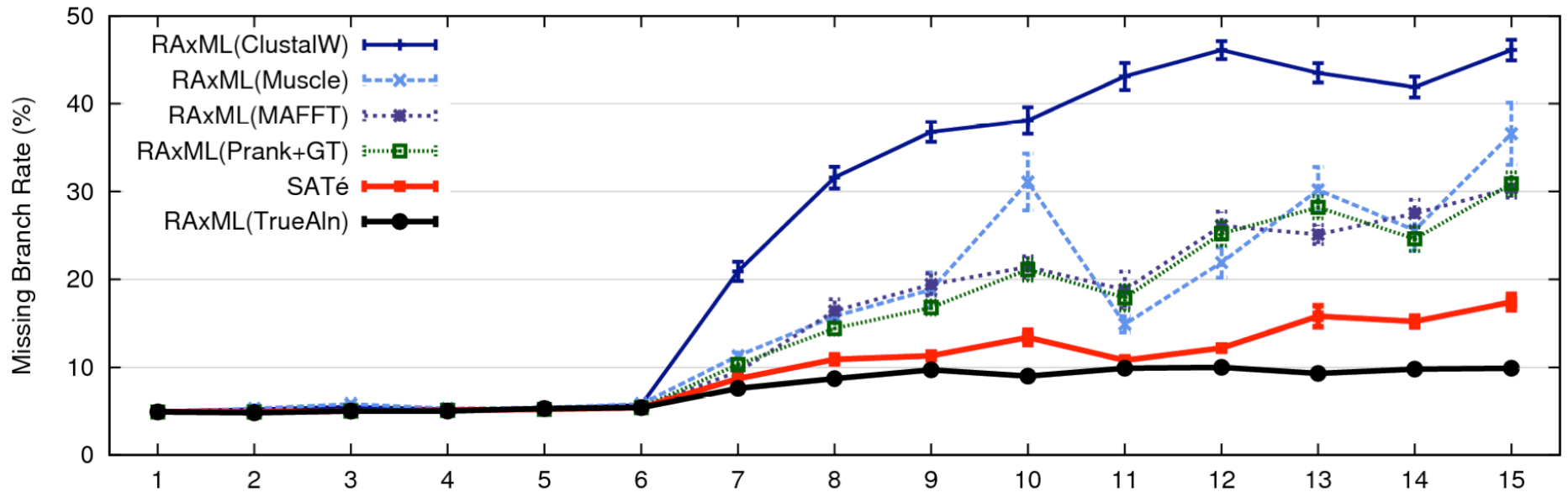
# SATé Algorithm

Obtain initial alignment  
and estimated ML tree



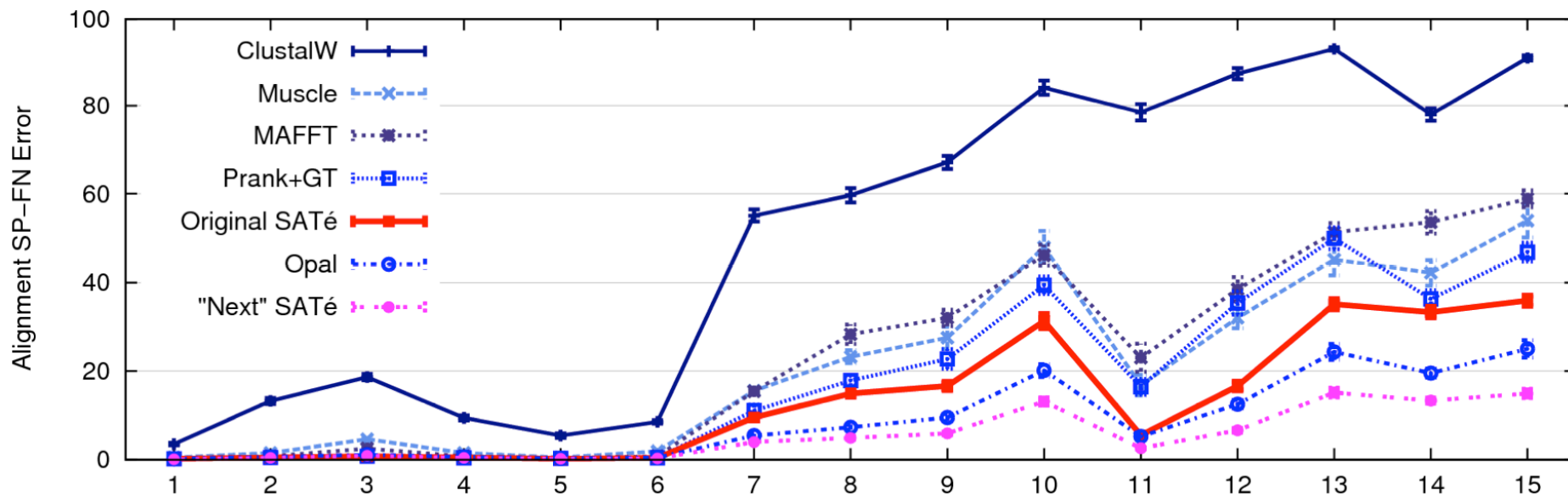
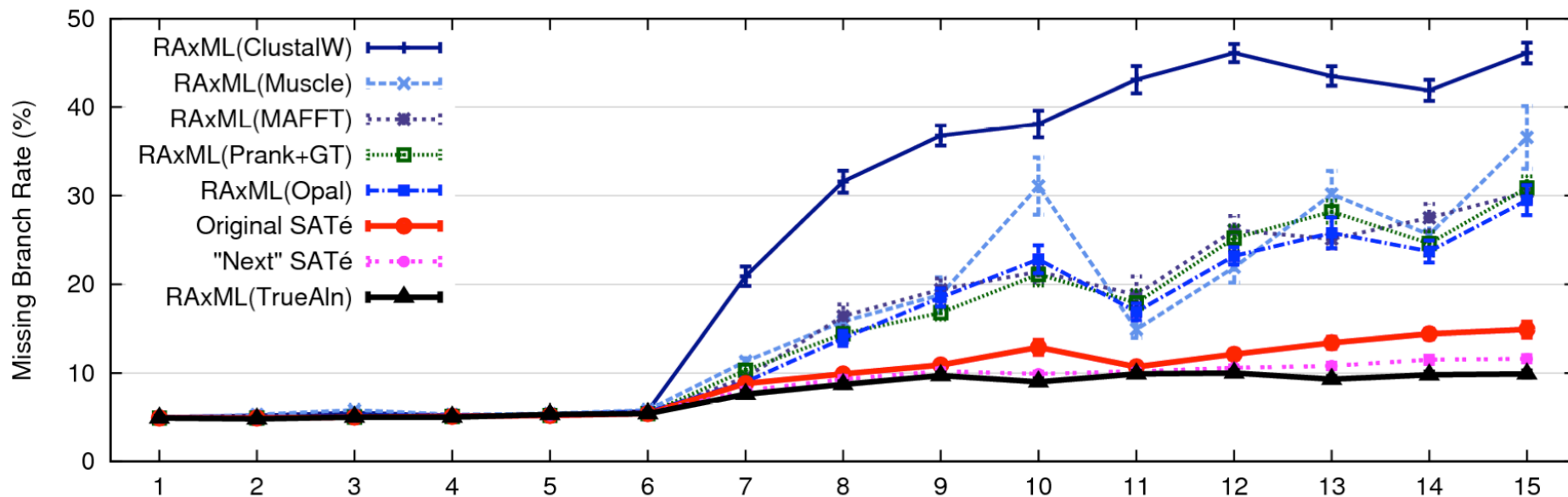
# Re-aligning on a tree





1000 taxon models, ordered by difficulty

24 hour SATé analysis, on desktop machines  
 (Similar improvements for biological datasets)



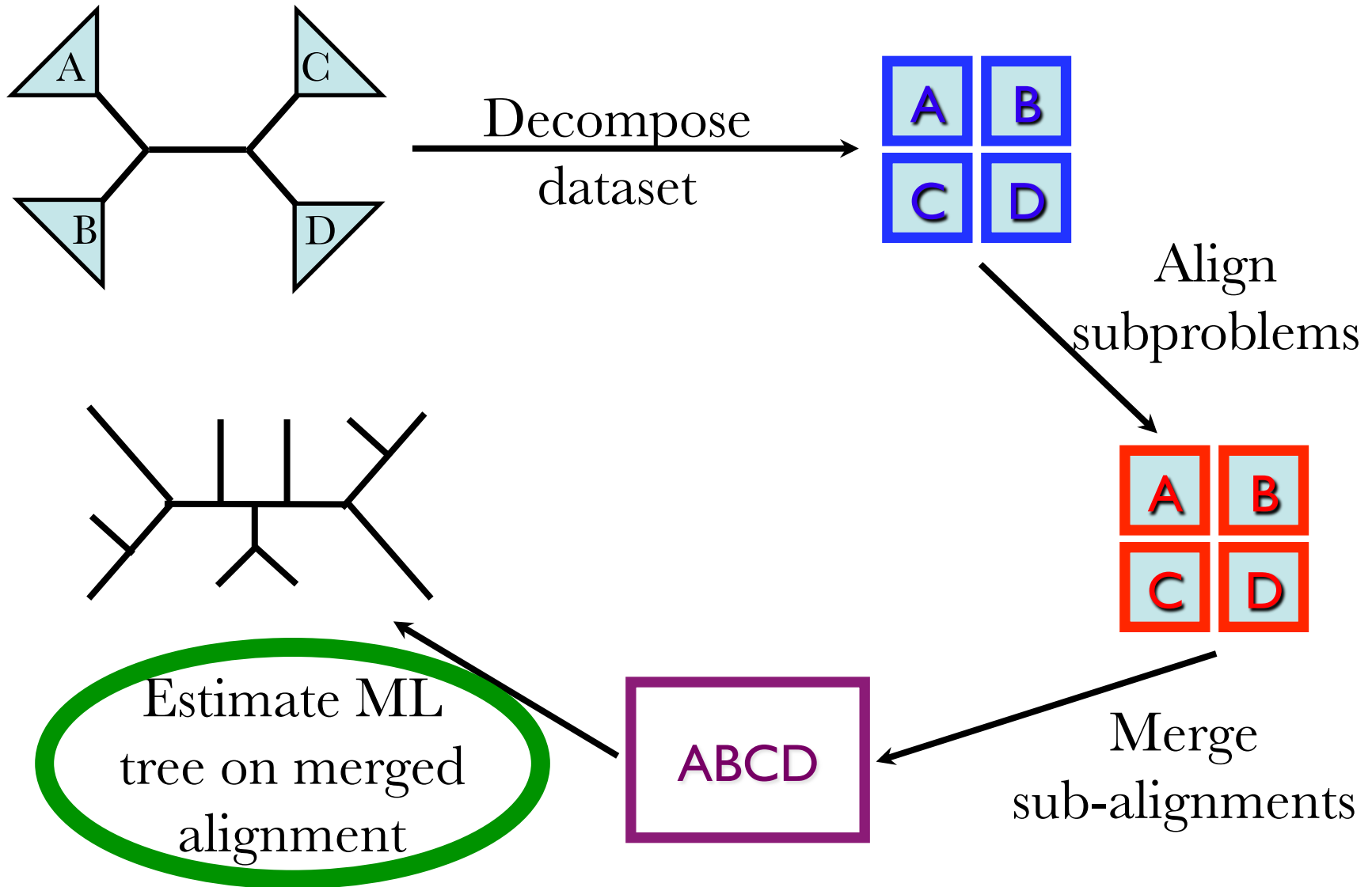
1000 taxon models ranked by difficulty



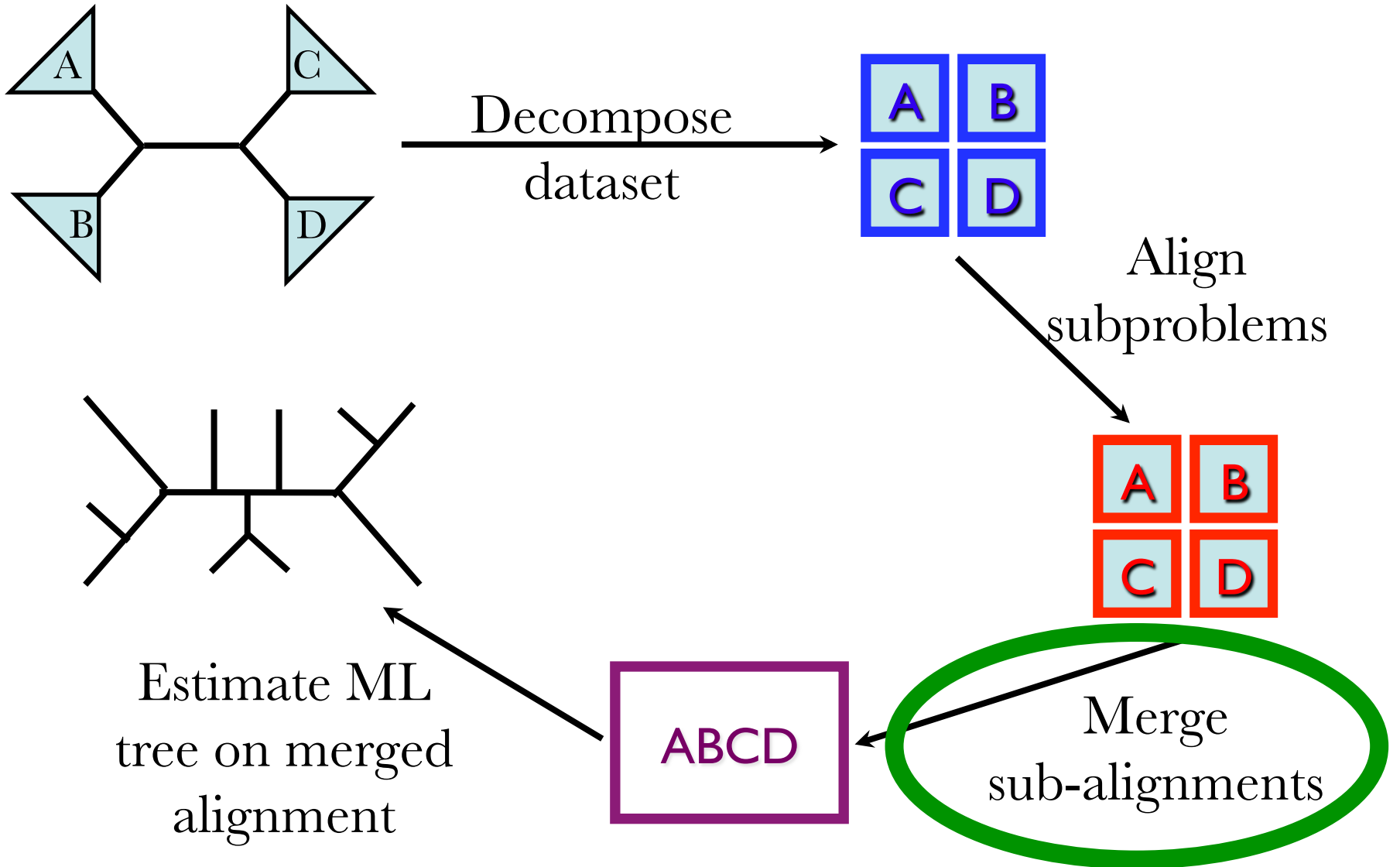
# Brief discussion

- **SATé “boosts” the base methods.** Results shown are for SATé used with MAFFT and Muscle. Similar improvements seen for use with Prank, Opal, Muscle, ClustalW, etc.
- **Biological datasets:** Similar results on large benchmark datasets (structurally-based rRNA alignments)
- **No statistical guarantees!!!** In fact, it’s all bad news: ML, treating gaps as missing data (even given the true alignment), can be inconsistent!
- **Performance in practice** results from use of base methods (and ability to use best versions of base methods).
- *Alignment of genome-scale sequences is a different problem.*
- *SATé is designed for full-length sequences, not fragmentary datasets*

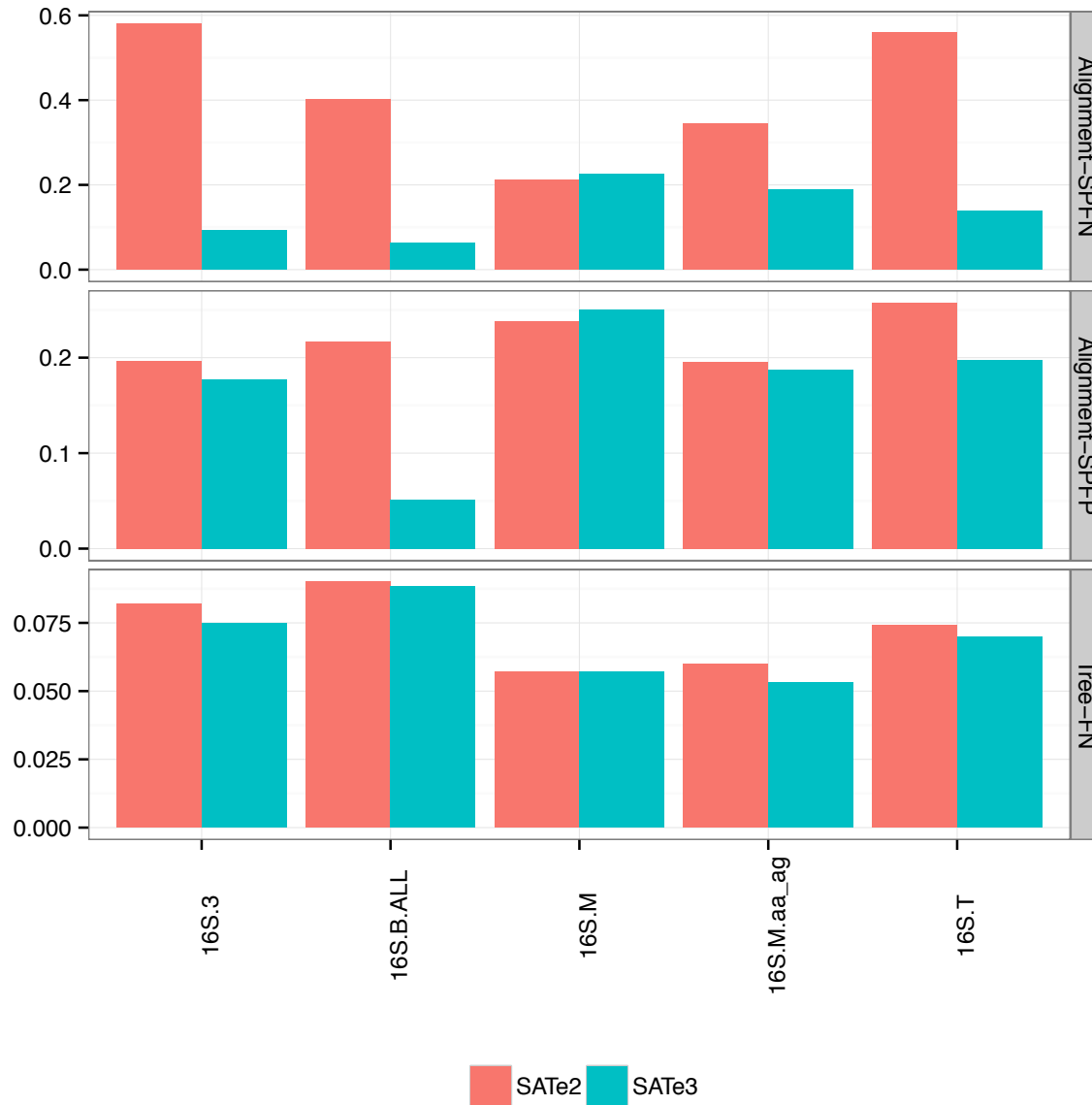
# Limitations



# Limitations



# SATe-3 improvement over SATe-2



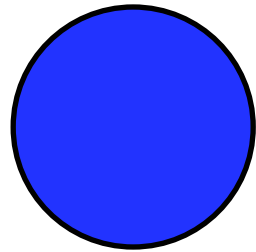
# **Part II: DACTAL**

## **Divide-And-Conquer Trees (Almost) without alignments**

- Input: set  $S$  of unaligned sequences
- Output: tree on  $S$  (but no alignment)

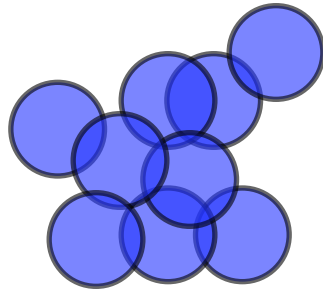
Nelesen, Liu, Wang, Linder, and Warnow,  
ISMB 2012 and Bioinformatics 2012

# DACTAL



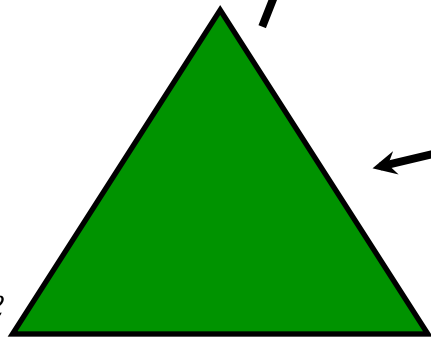
*Unaligned  
Sequences*

**BLAST-  
based**



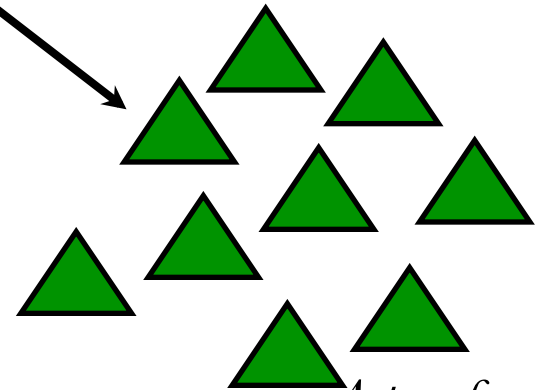
*Overlapping  
subsets*

**pRecDCM3**



*A tree for the  
entire dataset*

Existing Method:  
RAxML(MAFFT)



*A tree for each  
subset*

New supertree method:  
**SuperFine**

# Average of 3 Largest CRW Datasets

CRW: Comparative RNA database,  
Three 16S datasets with **6,323** to **27,643**  
sequences

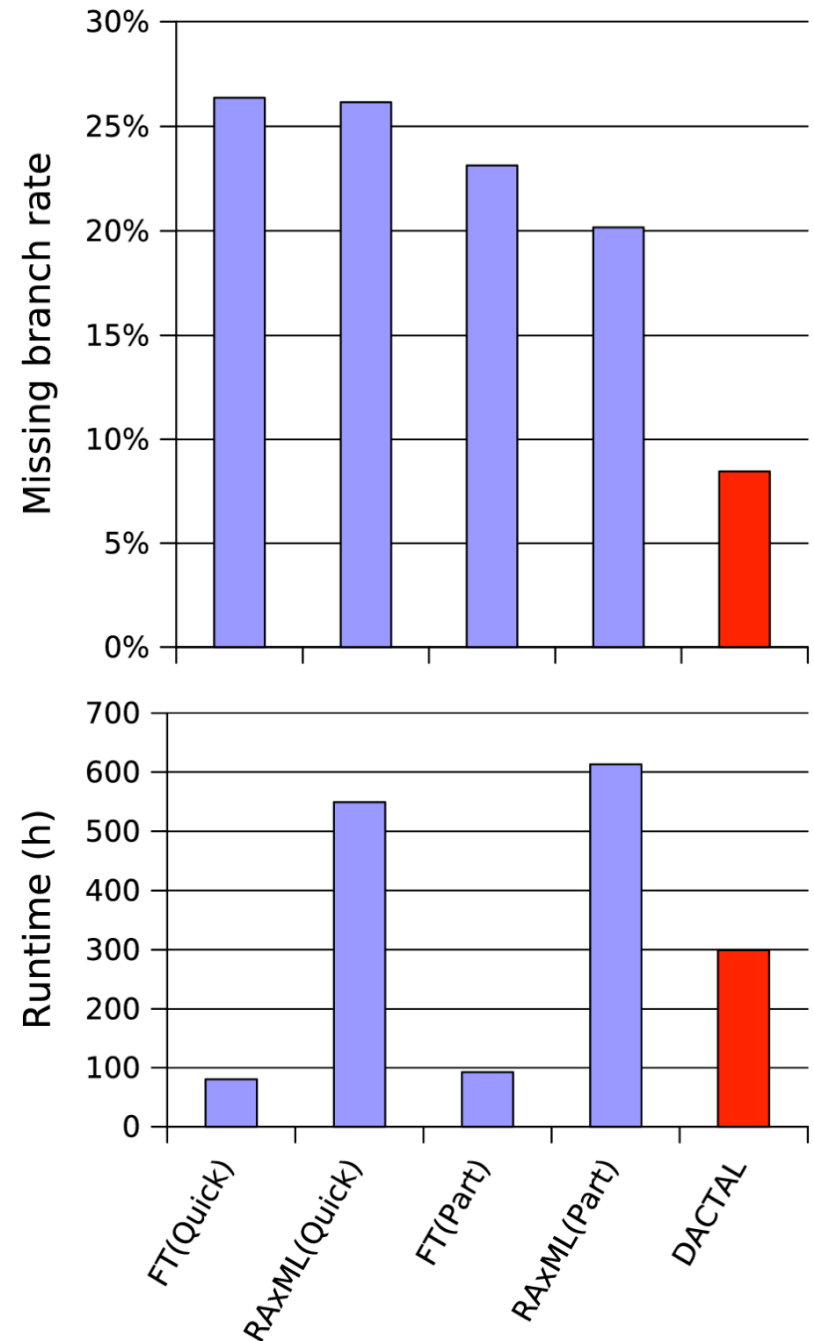
Reference alignments based on  
secondary structure

Reference trees are 75% RAxML  
bootstrap trees

DACTAL (shown in red) run for 5  
iterations starting from FT(Part)

FastTree (FT) and RAxML are ML  
methods

DACTAL and SATe have comparable accuracy



# Part III: SEPP

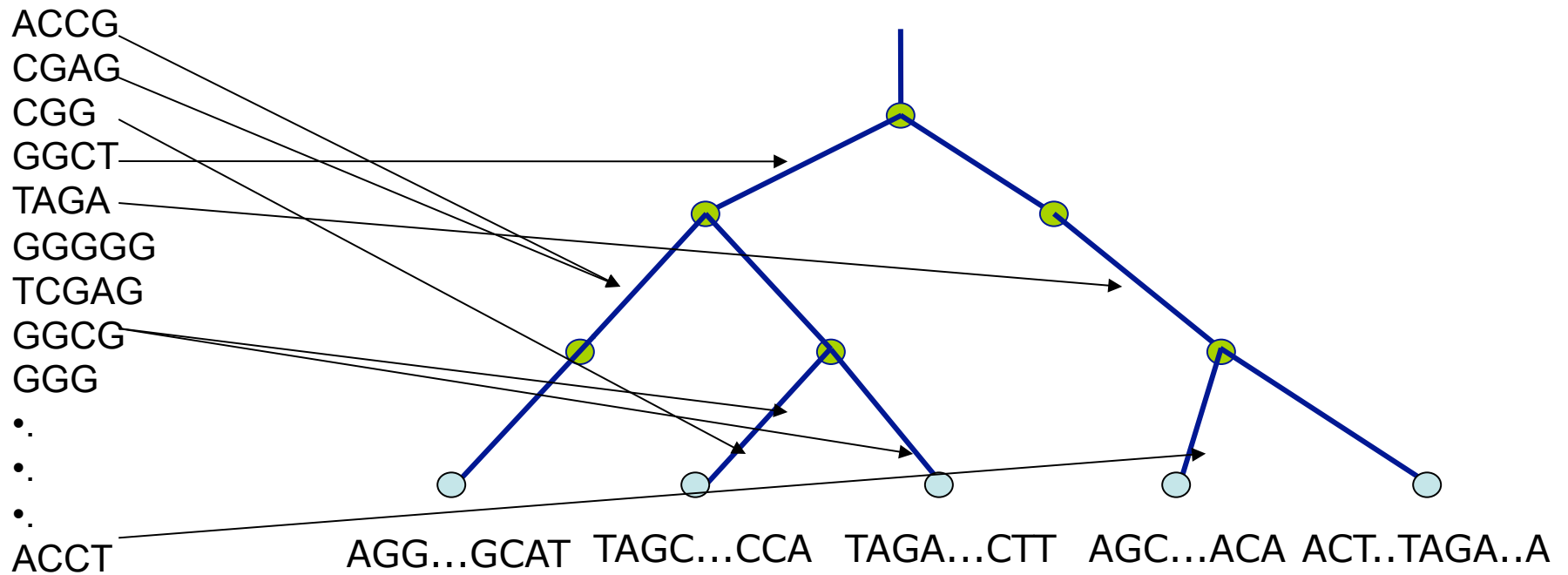
- **SEPP: SATé-enabled Phylogenetic Placement**, by Mirarab, Nguyen, and Warnow
- Pacific Symposium on Biocomputing, 2012 (special session on the Human Microbiome)
- Objective: phylogenetic analysis of single-gene datasets with **fragmentary sequences**



# Phylogenetic Placement

Fragmentary sequences  
from some gene

Full-length sequences for same  
gene, and an alignment and a tree



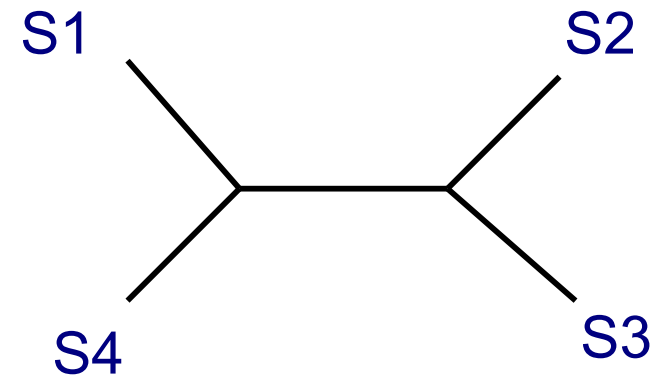
# Phylogenetic Placement

Step 1: Align each query sequence to backbone alignment

Step 2: Place each query sequence into backbone tree, using extended alignment

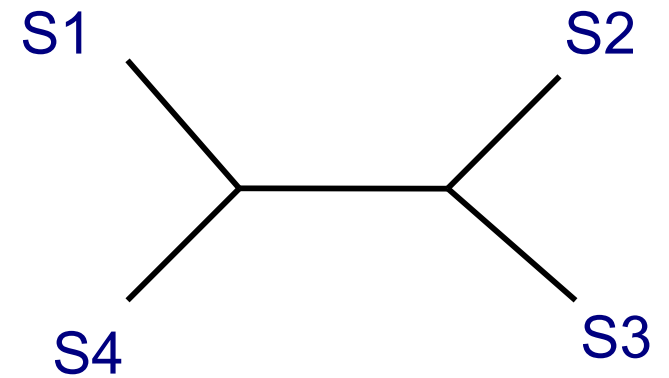
# Align Sequence

S1 = -AGGCTATCACCTGACCTCCA-AA  
S2 = TAG-CTATCAC--GACCGC--GCA  
S3 = TAG-CT-----GACCGC--GCT  
S4 = TAC----TCAC--GACCGACAGCT  
Q1 = TAAAAC



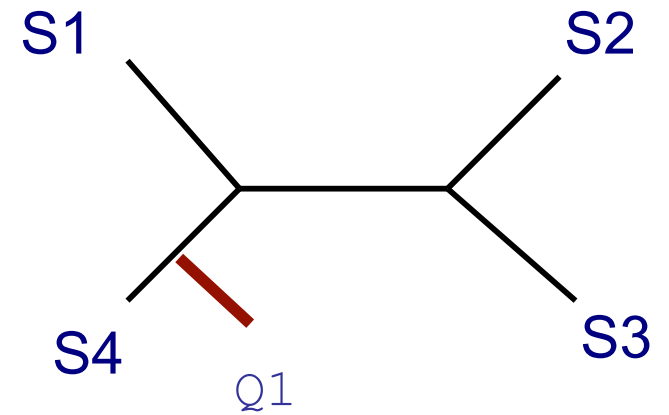
# Align Sequence

S1 = -AGGCTATCACCTGACCTCCA-AA  
S2 = TAG-CTATCAC--GACCGC--GCA  
S3 = TAG-CT-----GACCGC--GCT  
S4 = TAC----TCAC--GACCGACAGCT  
Q1 = -----T-A--AAAC-----



# Place Sequence

S1 = -AGGCTATCACCTGACCTCCA-AA  
S2 = TAG-CTATCAC--GACCGC--GCA  
S3 = TAG-CT-----GACCGC--GCT  
S4 = TAC----TCAC--GACCGACAGCT  
Q1 = -----T-A--AAAC-----

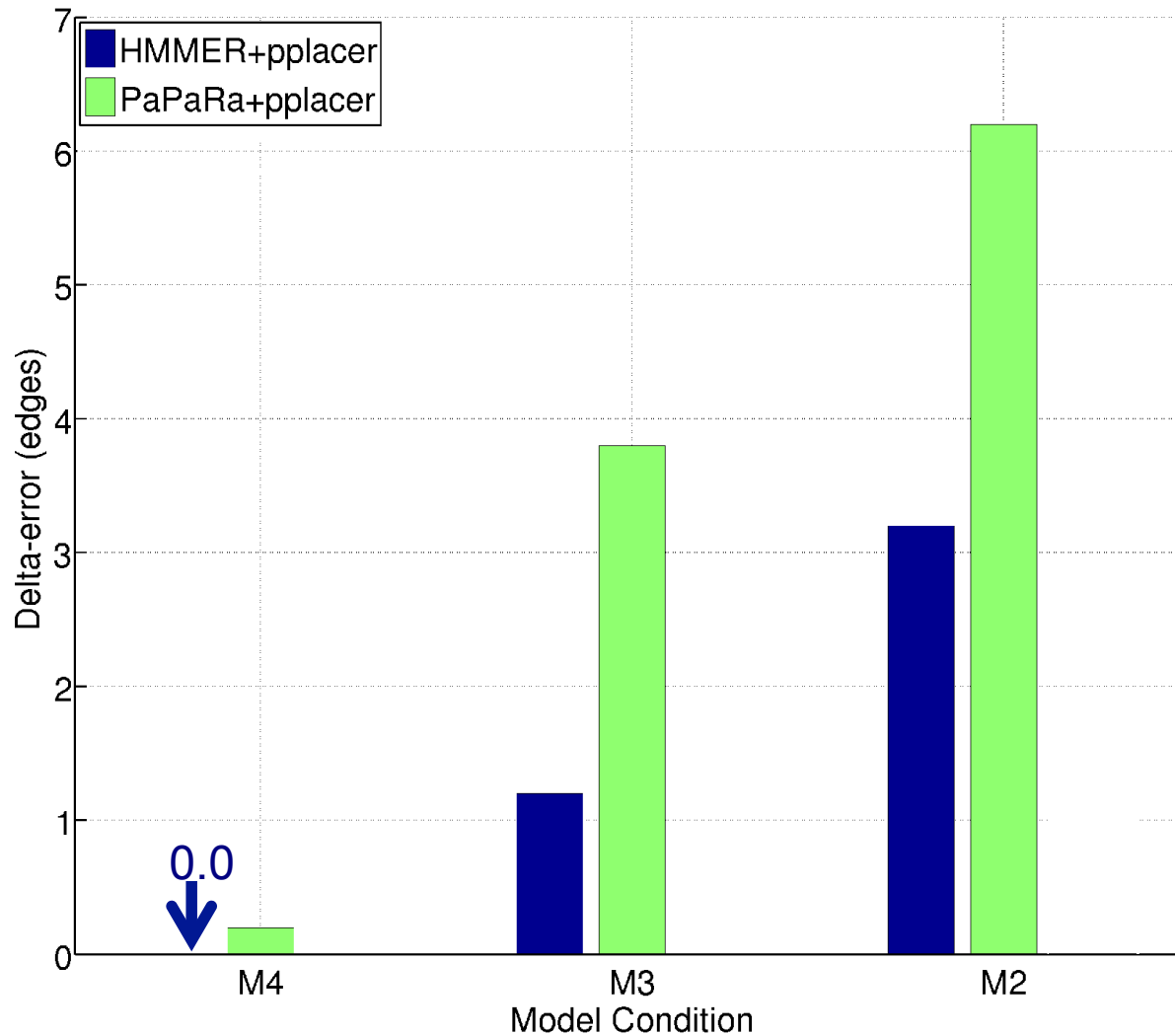


# Phylogenetic Placement

- Align each query sequence to backbone alignment
  - **HMMALIGN** (Eddy, Bioinformatics 1998)
  - **PaPaRa** (Berger and Stamatakis, Bioinformatics 2011)
- Place each query sequence into backbone tree
  - **Pplacer** (Matsen et al., BMC Bioinformatics, 2011)
  - EPA (Berger and Stamatakis, Systematic Biology 2011)

Note: pplacer and EPA use maximum likelihood, and are reported to have the same accuracy.

# HMMER vs. PaPaRa

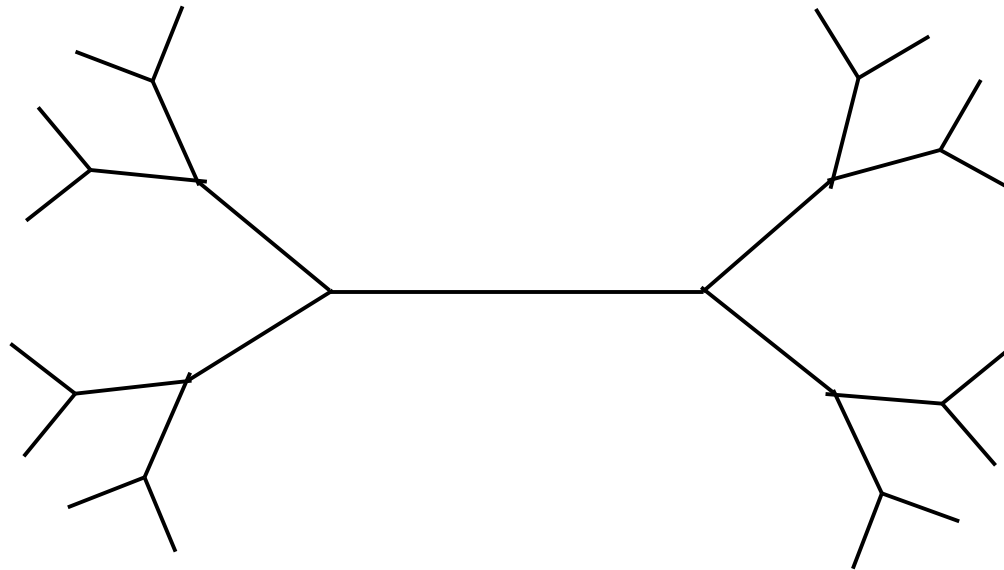


Increasing rate of evolution



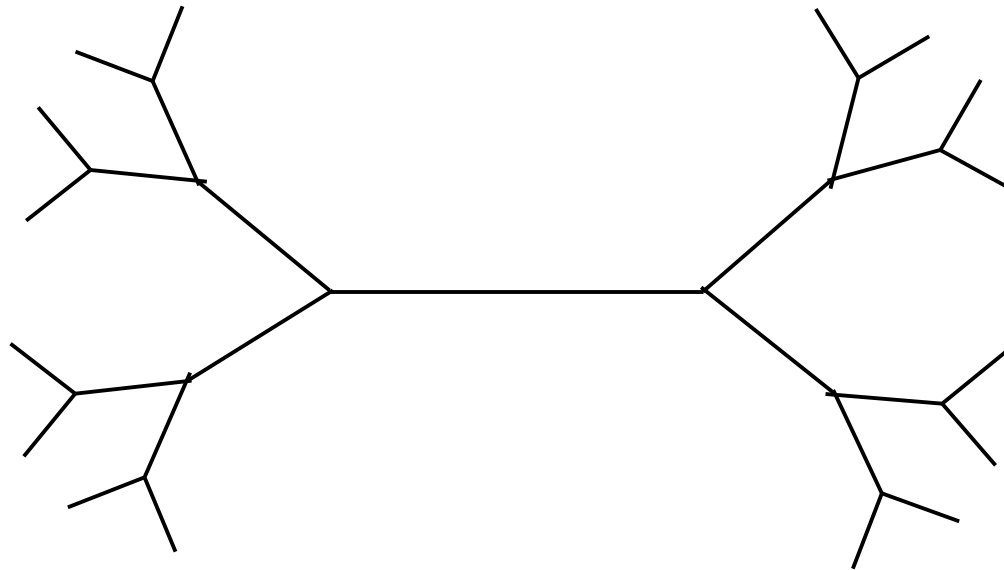
## HMMER+pplacer:

- 1) build one HMM for the entire alignment
- 2) Align fragment to the HMM, and insert into alignment
- 3) Insert fragment into tree to optimize likelihood

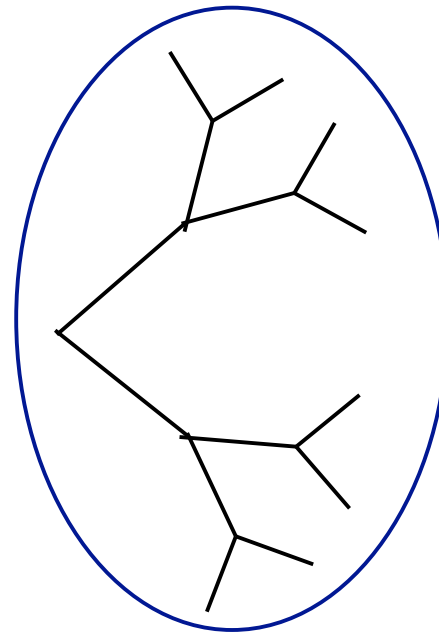
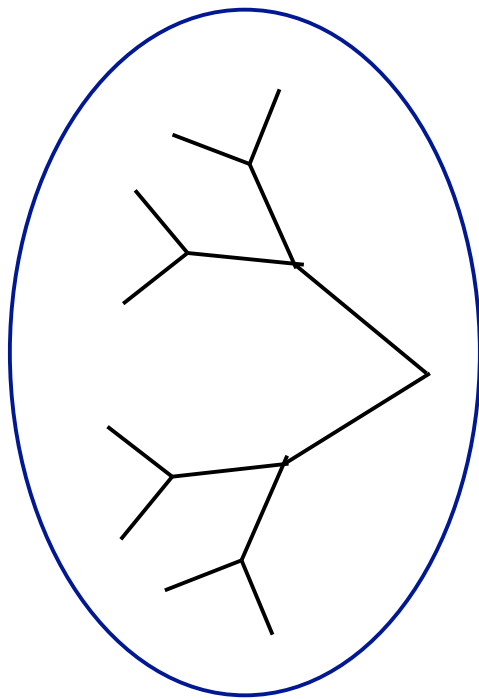




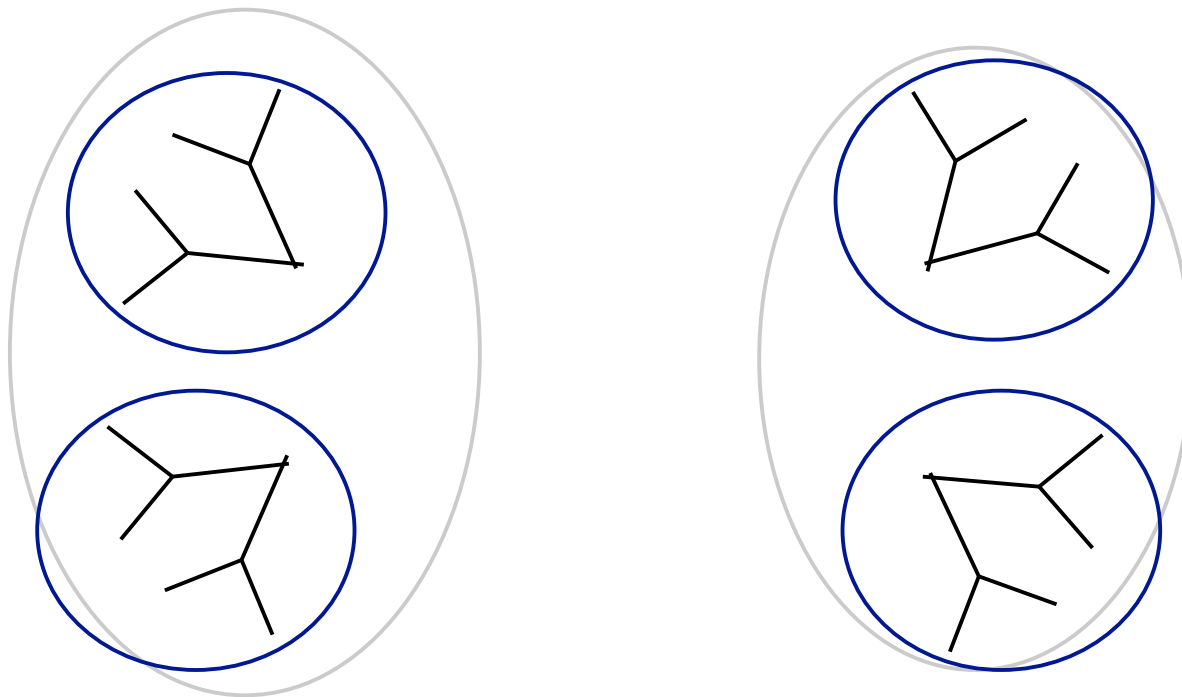
One Hidden Markov Model  
for the entire alignment?



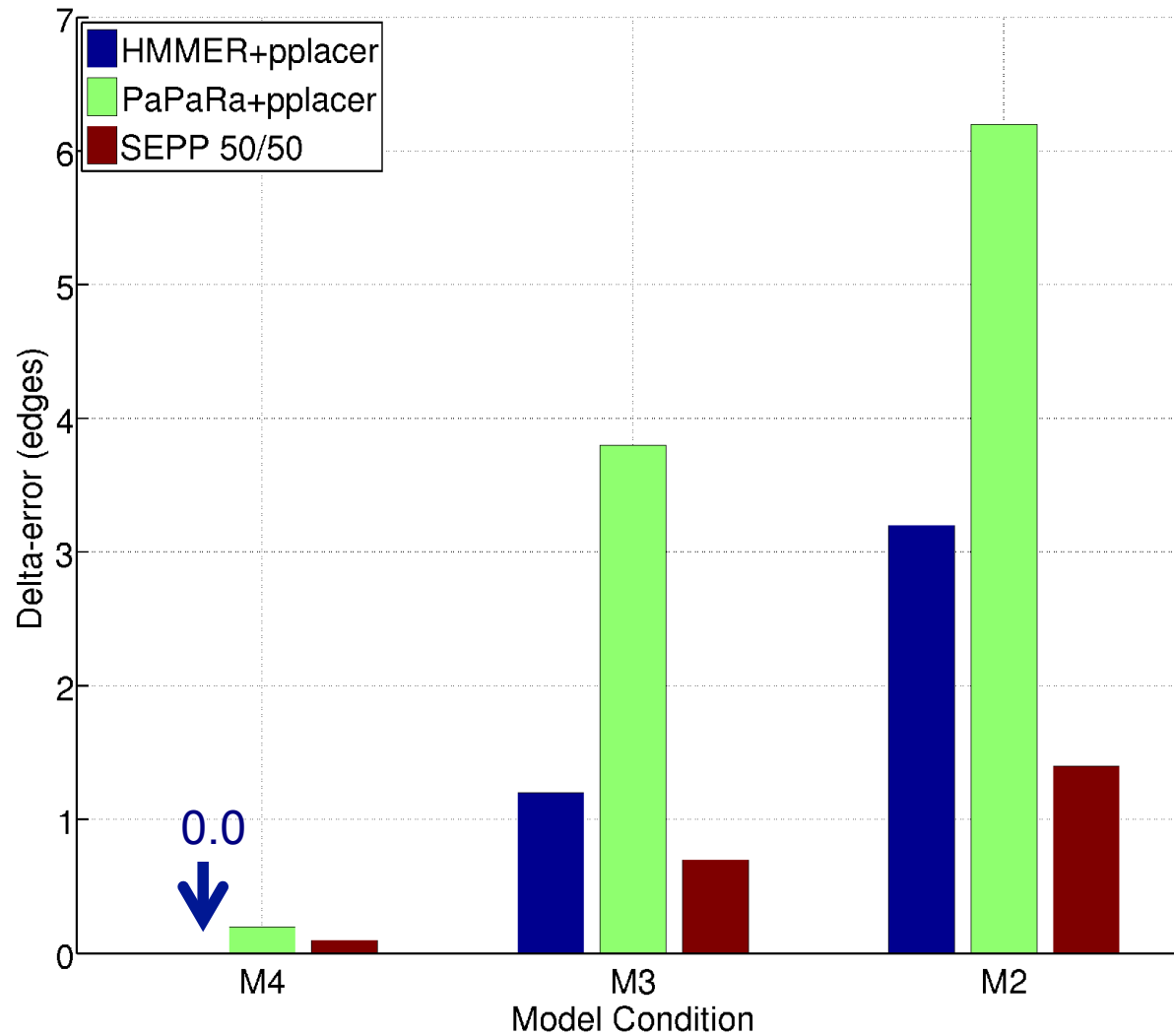
Or 2 HMMs?



Or 4 HMMs?



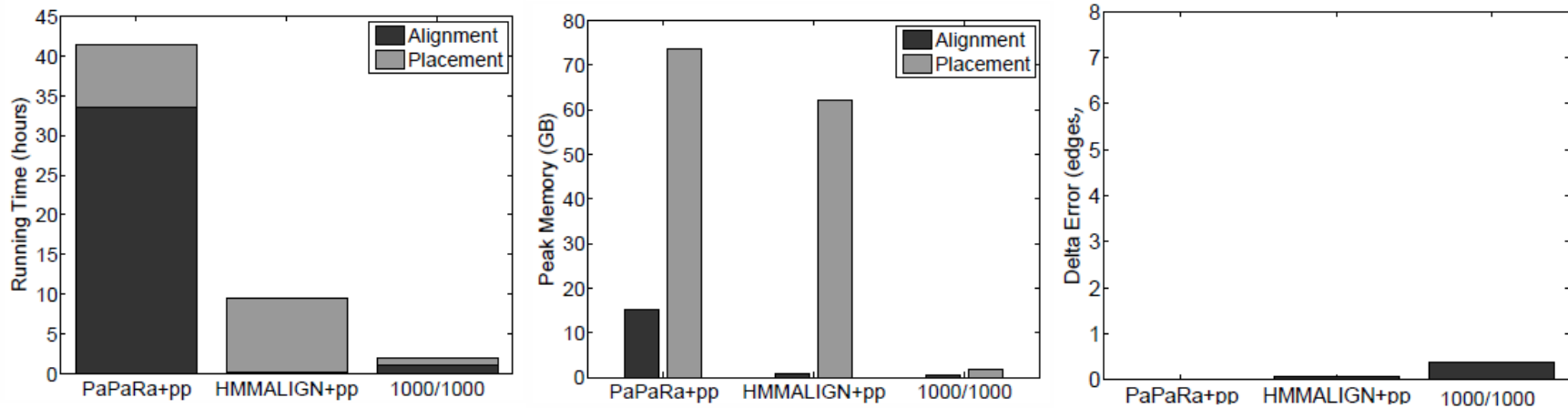
# SEPP(10%), based on ~10 HMMs



Increasing rate of evolution



# SEPP (10%) on Biological Data



16S.B.ALL dataset, 13k curated backbone tree, 13k total fragments

For 1 million fragments:

PaPaRa+pplacer: ~133 days

HMMALIGN+pplacer: ~30 days

SEPP 1000/1000: ~6 days

# Applications of SEPP (unpublished)

- **UPP**: Ultra-large alignment using SEPP
- **TIPP**: taxon identification of fragmentary data (for metagenomic analysis)

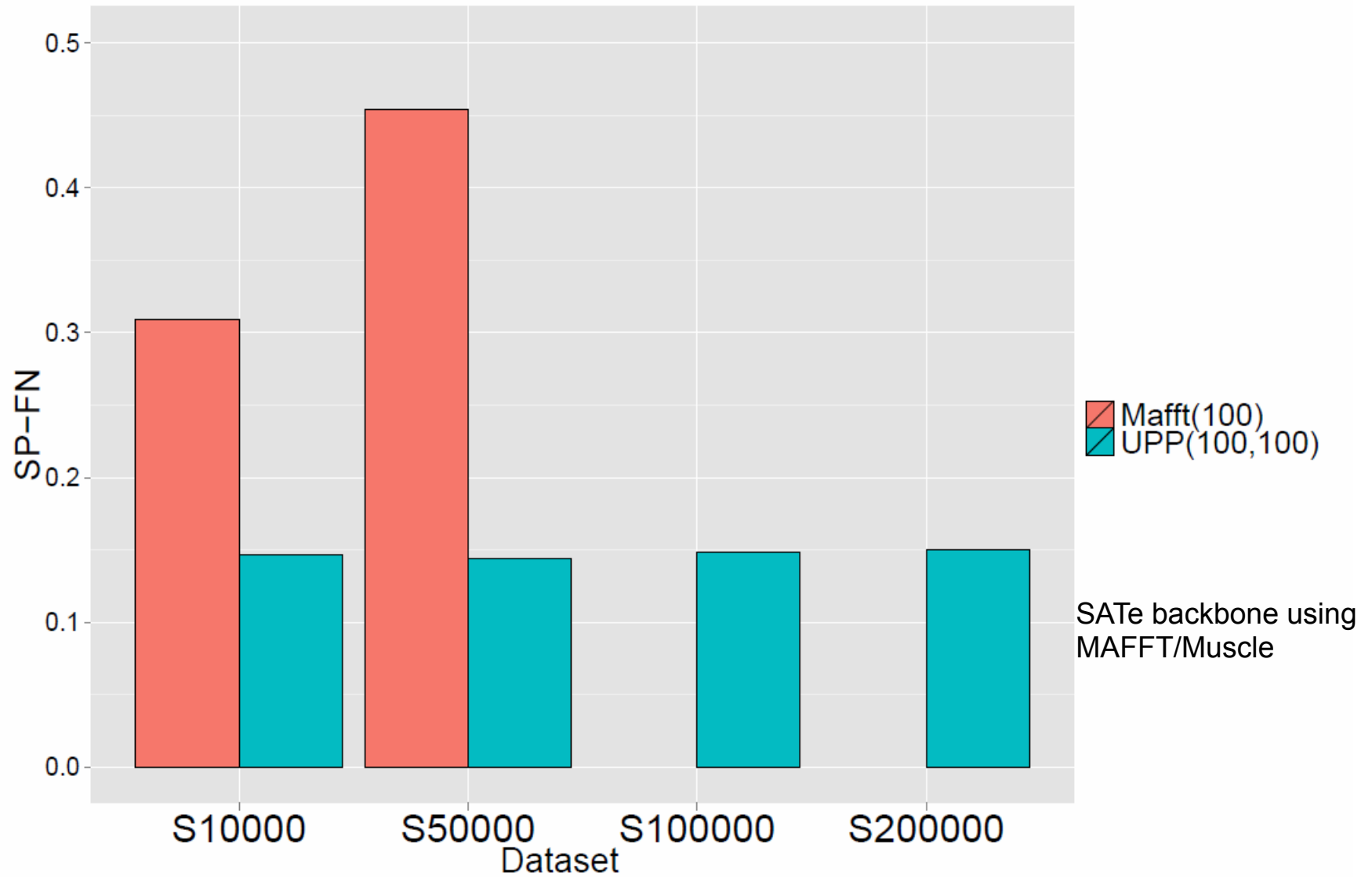
# Part IV: UPP: Ultra-large alignment using SEPP

Input: set  $S$  of unaligned sequences

Output: alignment and tree on  $S$

- Select random subset  $X$  of sequences
- Estimate alignment and tree on  $X$
- Run SEPP to align remaining sequences
- Run favorite tree estimation method on alignment
- UPP( $x,y$ ) refers to UPP using backbones of size  $y$  and alignment subsets of size  $x$

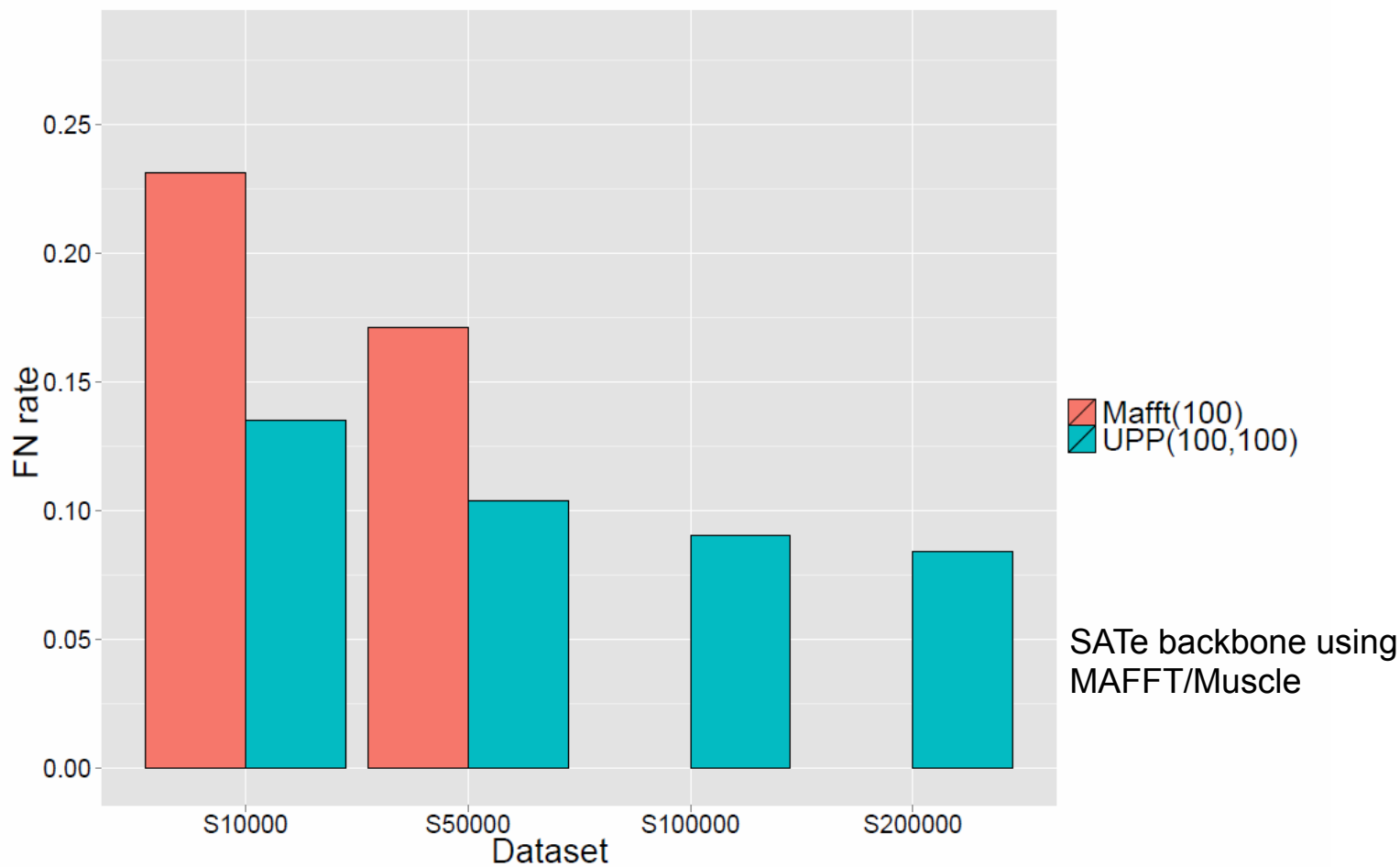
# RNASim: SP-FN Score



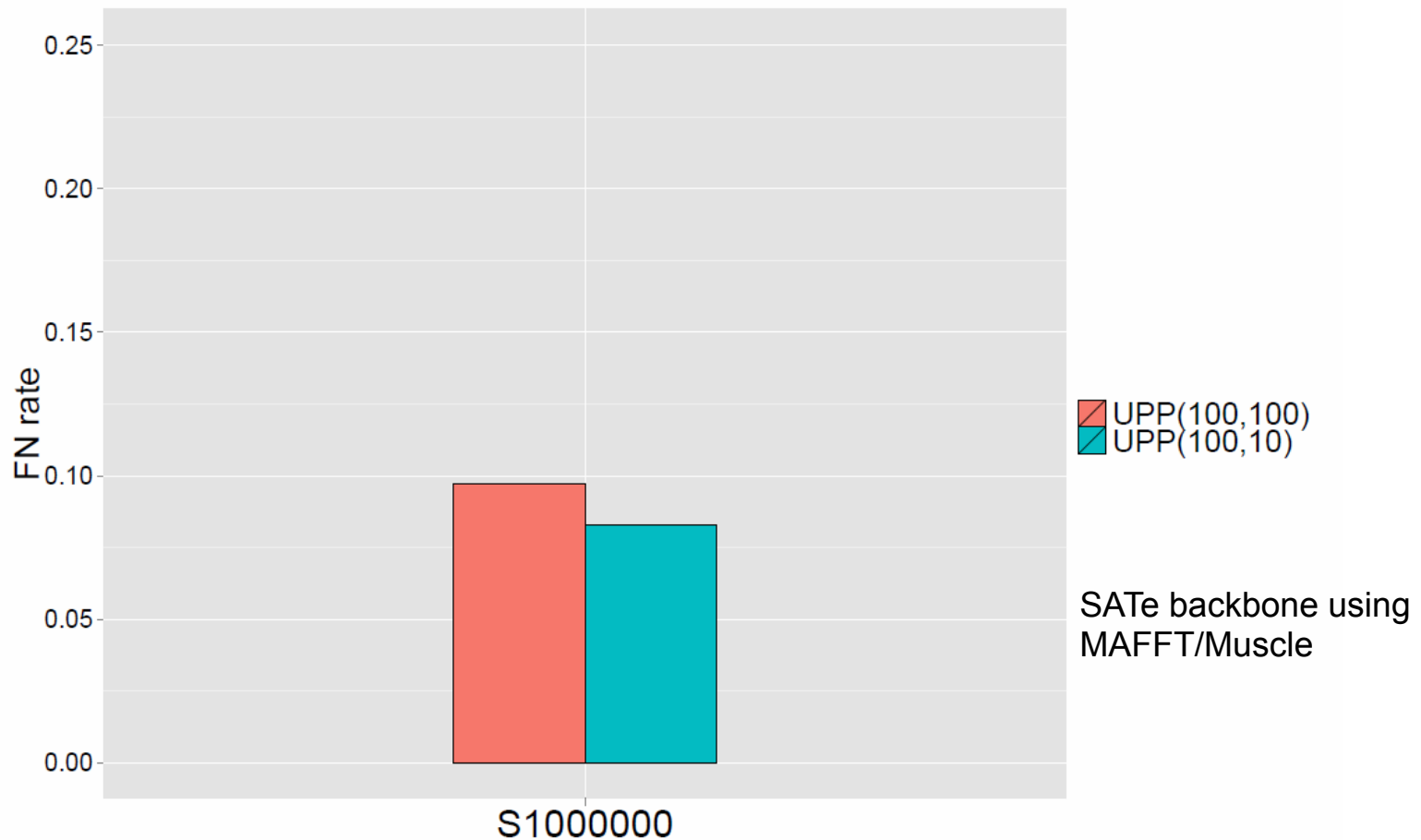


# UPP vs. MAFFT

## Tree error on 10K-200K sequences



# UPP(100,100) vs. UPP(100,10) One Million Taxa: Tree Error



Note improvement obtained by using SEPP decomposition

# Four “Boosters”

**SATé**: co-estimation of alignments and trees

**DACTAL**: tree estimation (almost) without alignments

**SEPP**: phylogenetic placement of short reads

**UPP**: ultra-large multiple sequence alignment

# Phylogenetic “boosters” (meta-methods)

Goal: improve accuracy, speed, robustness, or theoretical guarantees of base methods

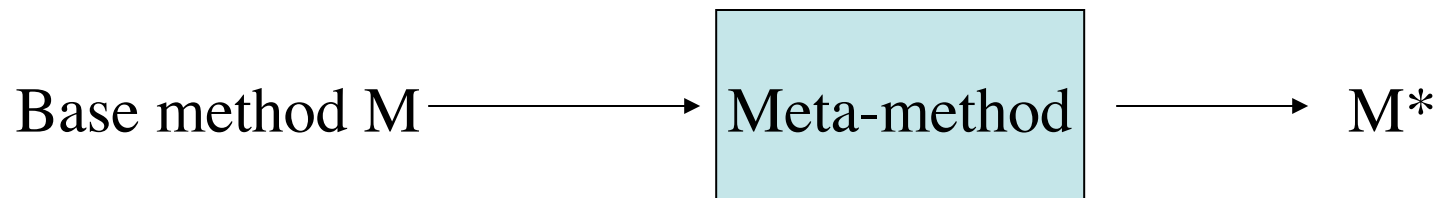
Techniques: divide-and-conquer, iteration, and “bin-and-conquer”

Examples:

- DCM-boosting for distance-based methods (1999)
- DCM-boosting for heuristics for NP-hard problems (1999)
- SATé-boosting for alignment methods (2009)
- SuperFine-boosting for supertree methods (2011)
- DACTAL-boosting: almost alignment-free phylogeny estimation methods (2011)
- SEPP-boosting for phylogenetic placement of short sequences (2012)
- UPP-boosting for alignment methods (unpublished)
- TIPPA-boosting for metagenomic taxon identification (unpublished)
- Binning to improve coalescent-based species tree estimation methods (2013)

# Meta-Methods

- Meta-methods “boost” the performance of base methods (phylogeny reconstruction, alignment estimation, etc).



# Algorithmic Strategies

- Divide-and-conquer
- Iteration
- Multiple HMMs instead of one (for classification problems)
- Bin-and-conquer

# Acknowledgments

- Guggenheim Foundation Fellowship
- Packard Fellowship
- NSF: ATOL, ITR, and IGERT grants
- David Bruton Jr. Professorship
- HHMI
- Microsoft Research
- Texas Advanced Computing Center (TACC)
  
- Collaborators:
  - SATé: Kevin Liu, Serita Nelesen, Sindhu Raghavan, and Randy Linder (and Mark Holder's lab at Kansas for public distribution)
  - DACTAL: Serita Nelesen, Kevin Liu, and Randy Linder
  - SEPP and UPP: Siavash Mirarab and Nam Nguyen

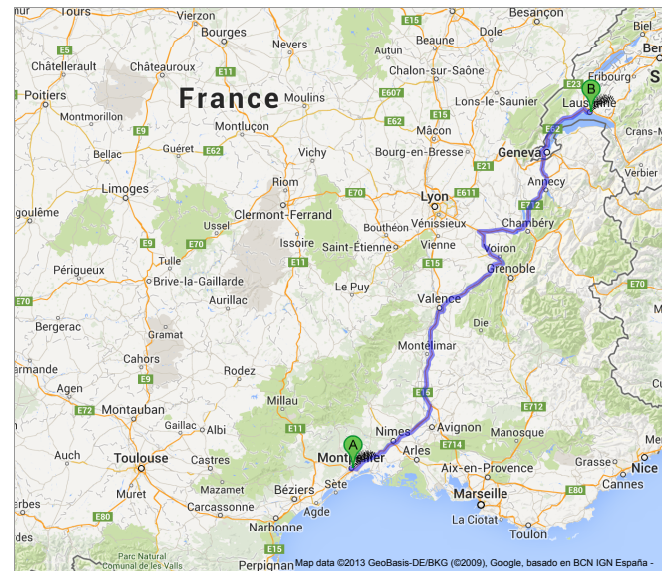
# 2008: David and I deal with a train strike in France

Montpellier, France to Lausanne, Switzerland - Google Maps

8/24/13 8:50 AM

To see all the details that are visible on the screen, use the "Print" link next to the map.

Google



Driving directions to Lausanne, Switzerland

This route has tolls.



Montpellier  
France

1. Head northeast on Pl. Martyrs de la Résistance toward Rue Foch

21 m

<https://maps.google.ca/maps?hl=en&q=montpellier+train+station&ie=UTF-8>

Page 1 of 4

## Montpellier, France to Lausanne, Switzerland



Thank you David!

