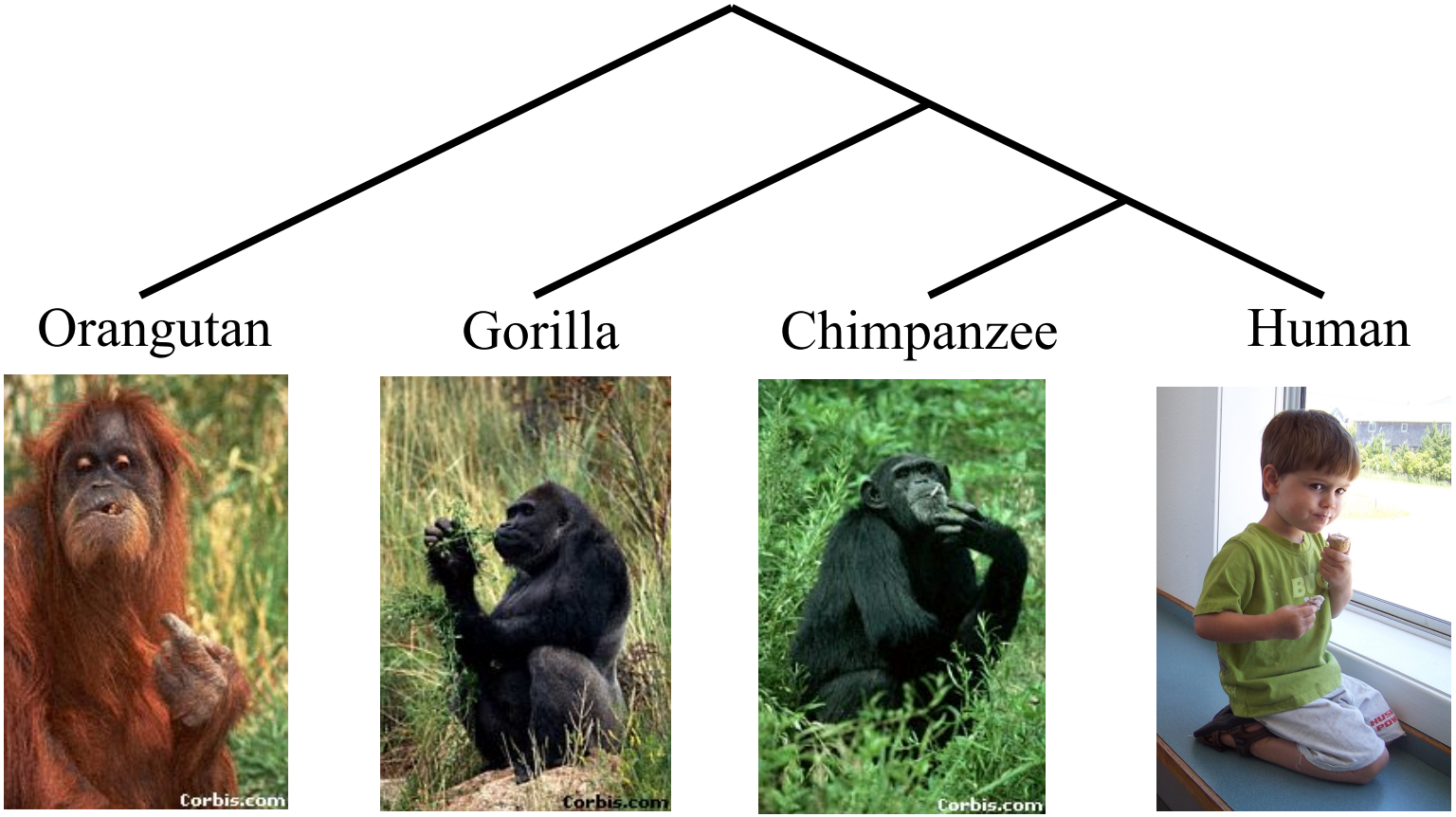
Large-scale Multiple Sequence Alignment and Phylogenetic Estimation

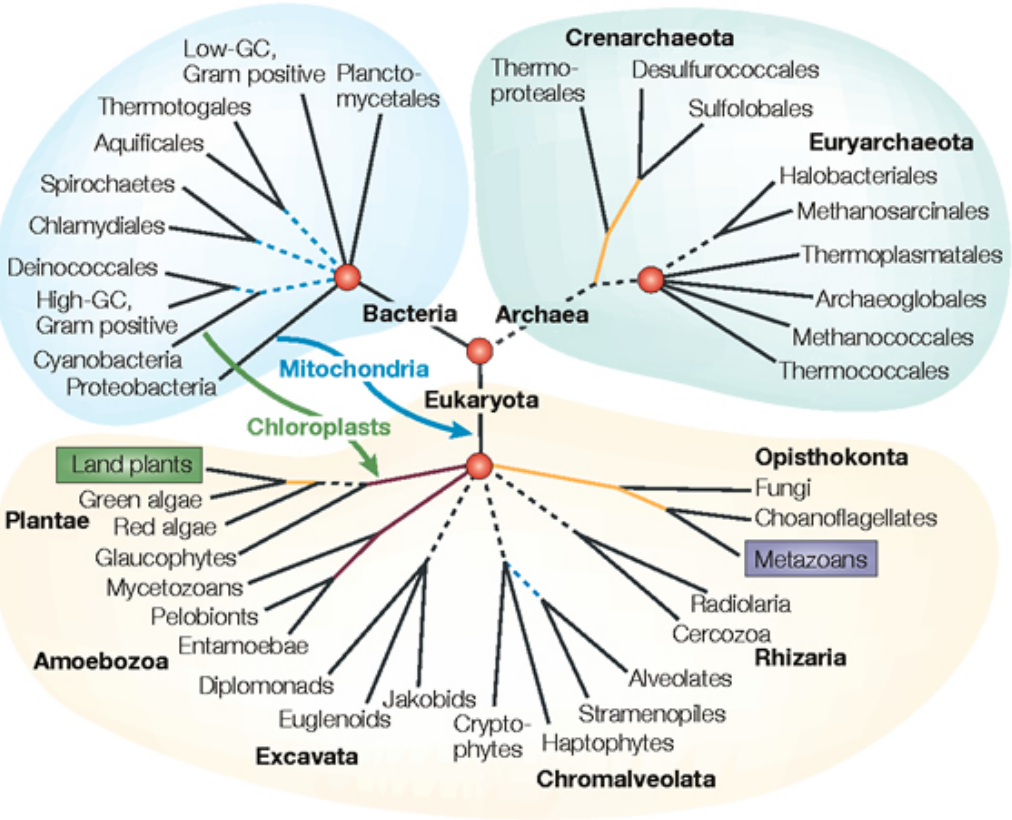
Tandy Warnow
Department of Computer Science
The University of Texas at Austin

Phylogeny (evolutionary tree)



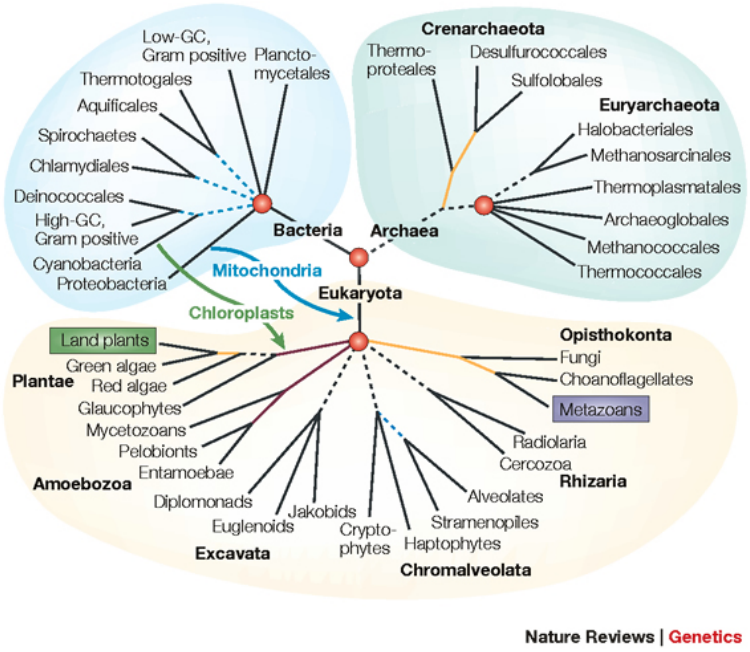
From the Tree of the Life Website,
University of Arizona

The “Tree of Life”



Nature Reviews | Genetics

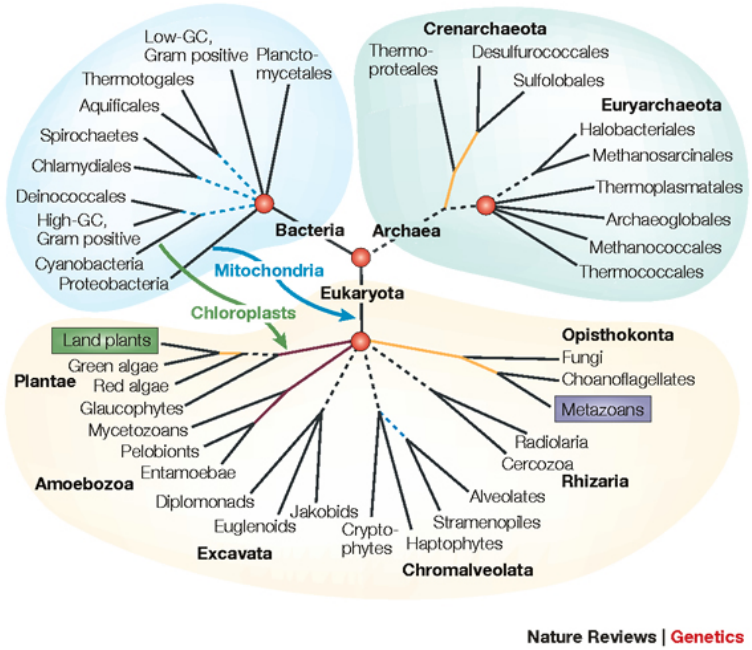
The Tree of Life: Applications to Biology



Biomedical applications
- Mechanisms of evolution
- Environmental influences
- Drug Design
- Protein structure and function
- Human migrations

"Nothing in biology makes sense except in the light of evolution"
Dobzhansky

Estimating the Tree of Life: a *Grand Challenge*

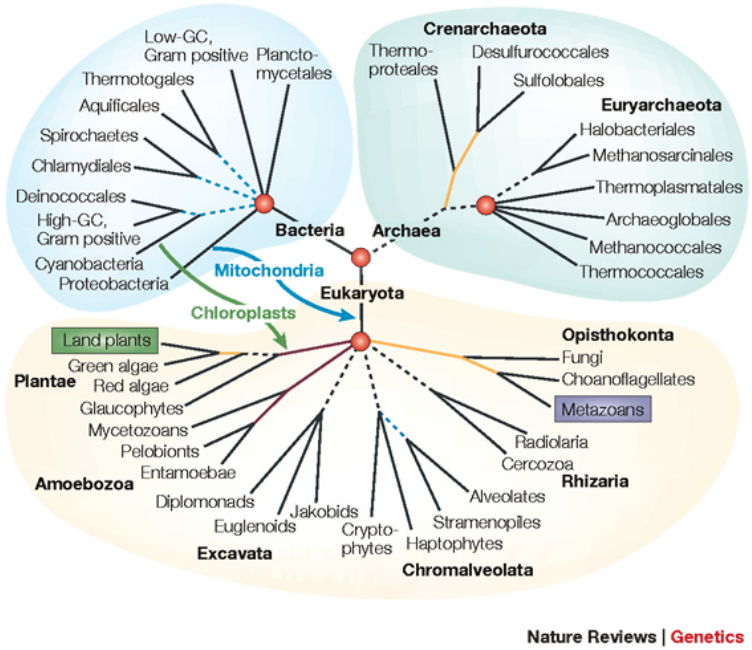


Most well studied problem:

Given DNA sequences, find the Maximum Likelihood Tree

NP-hard, lots of software (RAxML, FastTree-2, GARLI, etc.)

Estimating the Tree of Life: a *Grand Challenge*



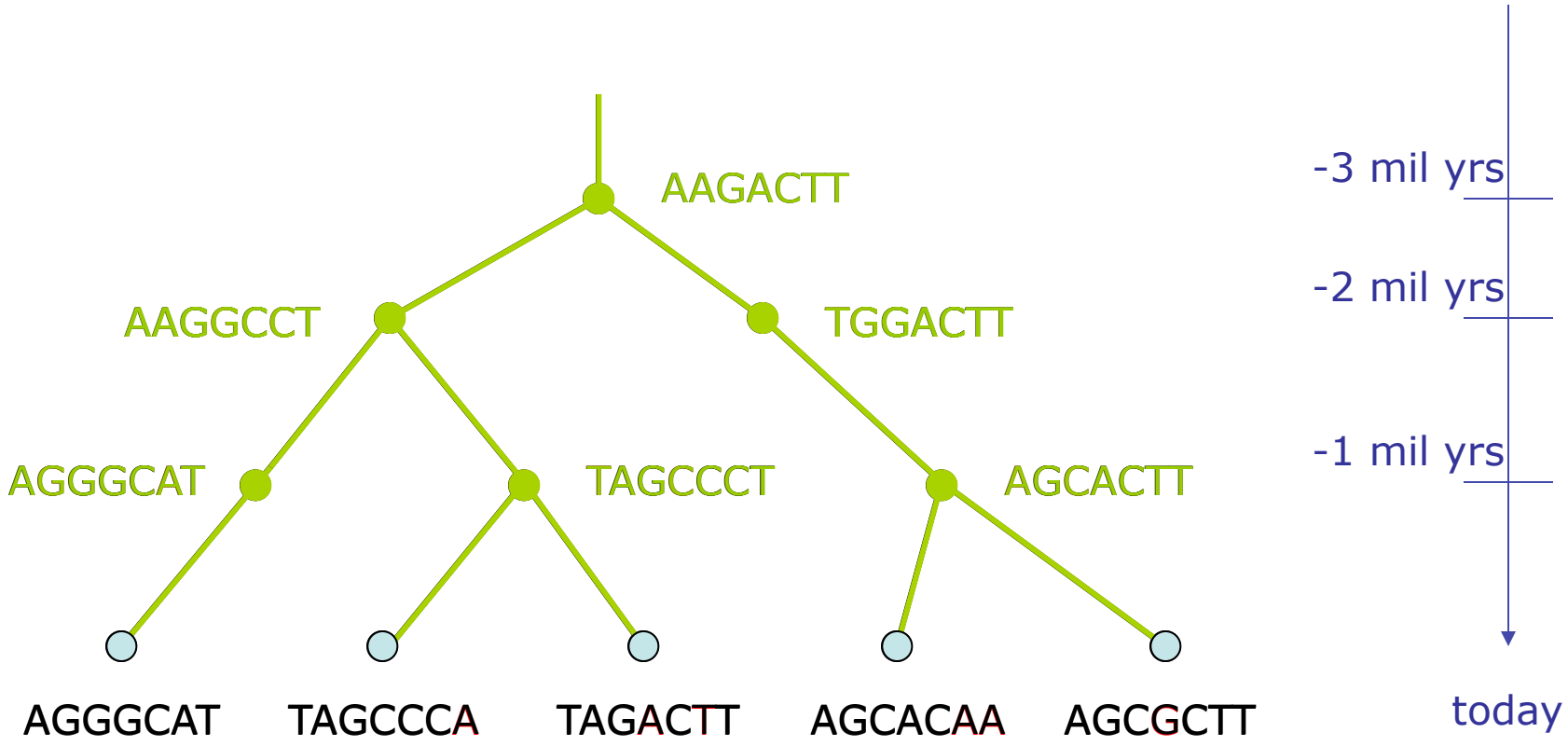
Novel techniques needed for scalability and accuracy

NP-hard problems and large datasets

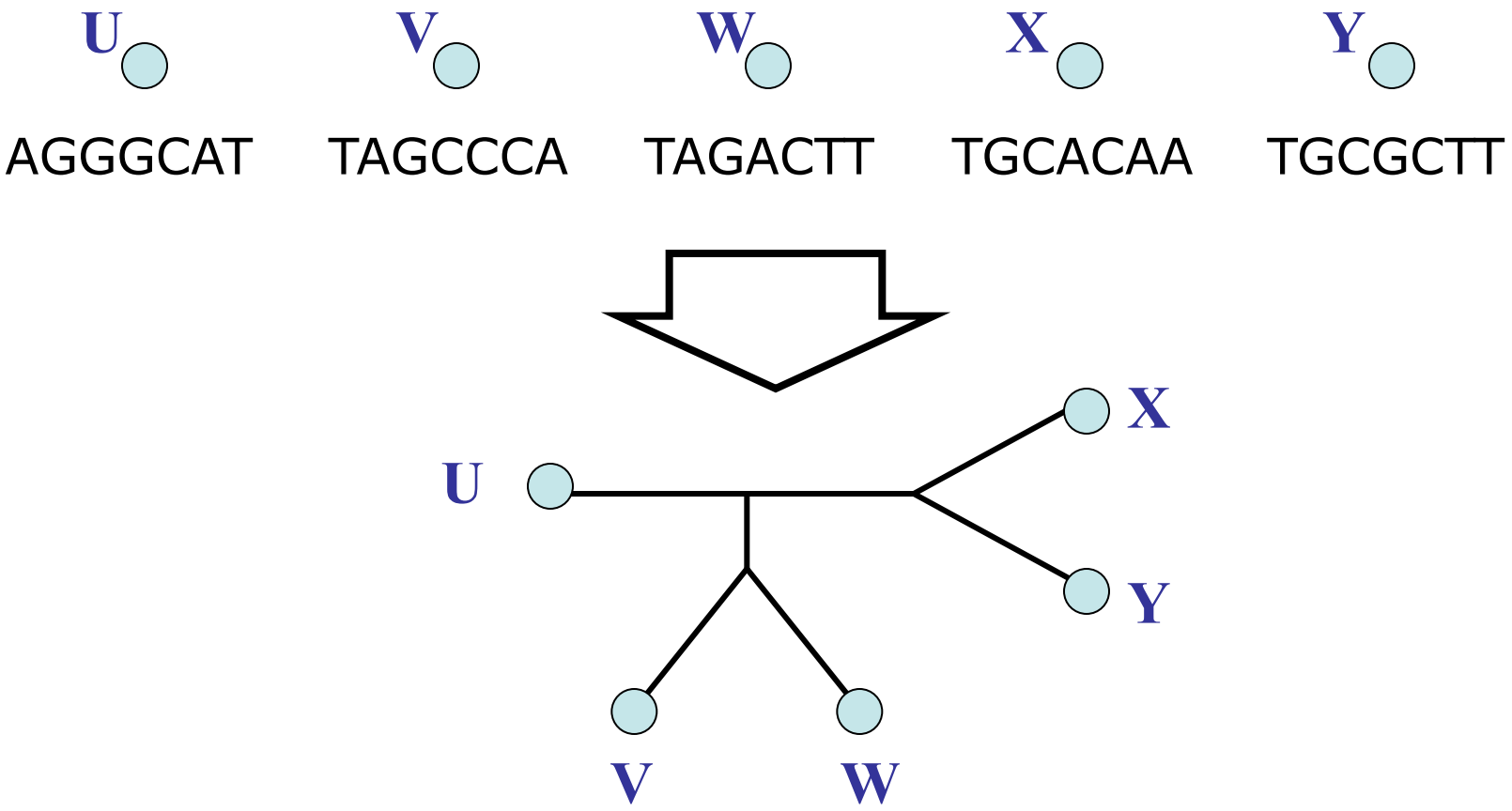
Current methods do not provide good accuracy

HPC is insufficient

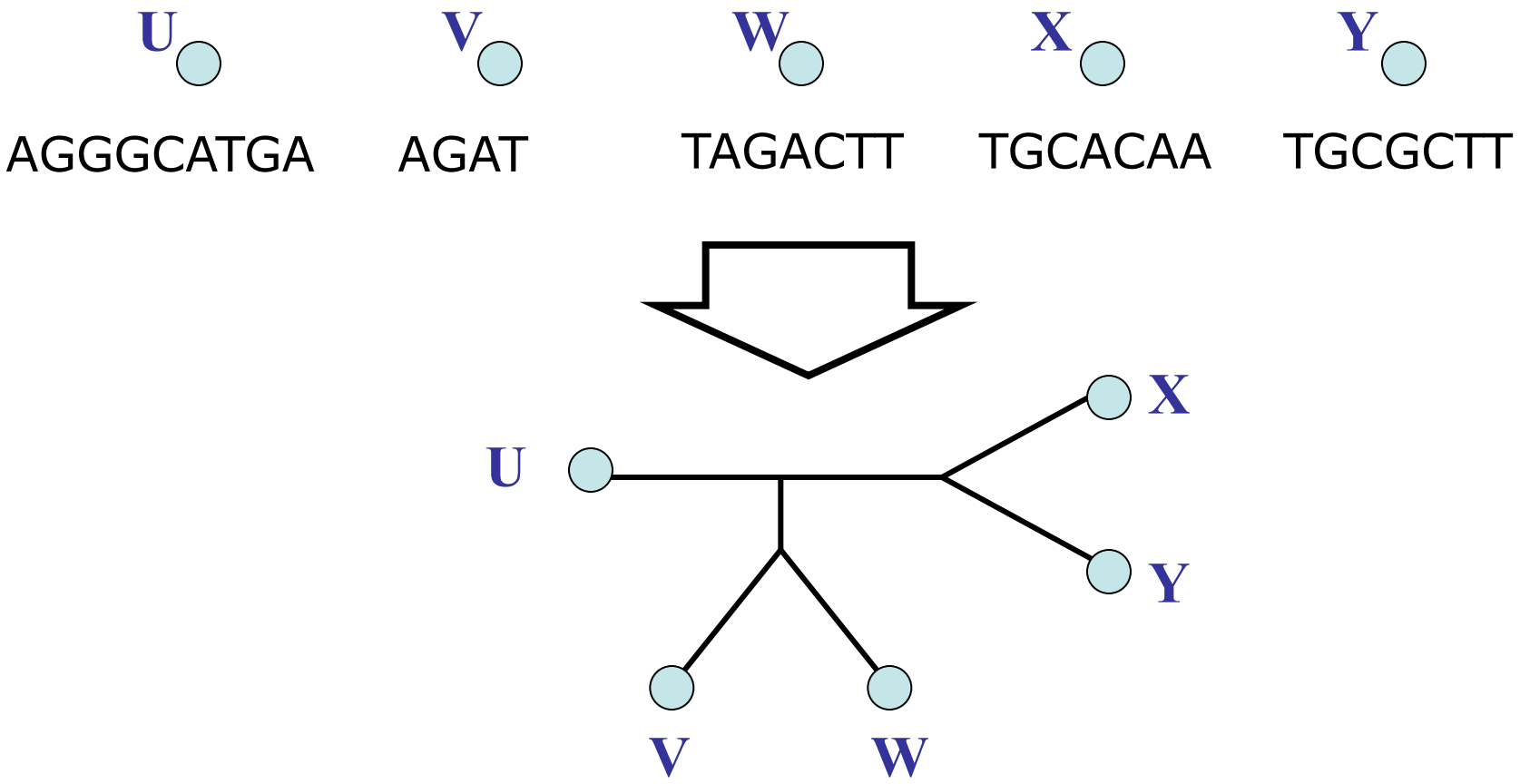
DNA Sequence Evolution



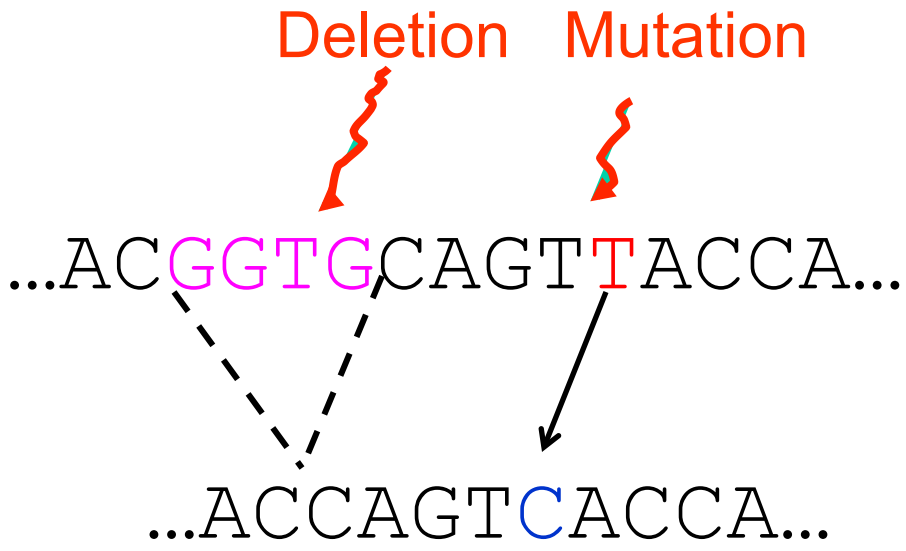
Phylogeny Problem

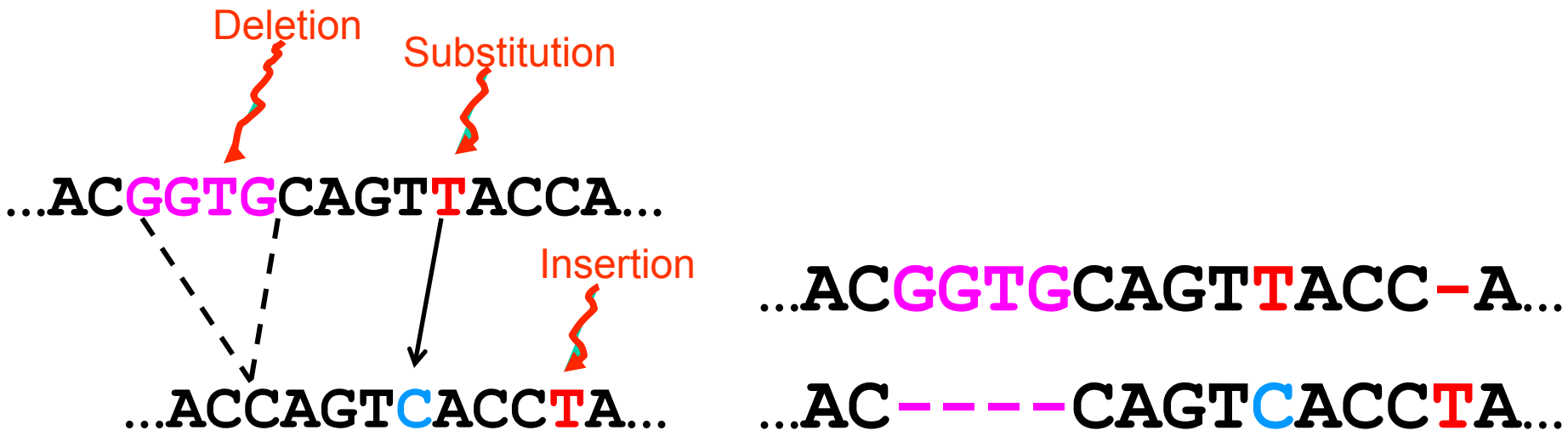


The “real” problem



Indels (insertions and deletions)

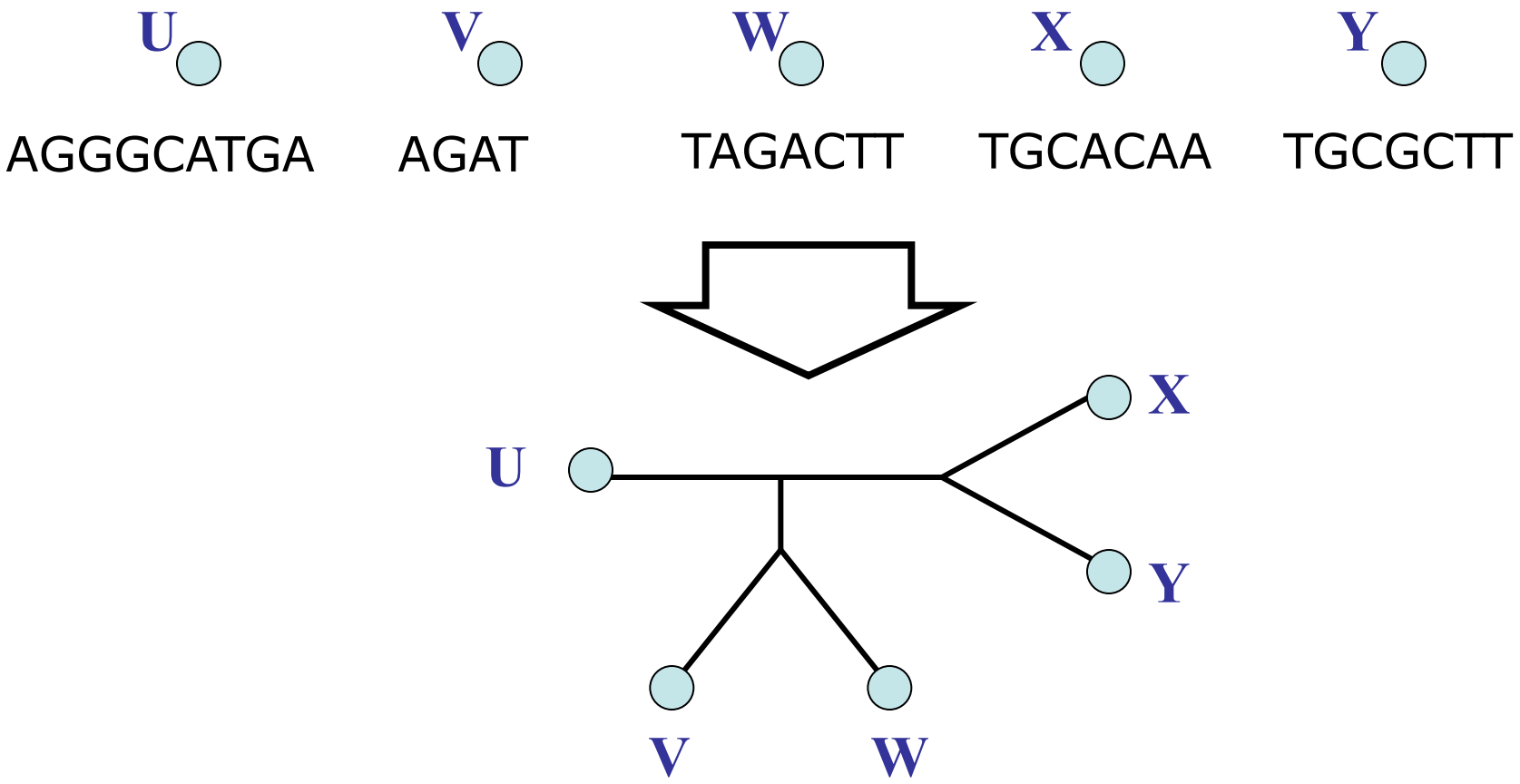




The true multiple alignment

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

The “real” problem



Input: unaligned sequences

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

Phase 1: Alignment

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



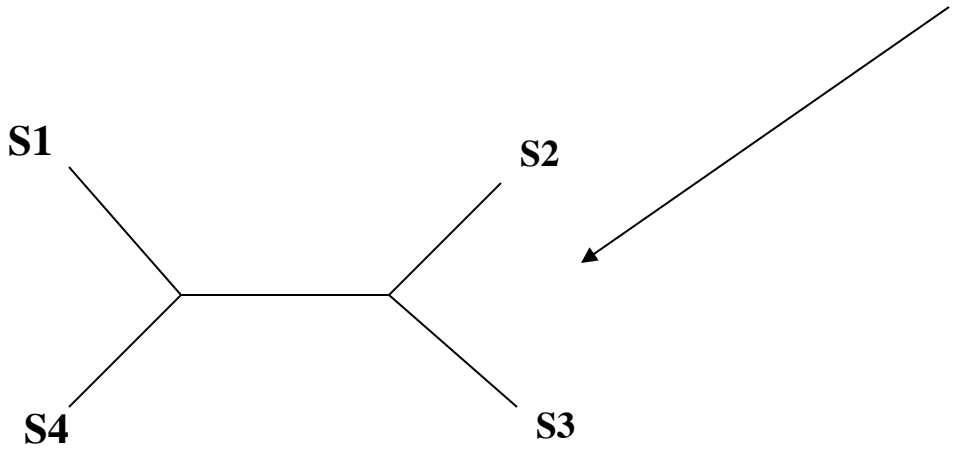
```
S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAC--GACCGC--
S3 = TAG-CT-------GACCGC--
S4 = -------TCAC--GACCGACA
```

Phase 2: Construct tree

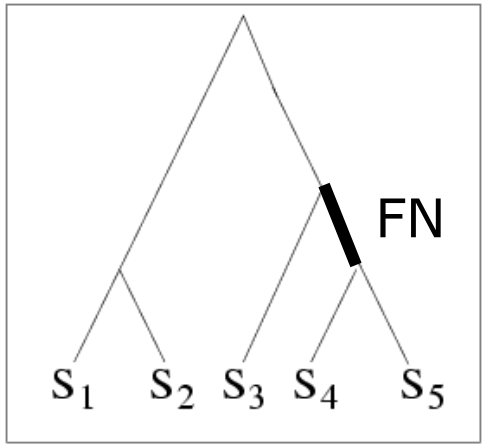
```
S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC
S3 = TAGCTGACCGC
S4 = TCACGACCGACA
```

→

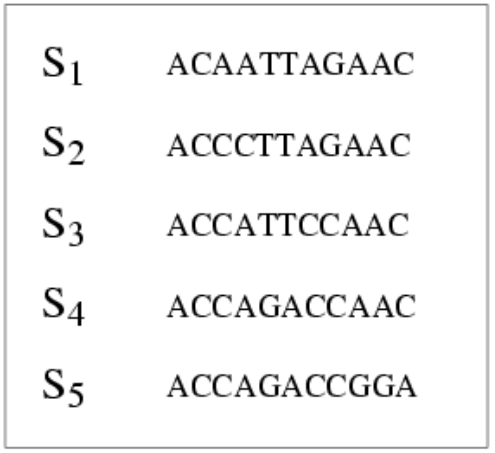
```
S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAC--GACCGC--
S3 = TAG-CT-------GACCGC--
S4 = -------TCAC--GACCGACA
```



Quantifying Error



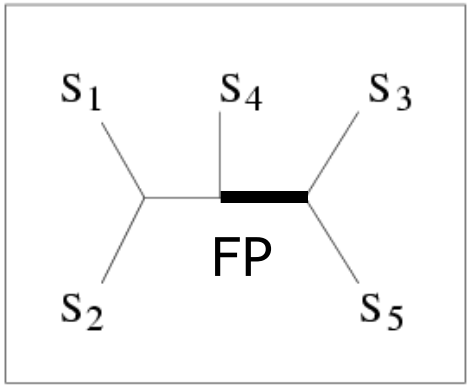
TRUE TREE



DNA SEQUENCES

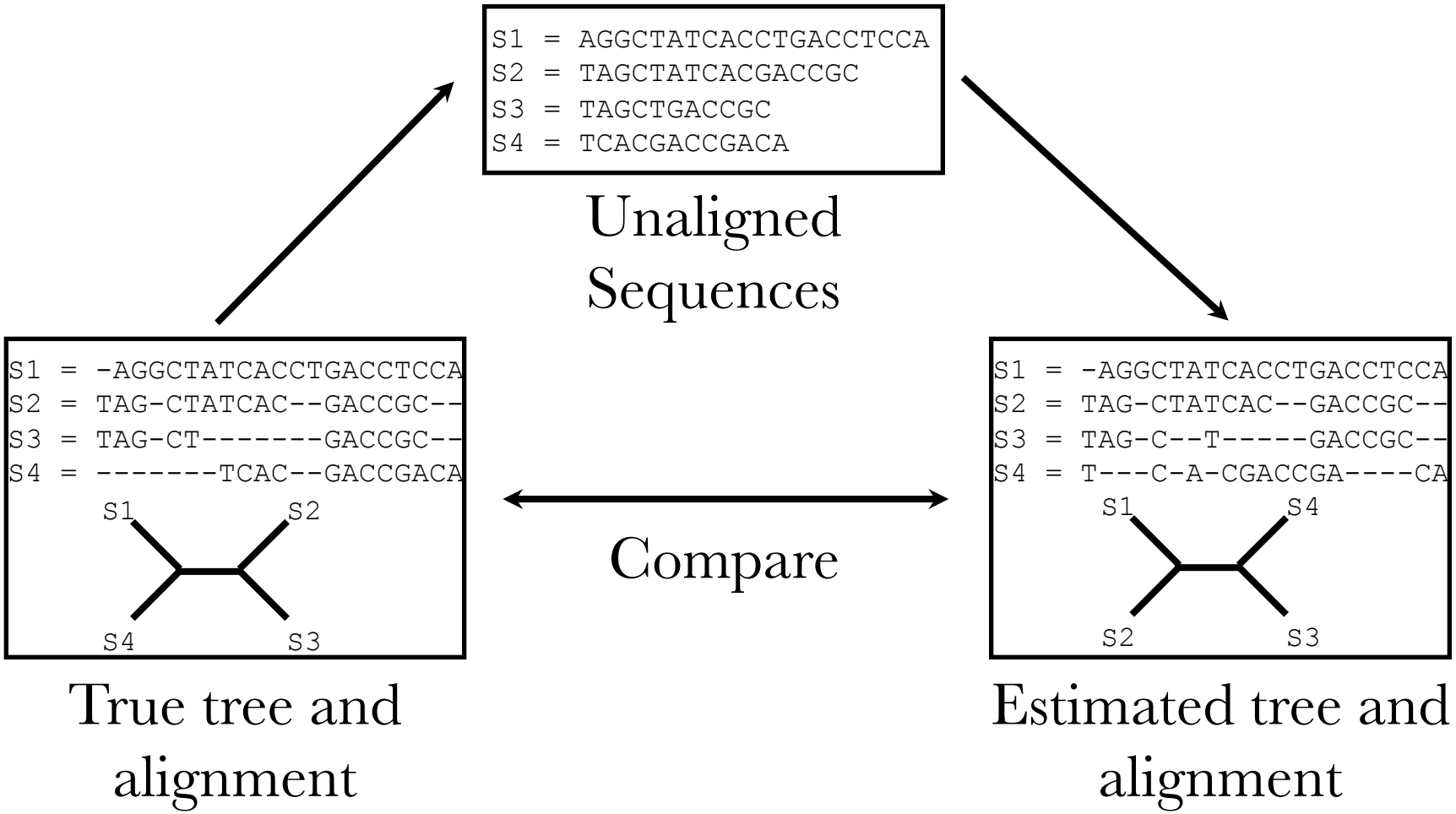
FN: false negative
(missing edge)
FP: false positive
(incorrect edge)

50% error rate



INFERRED TREE

Simulation Studies



Two-phase estimation

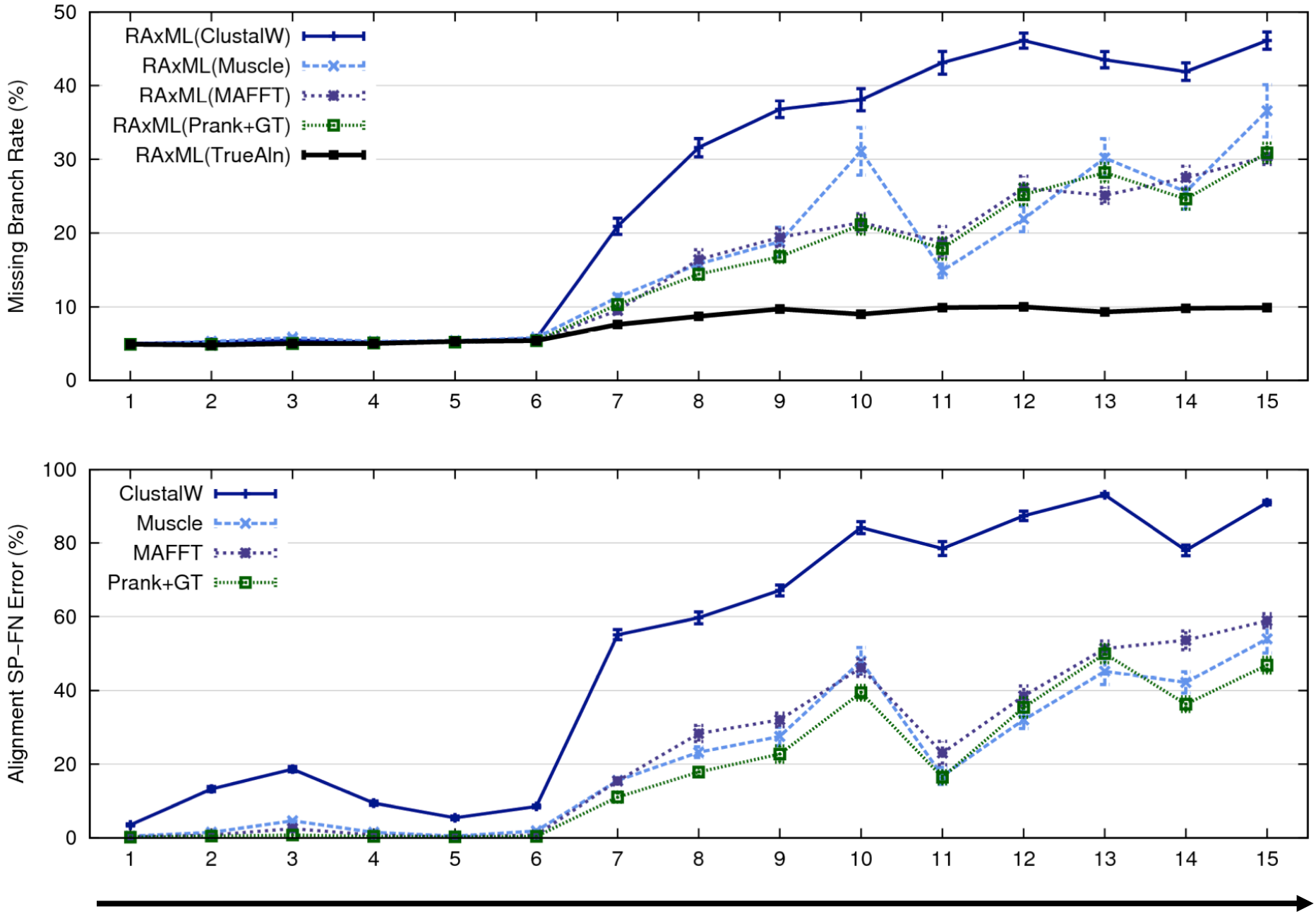
Alignment methods

- 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*

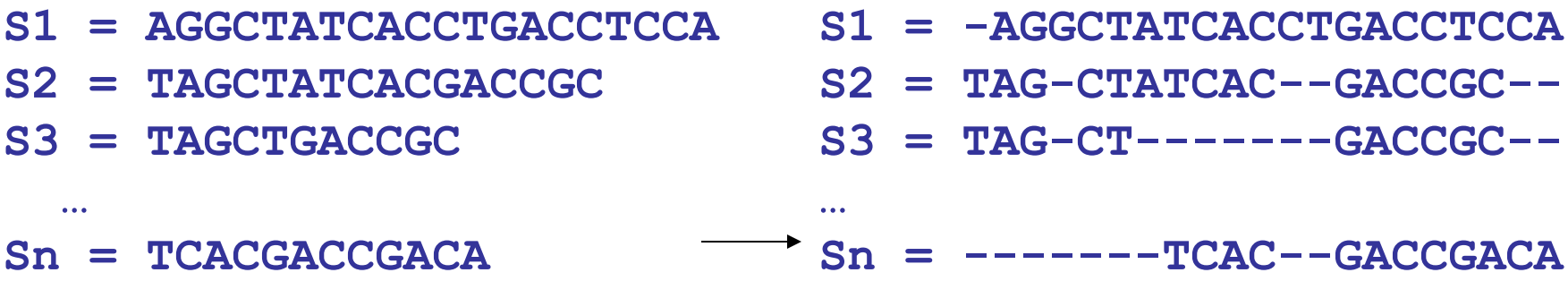

RAxML(ClustalW)
RAxML(Muscle)
RAxML(MAFFT)
RAxML(Prank+GT)
RAxML(TrueAln)
Missing Branch Rate (%)
0
10
20
30
40
50
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
ClustalW
Muscle
MAFFT
Prank+GT
Alignment SP–FN Error (%)
0
20
40
60
80
100
1000-taxon models, ordered by difficulty (Liu et al., 2009)

Problems with the two-phase approach

- Current alignment methods fail to return reasonable alignments on large datasets with high rates of indels and substitutions.
- Manual alignment is time consuming and subjective.
- *Systematists discard potentially useful markers* if they are difficult to align.

This issues seriously impact large-scale phylogeny estimation (and Tree of Life projects)

Multiple Sequence Alignment (MSA): *another grand challenge*[1]



Novel techniques needed for scalability and accuracy

NP-hard problems and large datasets

Current methods do not provide good accuracy

Few methods can analyze even moderately large datasets

Many important applications besides phylogenetic estimation

[1] Frontiers in Massive Data Analysis, National Academies Press, 2013

1kp: Thousand Transcriptome Project

G. Ka-Shu Wong
U Alberta

J. Leebens-Mack
U Georgia

N. Wickett
Northwestern

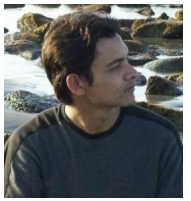
N. Matasci
iPlant

T. Warnow,
UT-Austin

S. Mirarab,
UT-Austin

N. Nguyen,
UT-Austin

Md. S.Bayzid
UT-Austin



Plus many many other people…

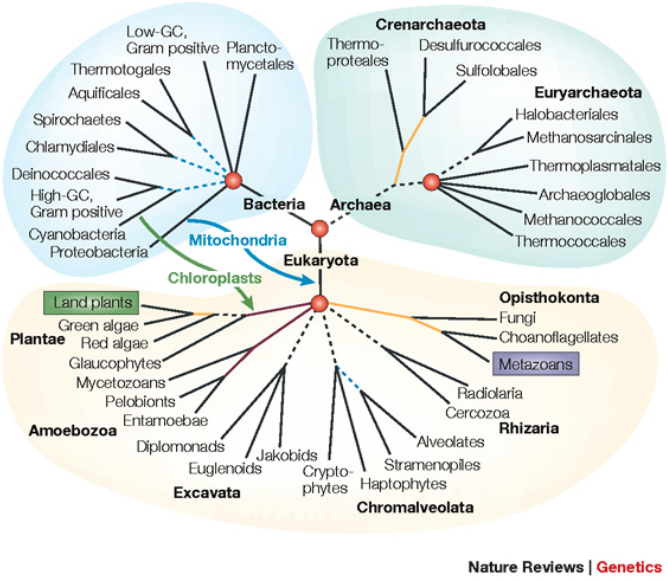
- Plant Tree of Life based on transcriptomes of ~1200 species
- More than 13,000 gene families (most not single copy)

Challenges:

Large datasets with > 100,000 sequences

Gene tree incongruence

The Tree of Life: *Multiple Challenges*



Large datasets:

100,000+ sequences

10,000+ genes

"BigData" complexity

Today's talk

Challenges:

Ultra-large multiple-sequence alignment

Alignment-free phylogeny estimation

Supertree estimation

Estimating species trees from incongruent gene trees

Absolute fast converging methods

Genome rearrangement phylogeny

Reticulate evolution

Visualization of large trees and alignments

Data mining techniques to explore multiple optima

This Talk

SATé - co-estimating trees and alignments (Science, 2009 and Systematic Biology 2012)

UPP - ultra-large alignment estimation using SEPP (unpublished)

SEPP – phylogenetic placement (PSB 2012)

TIPP – taxon identification using SEPP (unpublished)

Part I: SATé

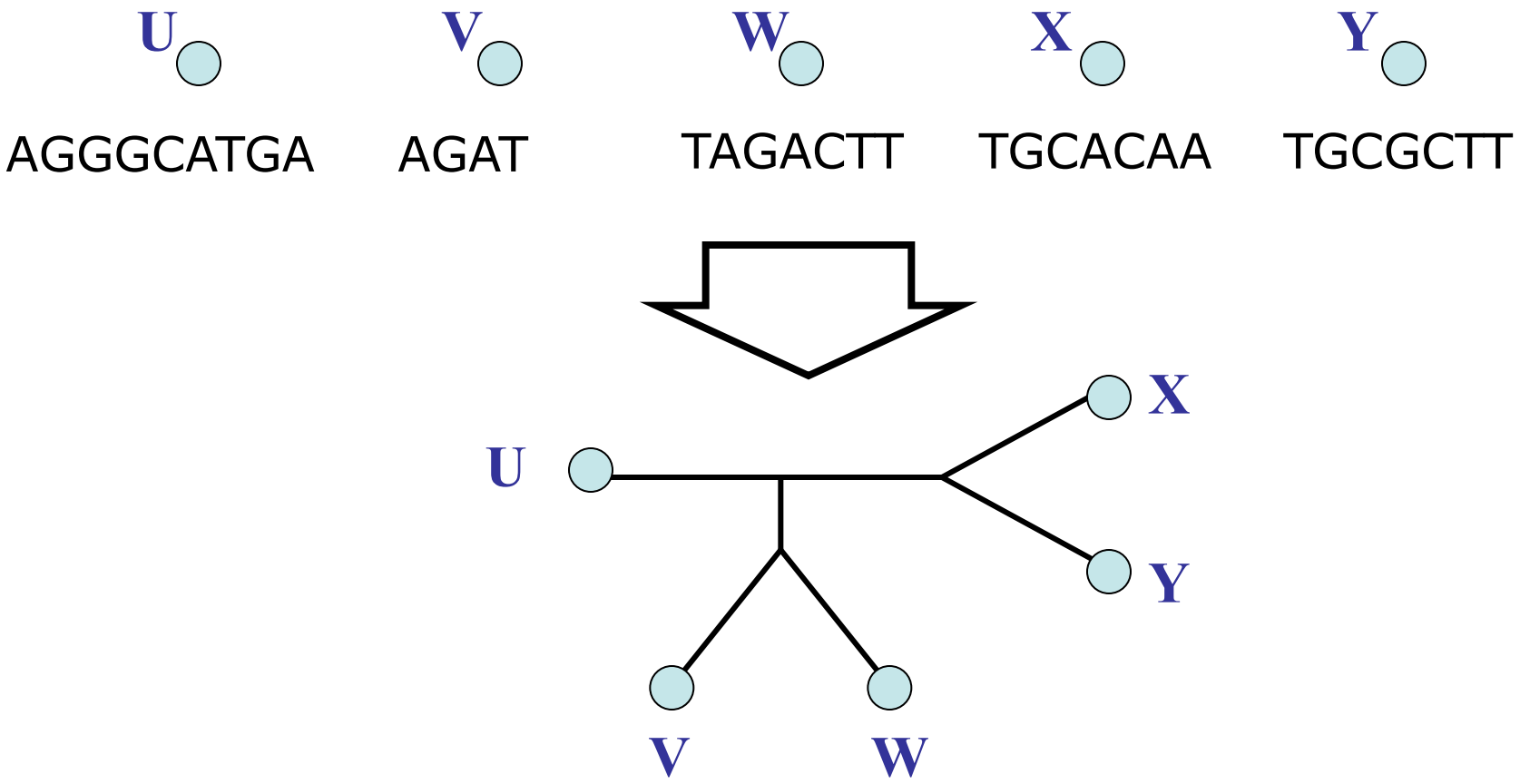
Simultaneous Alignment and Tree Estimation

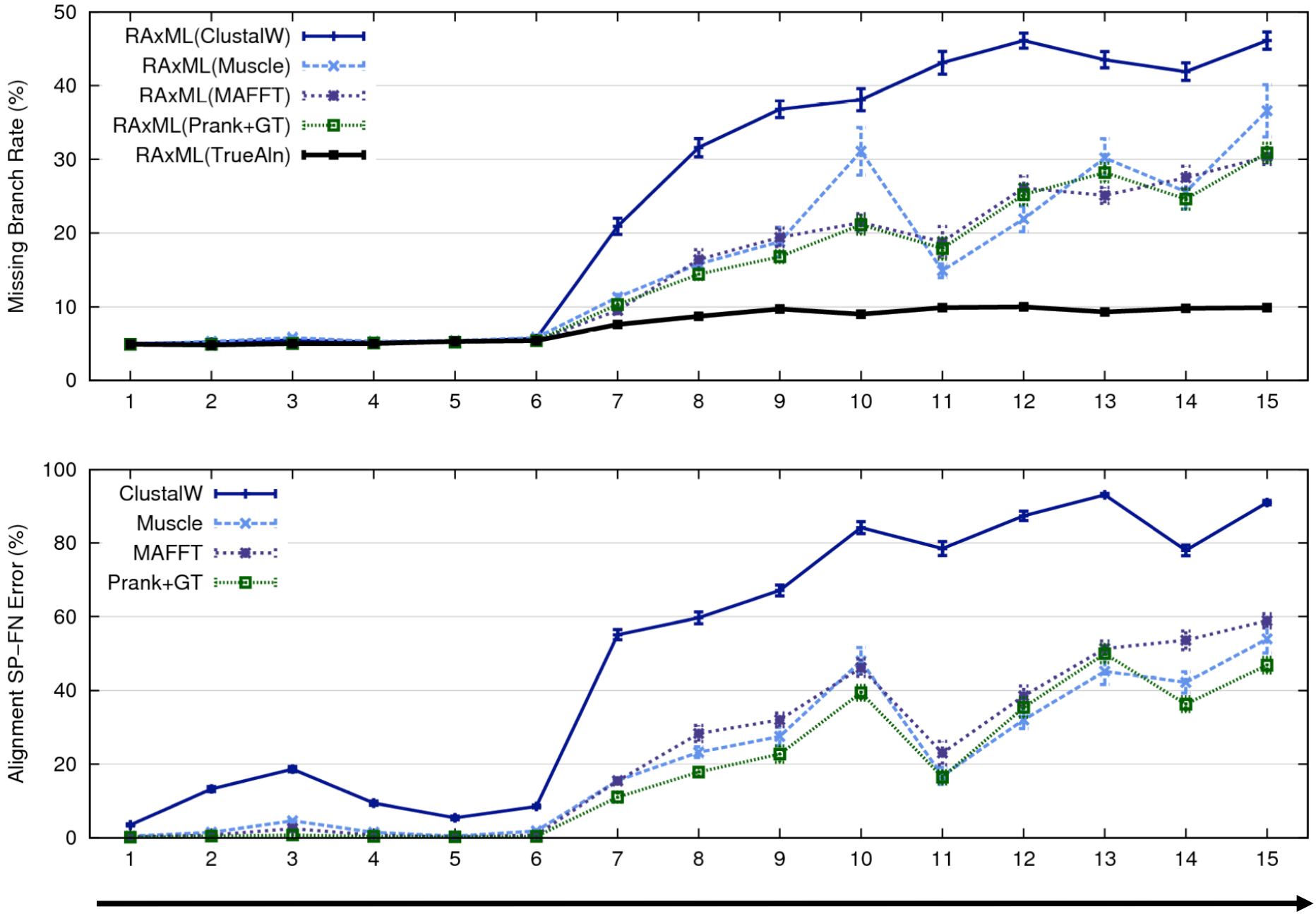
Liu, Nelesen, Raghavan, Linder, and Warnow, *Science*, 19 June 2009, pp. 1561-1564.

Liu et al., Systematic Biology 2012

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

The “real” problem




RAxML(ClustalW)
RAxML(Muscle)
RAxML(MAFFT)
RAxML(Prank+GT)
RAxML(TrueAln)
Missing Branch Rate (%)
0
10
20
30
40
50
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
ClustalW
Muscle
MAFFT
Prank+GT
Alignment SP–FN Error (%)
0
20
40
60
80
100
1000-taxon models, ordered by difficulty (Liu et al., 2009)

Two-phase estimation

- Alignment error increases with the rate of evolution, and poor alignments result in poor trees.

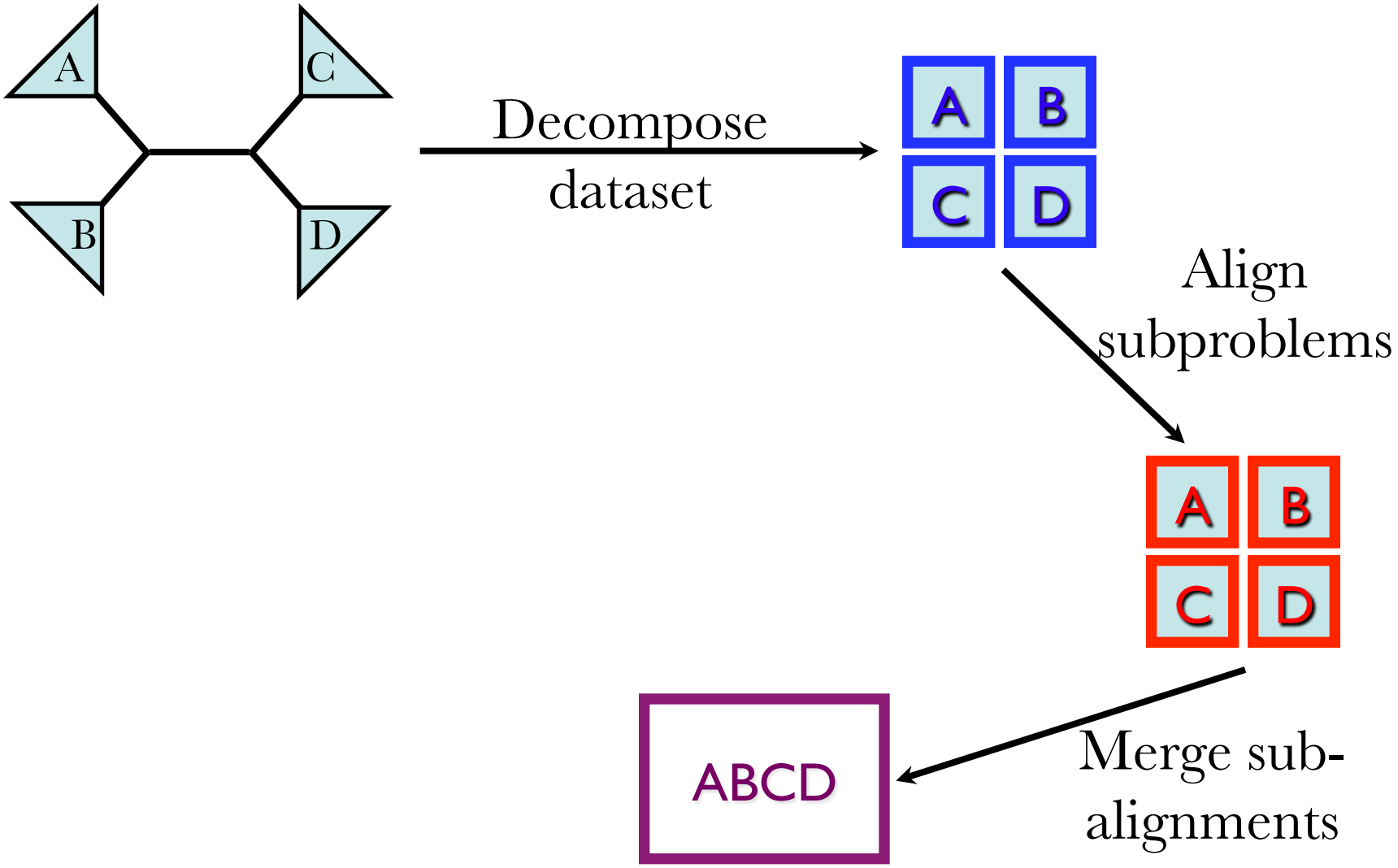
- Datasets with small enough "evolutionary diameters" are easy to align with high accuracy.

Alignment on the tree

- Idea: better (more accurate) alignments will be found if we align subsets with smaller diameters, and then combine alignments on these subsets

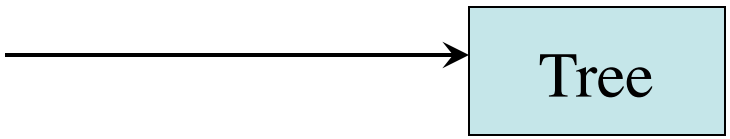
- Approach: use the tree topology to divide-and-conquer

Cartoon (real decomposition is different)



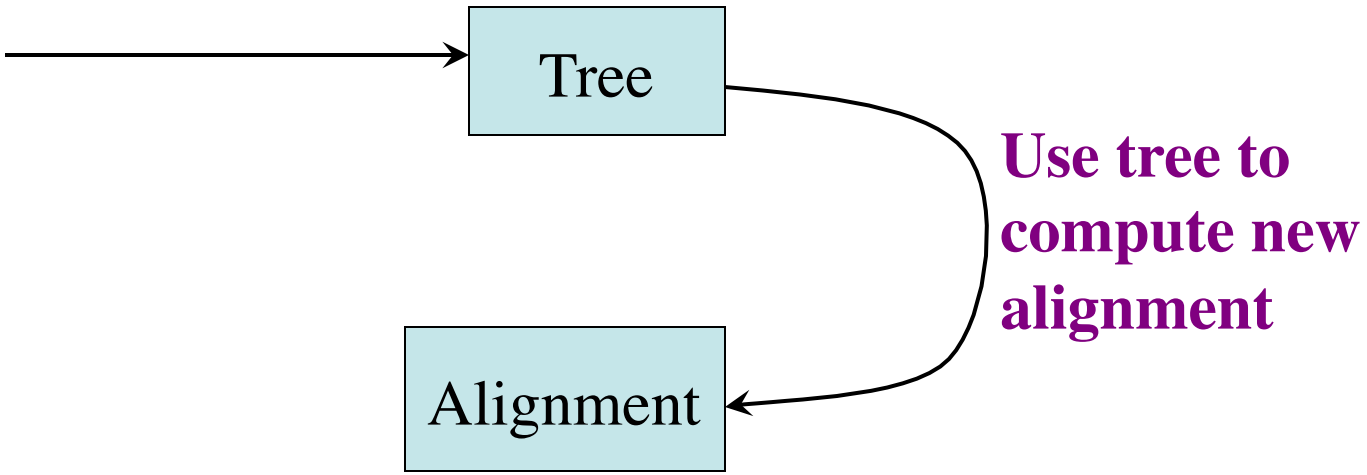
SATé Algorithm

Obtain initial alignment
and estimated ML tree



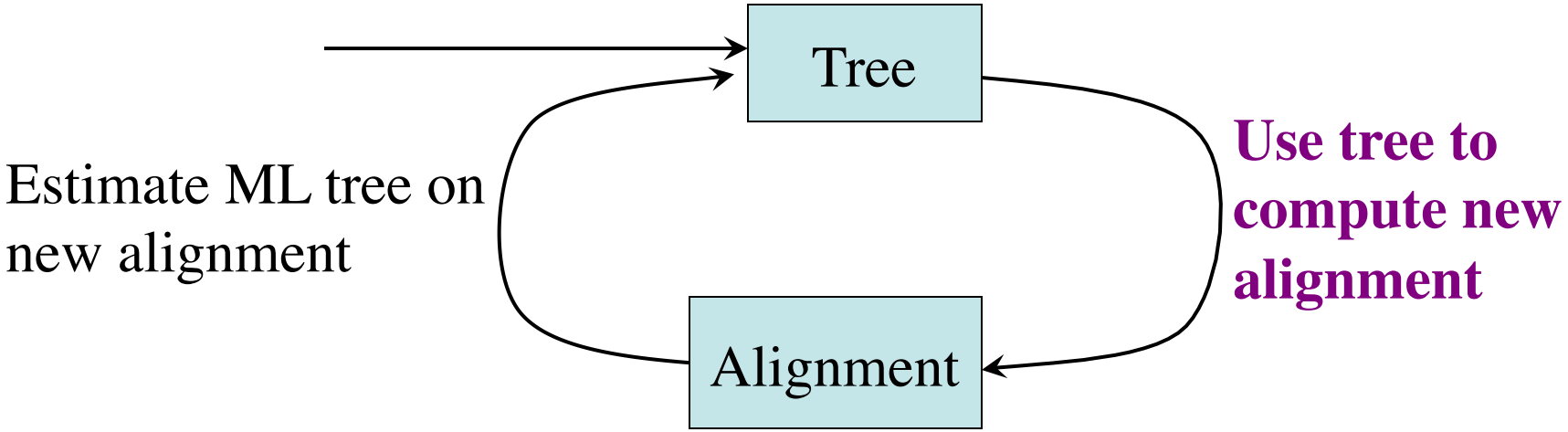
SATé Algorithm

Obtain initial alignment
and estimated ML tree

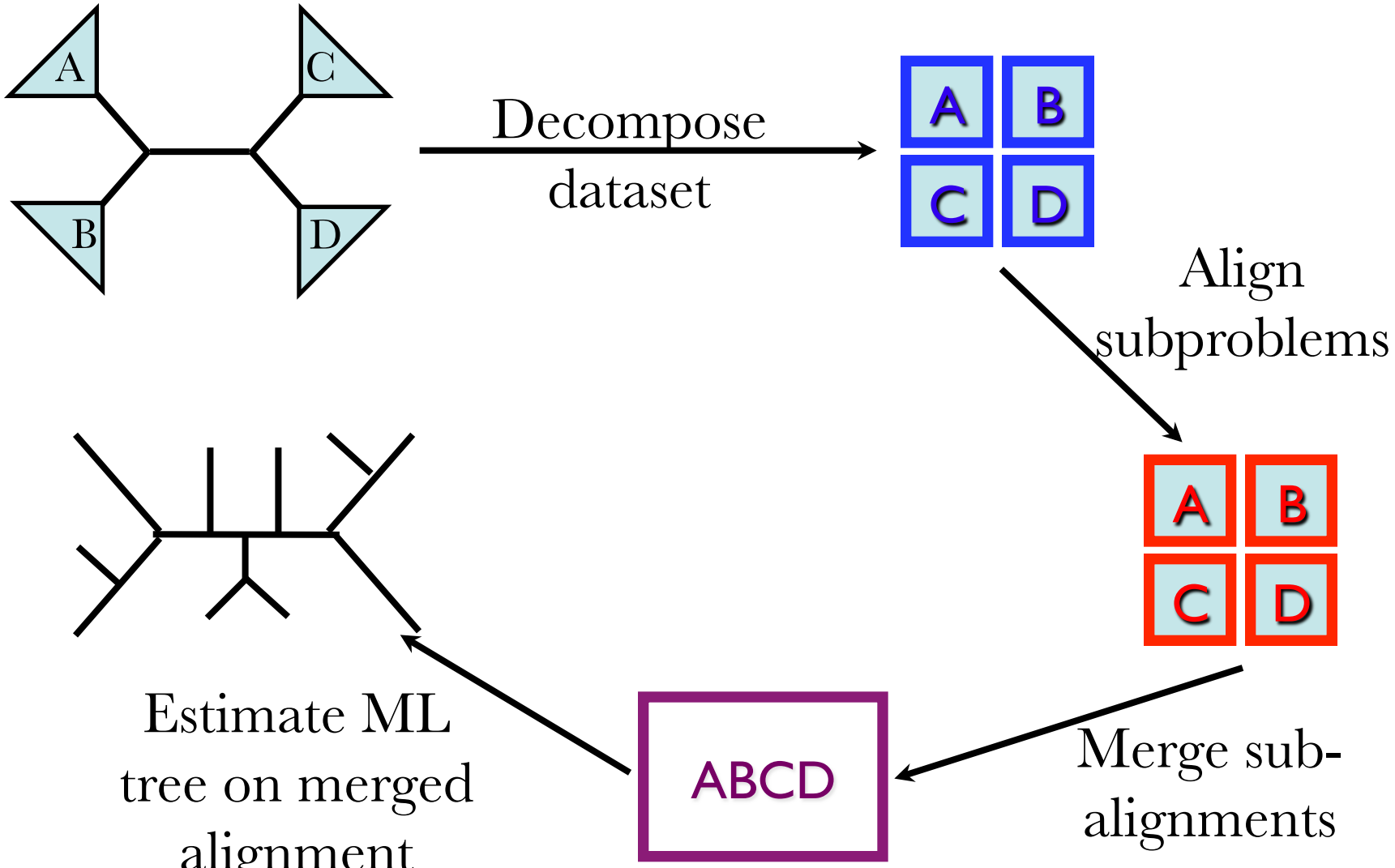


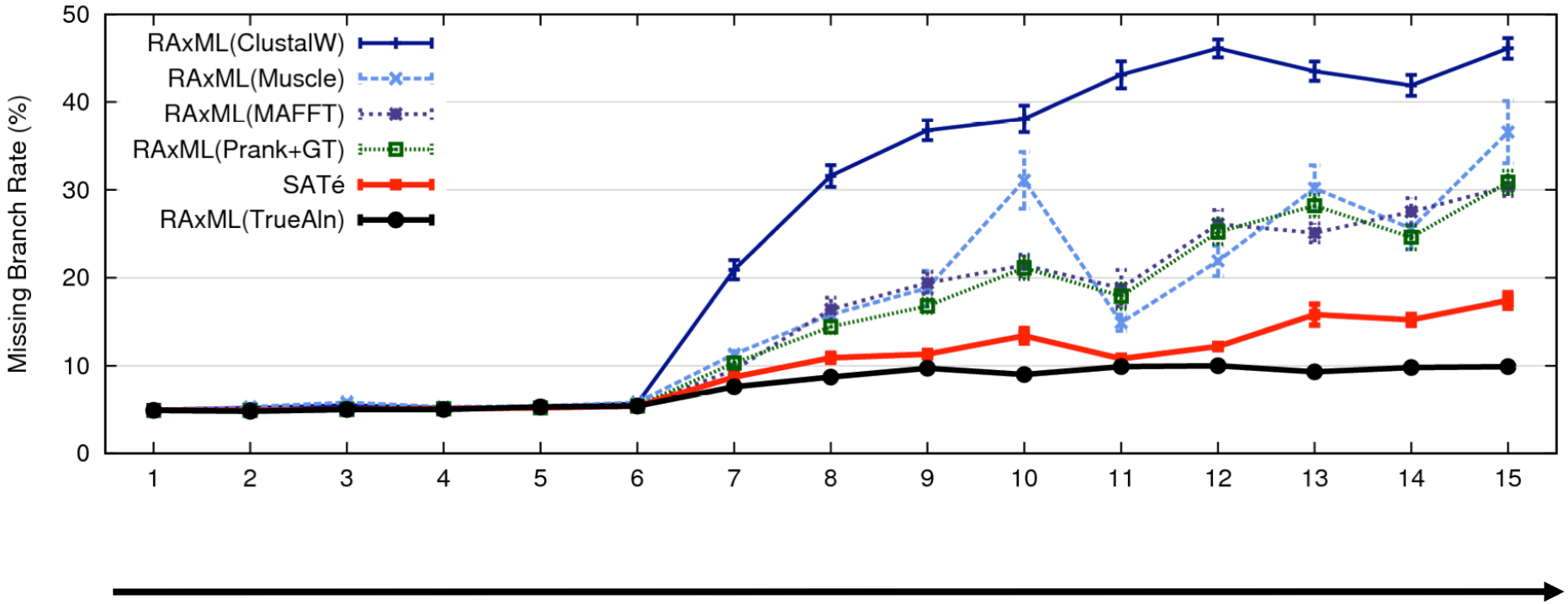
SATé Algorithm

Obtain initial alignment
and estimated ML tree



One iteration (cartoon)

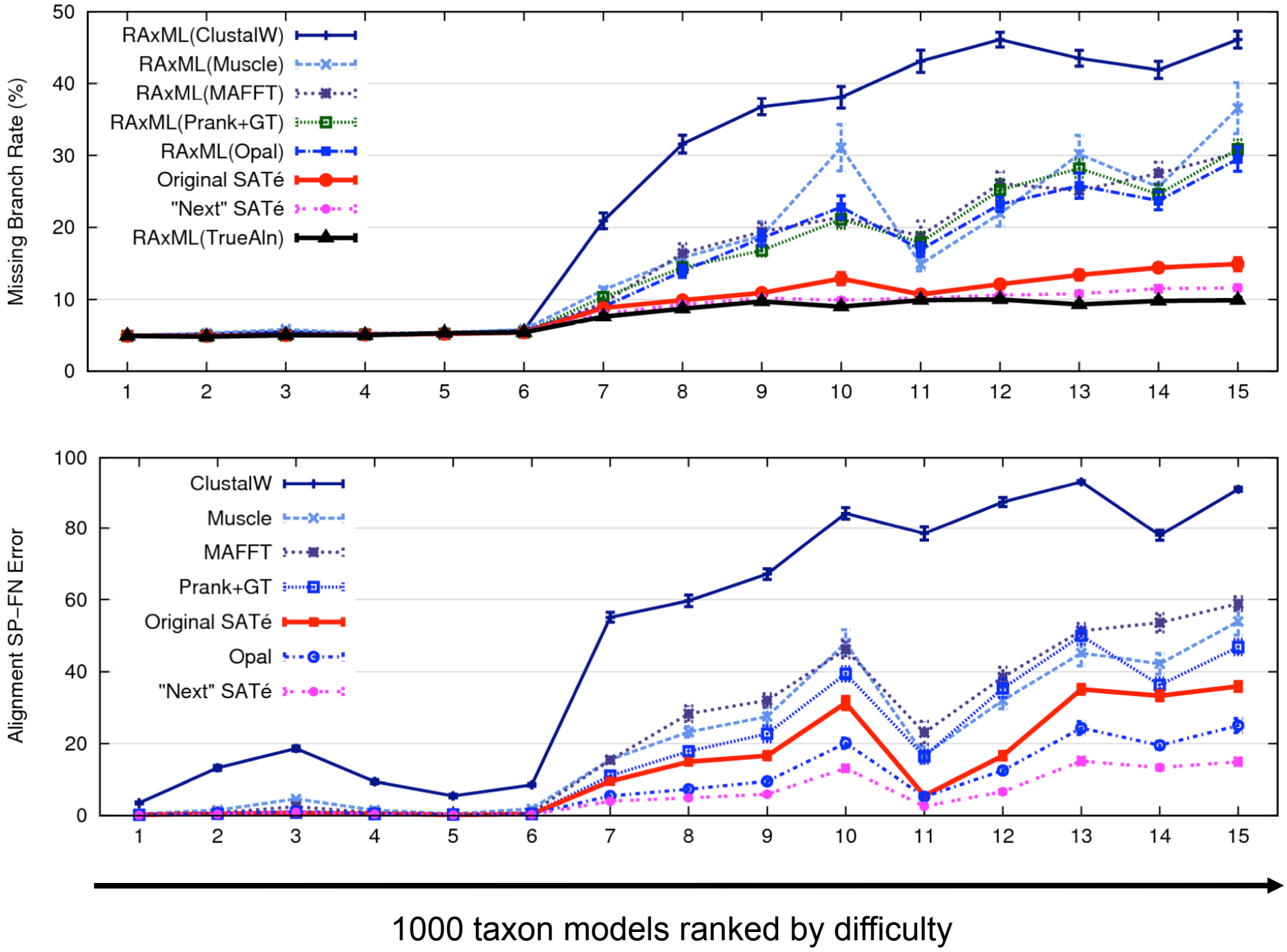




1000 taxon models, ordered by difficulty

24 hour SATé analysis, on desktop machines

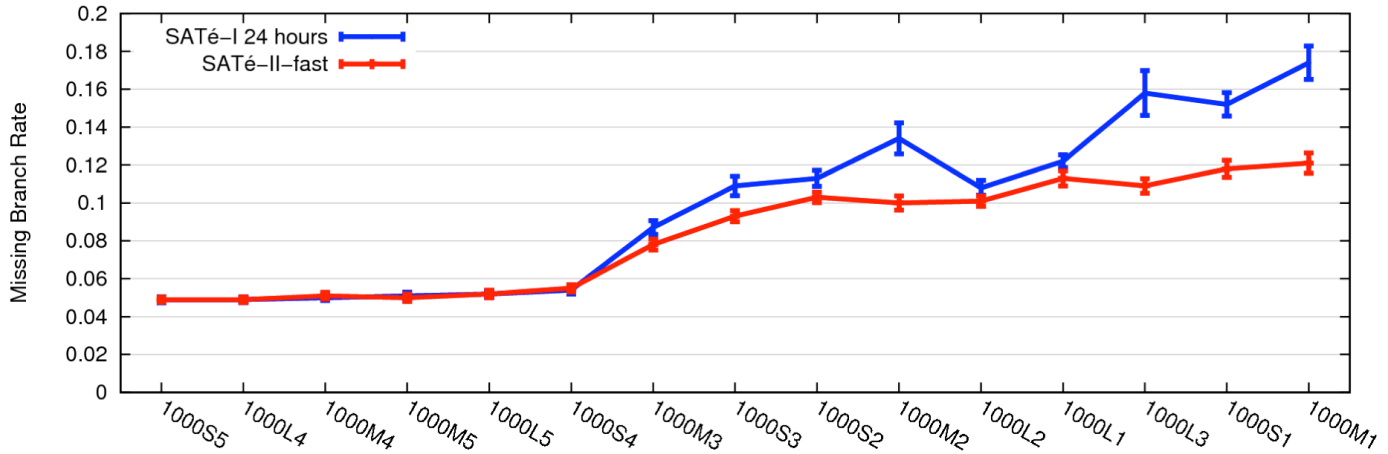
(Similar improvements for biological datasets)

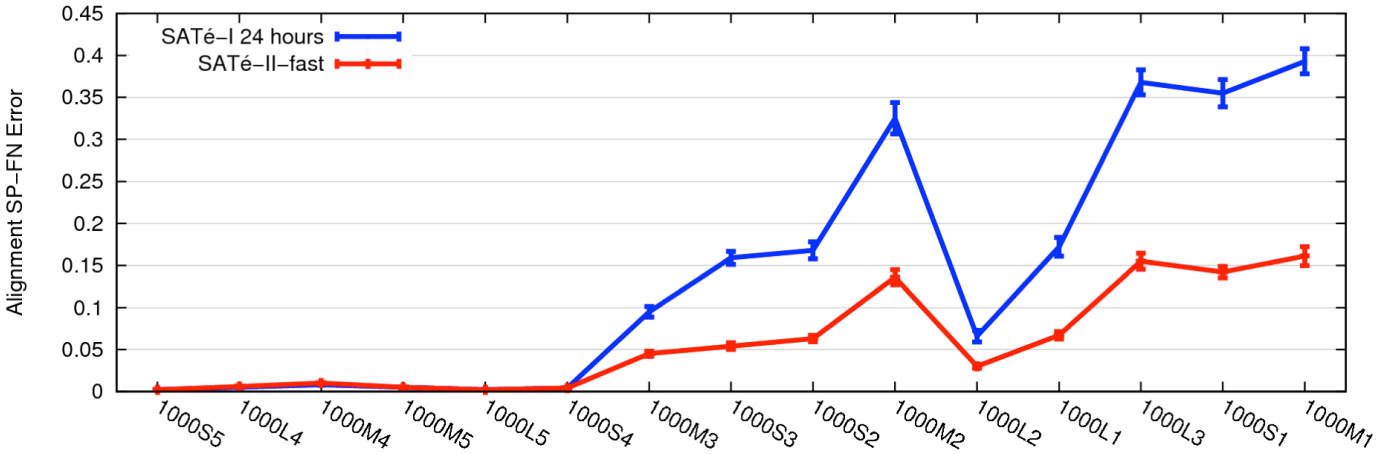

Missing Branch Rate (%)
RAxML(ClustalW)
RAxML(Muscle)
RAxML(MAFFT)
RAxML(Prank+GT)
RAxML(Opal)
Original SATé
"Next" SATé
RAxML(TrueAln)
Alignment SP–FN Error
ClustalW
Muscle
MAFFT
Prank+GT
Original SATé
Opal
"Next" SATé
1000 taxon models ranked by difficulty

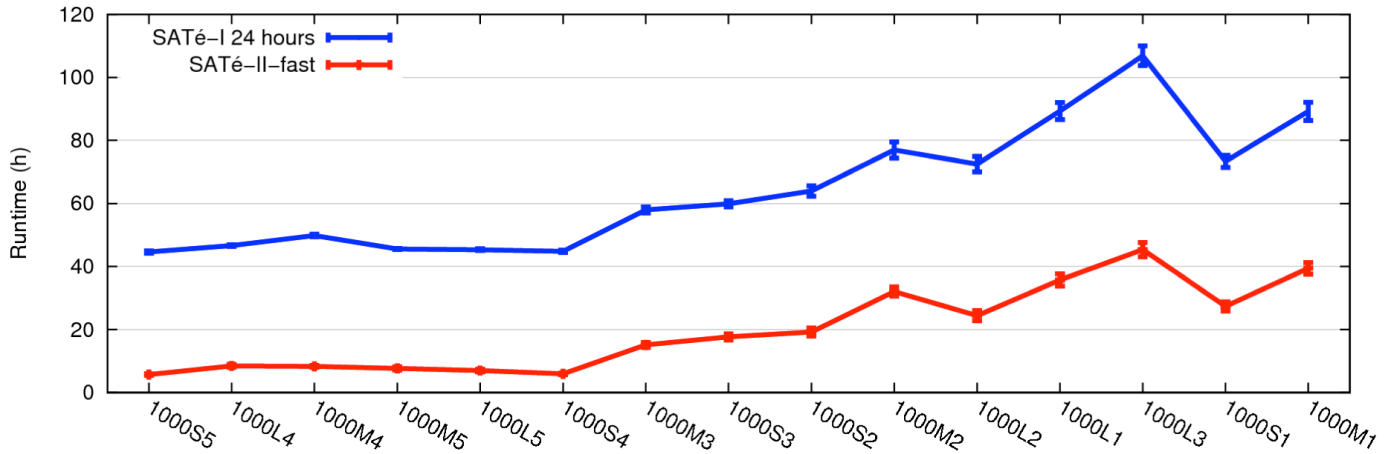
SATé-I vs. SATé-II

SATé-II

- Faster and more accurate than SATé-I
- Longer analyses or use of ML to select tree/alignment pair slightly better results





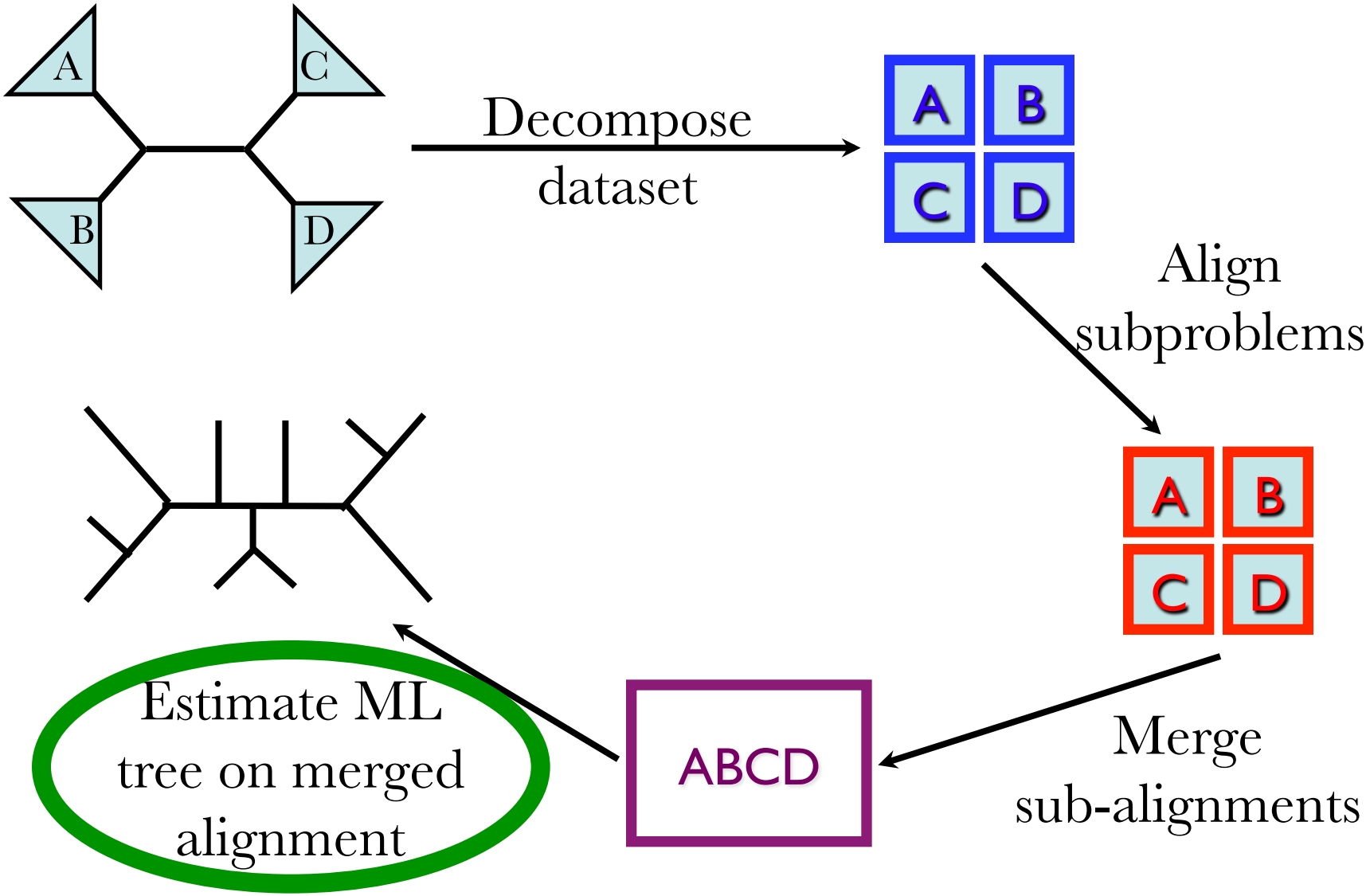


Brief discussion

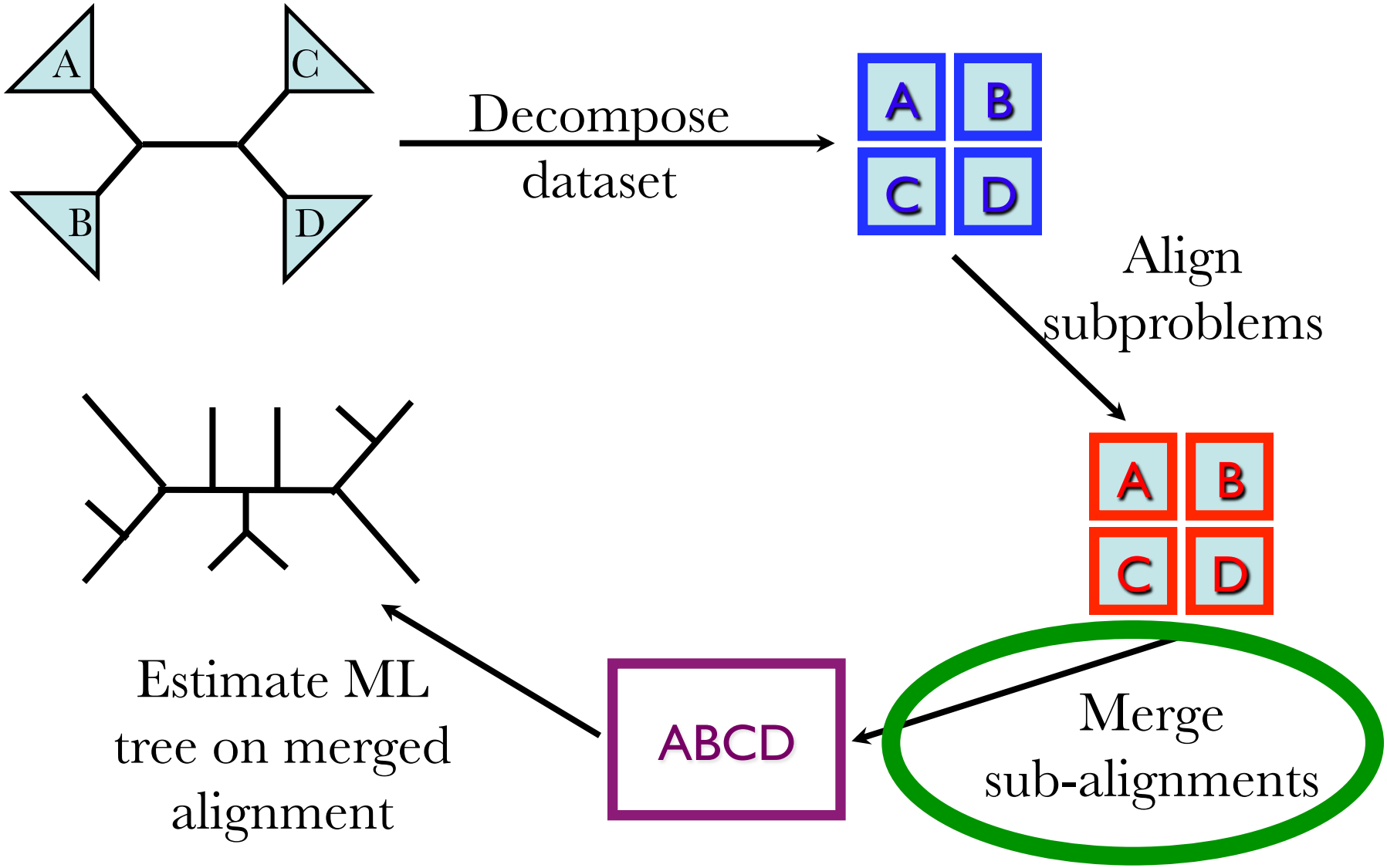
- SATé "boosts" the base methods. Results shown are for SATé used with MAFFT. Similar improvements seen for use with other MSA methods (e.g., Prank, Opal, Muscle, ClustalW).

- Biological datasets: Similar results on large benchmark datasets (structurally-based rRNA alignments)

Limitations



Limitations



II: UPP: Ultra-large alignment using SEPP[1]

Objective: highly accurate multiple sequence alignments and trees on ultra-large datasets

Authors: Nam Nguyen, Siavash Mirarab, and Tandy Warnow

In preparation – expected submission Fall 2013

[1] SEPP: SATe-enabled phylogenetic placement, Nguyen, Mirarab, and Warnow, PSB 2012

UPP: basic idea

Input: set S of unaligned sequences

Output: alignment on S

- Select random subset X of S
- Estimate “backbone” alignment A and tree T on X
- Independently align each sequence in S-X to A
- Use transitivity to produce multiple sequence alignment A* for entire set S

Input: Unaligned Sequences

```
S1  = AGGCTATCACCTGACCTCCAAT
S2  = TAGCTATCACGACCGCGCT
S3  = TAGCTGACCGCGCT
S4  = TACTCACGACCGACAGCT
S5  = TAGGTACAACCTAGATC
S6  = AGATACGTCGACATATC
```

Step 1: Pick random subset (backbone)

```
S1  = AGGCTATCACCTGACCTCCAAT
S2  = TAGCTATCACGACCGCGCT
S3  = TAGCTGACCGCGCT
S4  = TACTCACGACCGACAGCT
S5  = TAGGTACAACCTAGATC
S6  = AGATACGTCGACATATC
```

Step 2: Compute backbone alignment

```
S1  = -AGGCTATCACCTGACCTCCA-AT
S2  = TAG-CTATCAC--GACCGC--GCT
S3  = TAG-CT-------GACCGC--GCT
S4  = TAC----TCAC--GACCGACAGCT
S5  = TAGGTAAAACCTAGATC
S6  = AGATAAAACTACATATC
```

Step 3: Align each remaining sequence to backbone

First we add S5 to the backbone alignment

```
S1  = -AGGCTATCACCTGACCTCCA-AT-
S2  = TAG-CTATCAC--GACCGC--GCT-
S3  = TAG-CT-------GACCGC--GCT-
S4  = TAC----TCAC--GACCGACAGCT-
S5  = TAGG---T-A-CAA-CCTA--GATC
```

Step 3: Align each remaining sequence to backbone

Then we add S6 to the backbone alignment

```
S1  = -AGGCTATCACCTGACCTCCA-AT-
S2  = TAG-CTATCAC--GACCGC--GCT-
S3  = TAG-CT-------GACCGC--GCT-
S4  = TAC----TCAC--GACCGACAGCT-
S6  = -AG---AT-A-CGTC--GACATATC
```

Step 4: Use transitivity to obtain MSA on entire set

```
S1  = -AGGCTATCACCTGACCTCCA-AT--
S2  = TAG-CTATCAC--GACCGC--GCT--
S3  = TAG-CT-------GACCGC--GCT--
S4  = TAC----TCAC--GACCGACAGCT--
S5  = TAGG---T-A—CAA-CCTA--GATC-
S6  = -AG---AT-A-CGTC--GACATAT-C
```

UPP: details

Input: set S of unaligned sequences

Output: alignment on S

- Select random subset X of S
- Estimate “backbone” alignment A and tree T on X
- Independently align each sequence in S-X to A
- Use transitivity to produce multiple sequence alignment A* for entire set S

UPP: details

Input: set S of unaligned sequences

Output: alignment on S

- Select random subset X of S
- Estimate “backbone” alignment A and tree T on X
- Independently align each sequence in S-X to A
- Use transitivity to produce multiple sequence alignment A* for entire set S

How to align sequences to a backbone alignment?

Standard machine learning technique:

- Build HMM (Hidden Markov Model) for backbone alignment, and use it to align remaining sequences

We use HMMER (Sean Eddy, HHMI) for this purpose

Using HMMER

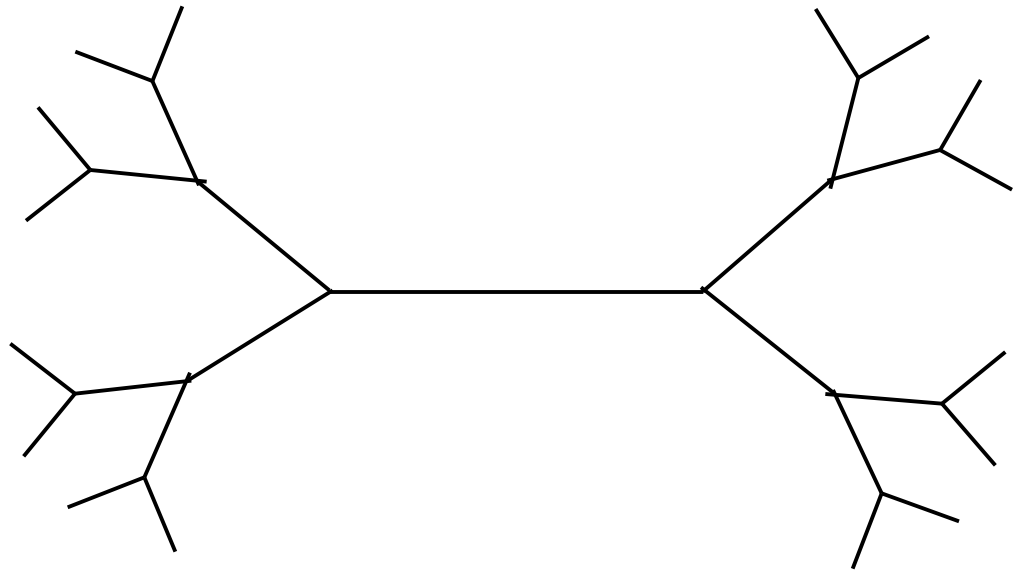
Using HMMER works well…

Using HMMER

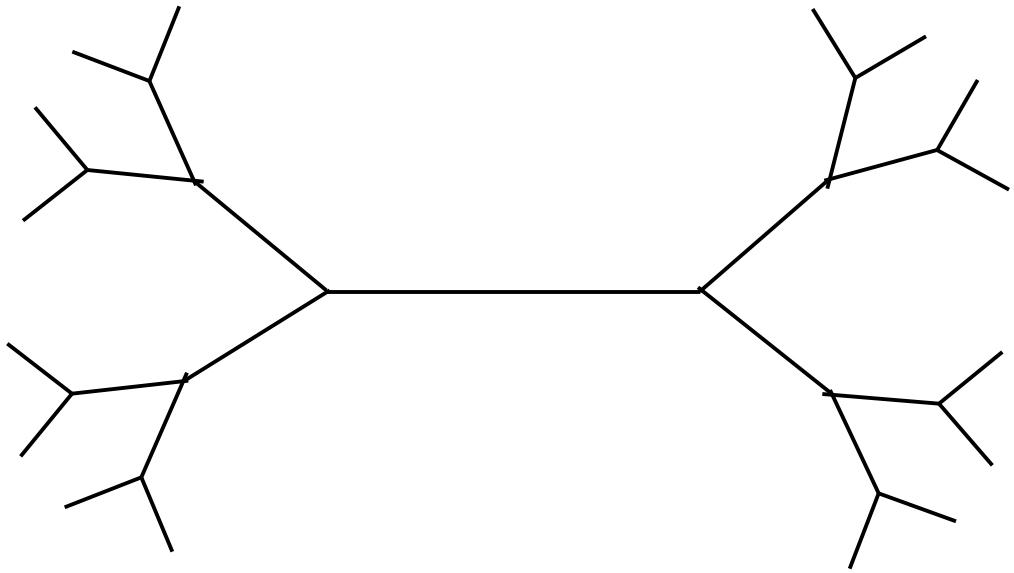
Using HMMER works well…except when the dataset is big!

Using HMMER to add sequences to an existing alignment

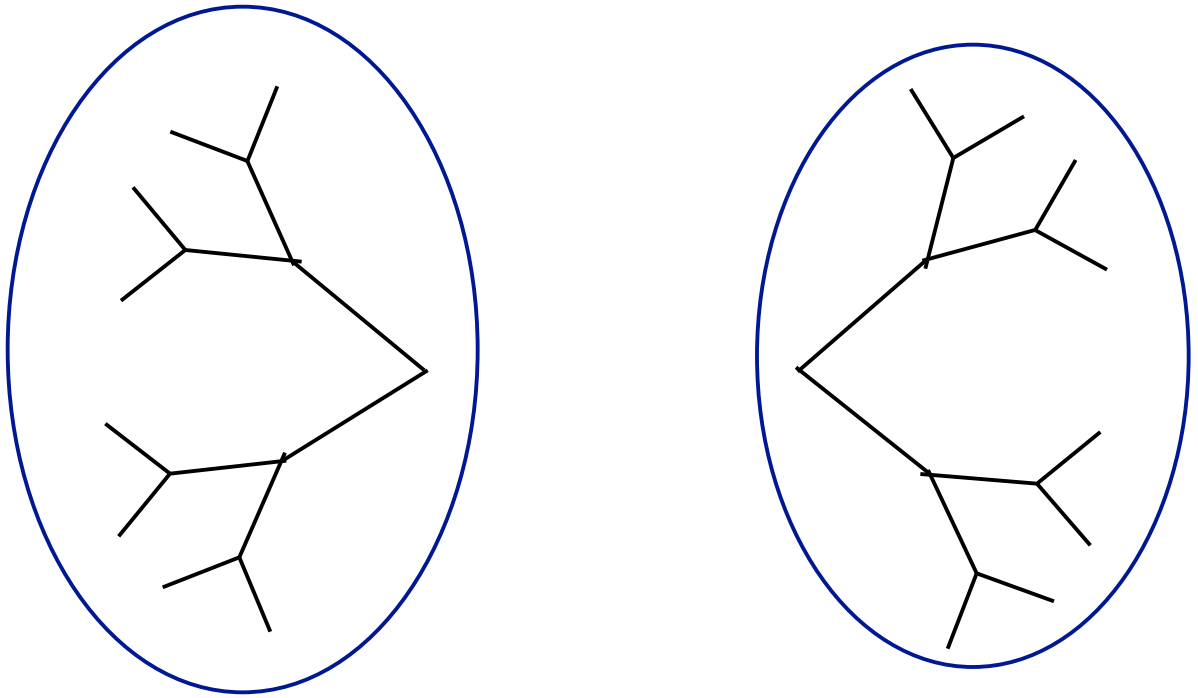
1) build one HMM for the backbone alignment
2) Align sequences to the HMM, and insert into backbone alignment



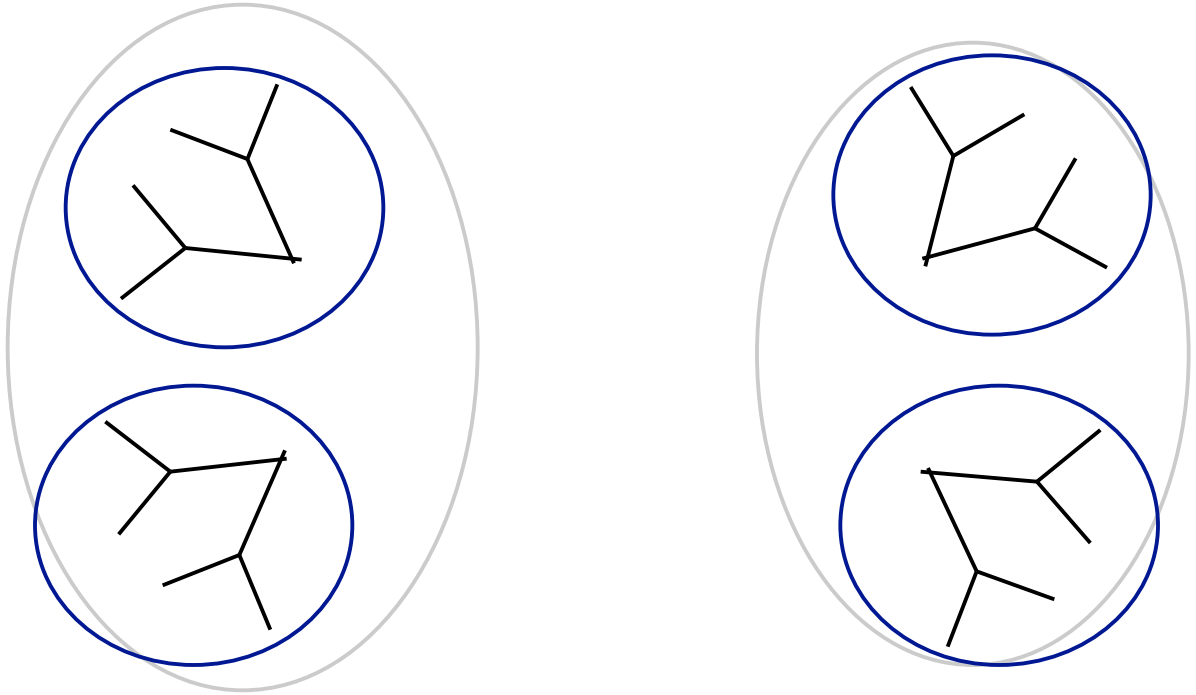
One Hidden Markov Model for the entire alignment?



Or 2 HMMs?



Or 4 HMMs?



UPP(x,y)

- Pick random subset X of size **x**
- Compute alignment A and tree T on X
- Use SATé decomposition on T to partition X into small “alignment subsets” of at most **y** sequences
- Build HMM on each alignment subset using HMMBUILD
- For each sequence s in S-X,
  - use HMMALIGN to produce alignment of s to each subset alignment and note the score of each alignment.
  - Pick the subset alignment that has the best score, and align s to that subset alignment.
  - Use transitivity to align s to the backbone alignment.

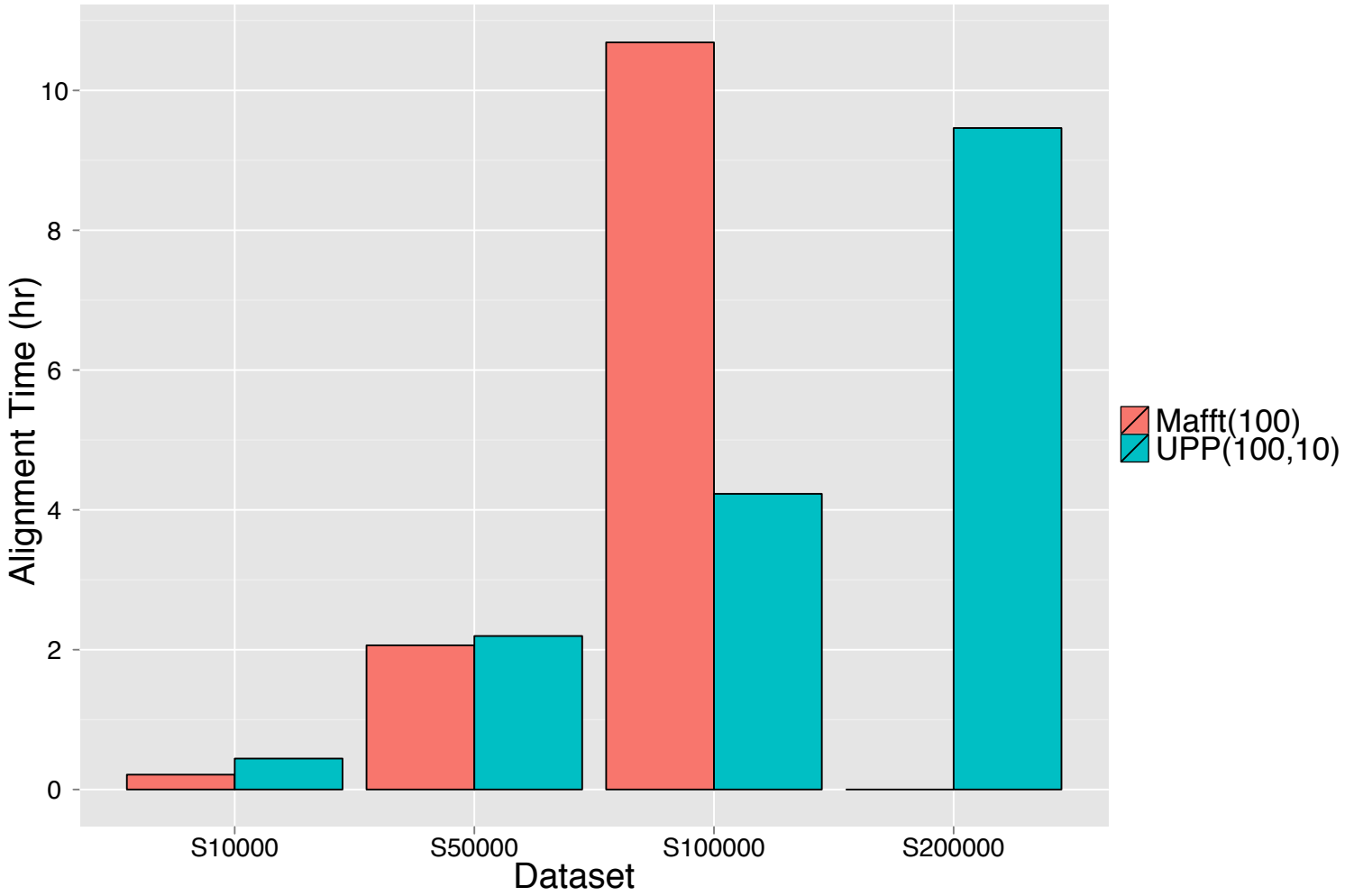
UPP design

- Size of backbone matters – small backbones are sufficient for most datasets (except for ones with very high rates of evolution). Random backbones are fine.
- Number of HMMs matters, and depends on the rate of evolution and number of taxa.
- Backbone alignment and tree matter; we use SATé.

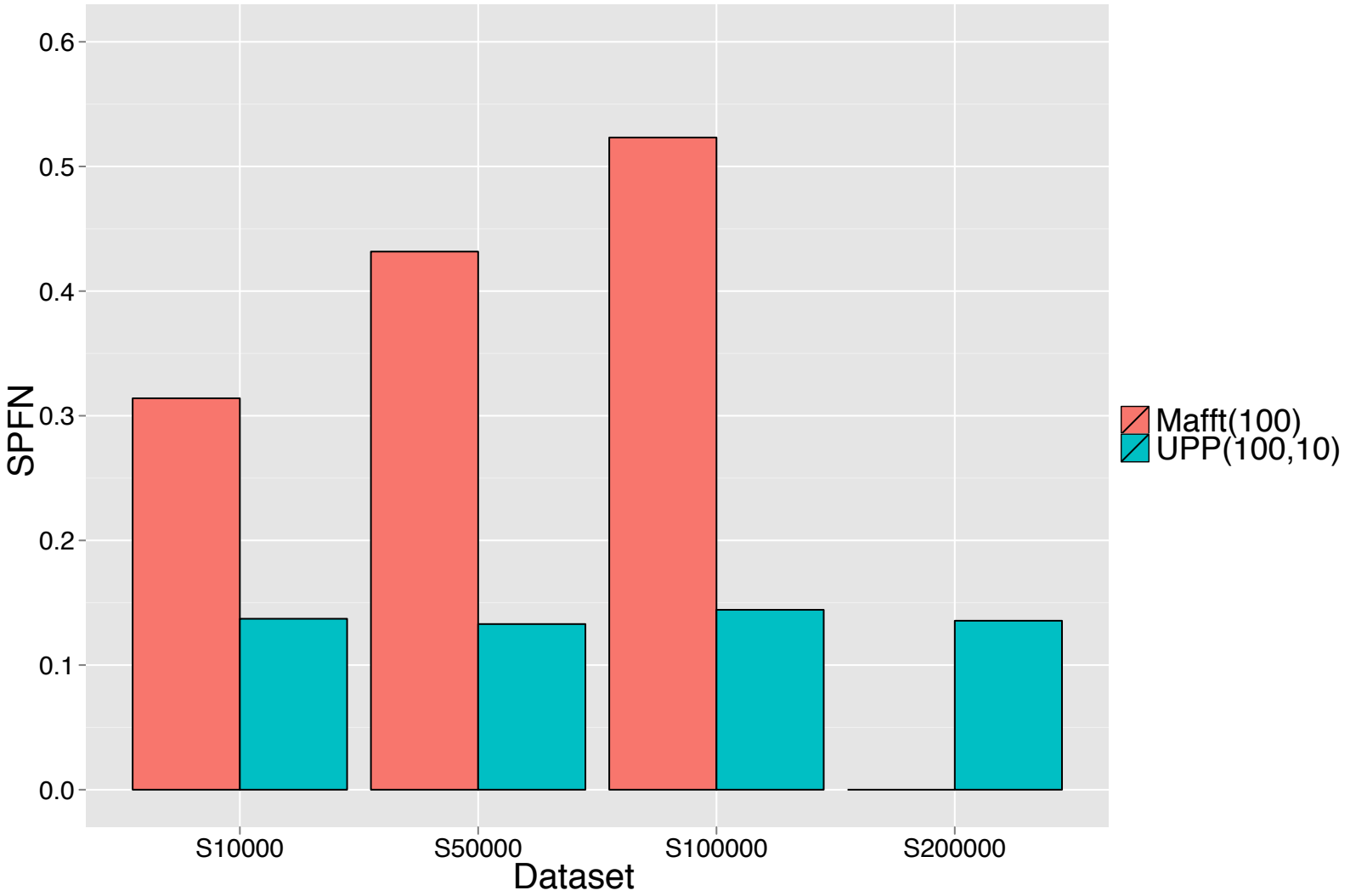
Evaluation of UPP

- Simulated Datasets: 10,000 to 1,000,000 sequences (RNASim, Junhyong Kim, U Penn)
- Biological datasets with reference alignments (Gutell's CRW data with up to 28,000 sequences)
- Criteria: Alignment error (SP-FN and SP-FP), tree error, and time

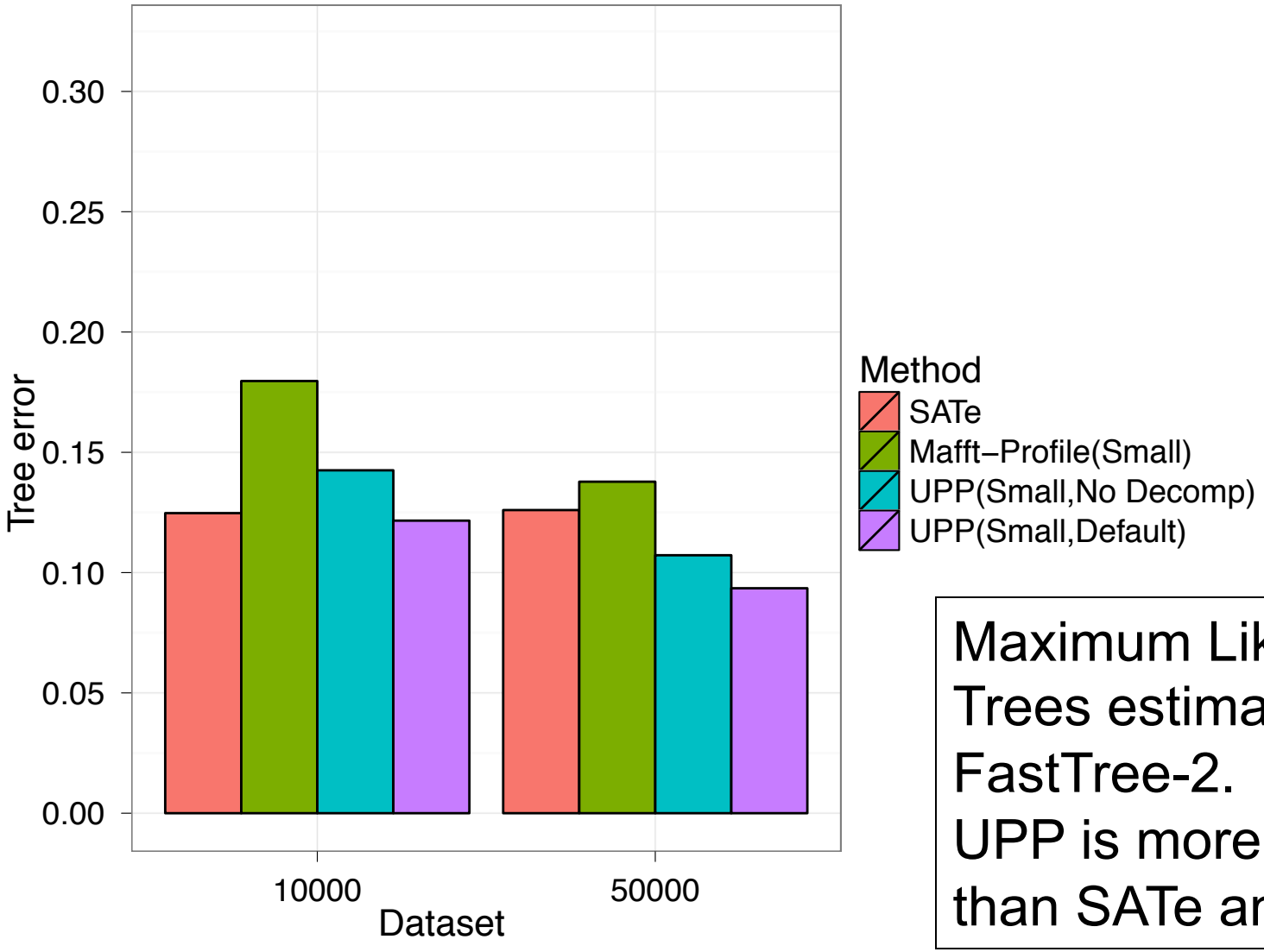
UPP vs. MAFFT-profile Running Time



UPP vs. MAFFT-profile Alignment Error

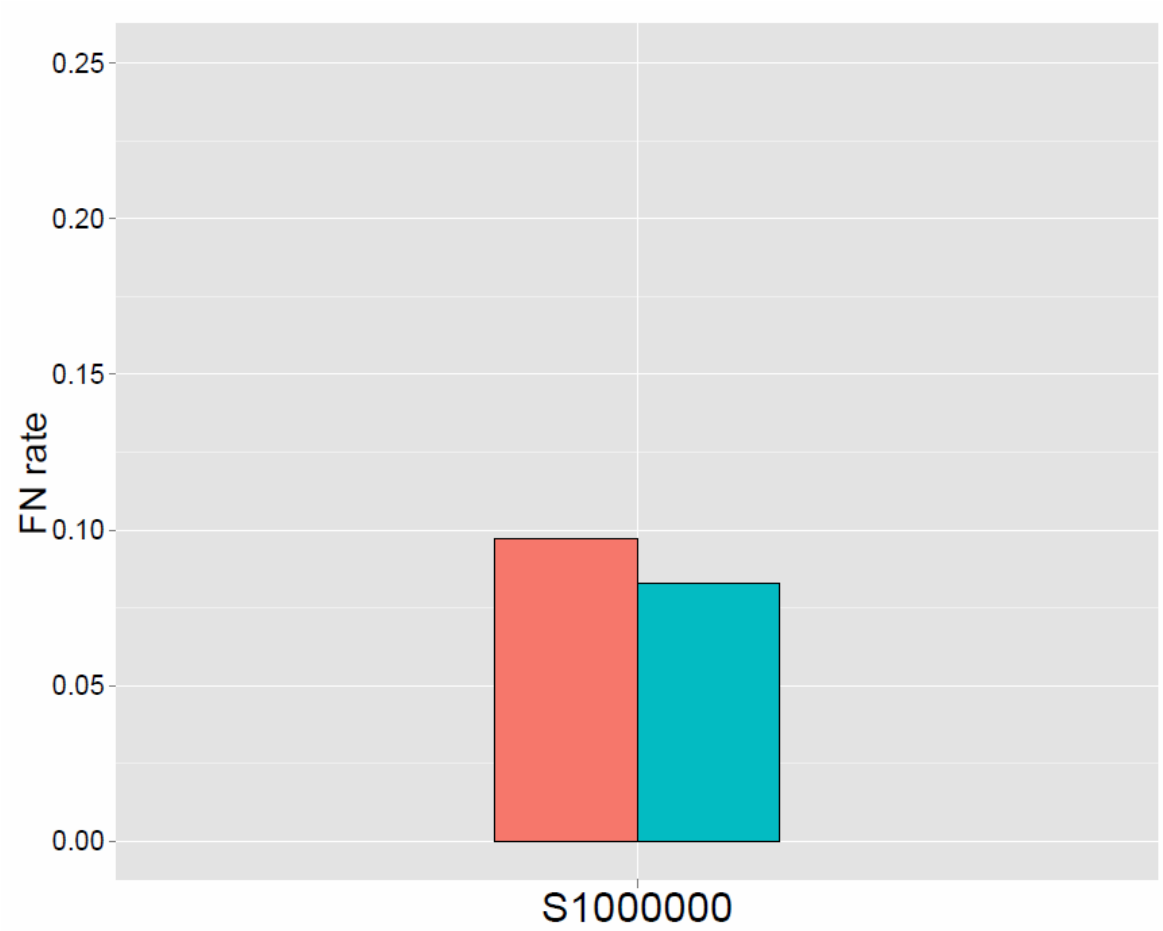


Tree Error on 10K and 50K RNASim datasets



Maximum Likelihood Trees estimated using FastTree-2.
UPP is more accurate than SATe and MAFFT

One Million Sequences: Tree Error



Note: Decomposition improves tree accuracy

UPP performance

- UPP is very fast, parallelizable, and scalable. UPP can analyze very large datasets (up to 1,000,000 sequences so far).
- On very large nucleotide datasets (>25,000 sequences)
  - UPP is generally the only method that can run on very large datasets in reasonable timeframes.
  - UPP alignments and trees are more accurate than other methods (e.g., MAFFT-Profile, Muscle, and SATé),
- On large (but not huge) datasets (1000-25000 sequences)
  - UPP alignments and trees are comparable to SATé and better than other alignment methods
- On smaller datasets (<1000 sequences)
  - UPP is either best or close to best for alignment accuracy, but its trees are generally less accurate than SATé trees.

UPP “HMM Family” technique

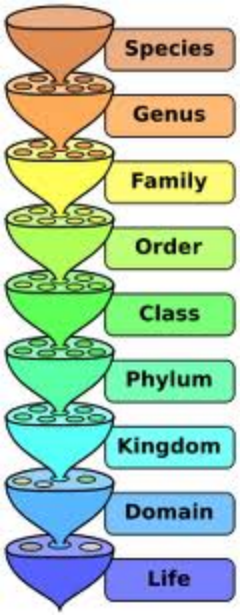
- Using multiple HMMs to represent a multiple sequence alignment (each on a different subset of the sequences) is key.
- Random subsets are not as helpful as tree-based decomposition.
- Note: the decomposition does not necessarily produce “clades”.

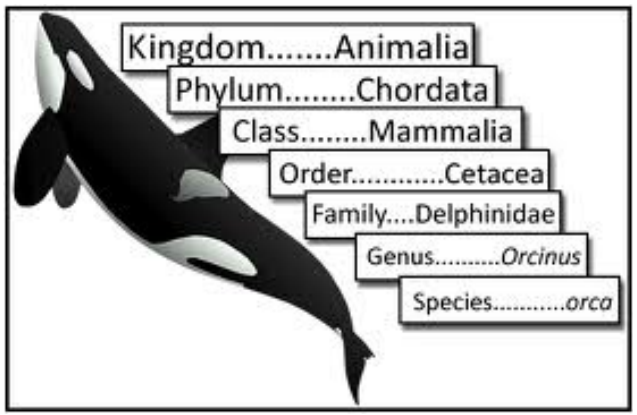
Two other applications

- SEPP: SATé-enabled phylogenetic placement (PSB 2012)
- TIPP: Taxonomic Identification using SEPP (in preparation)

Part III: Metagenomic Taxon Identification

Objective: classify short reads in a metagenomic sample

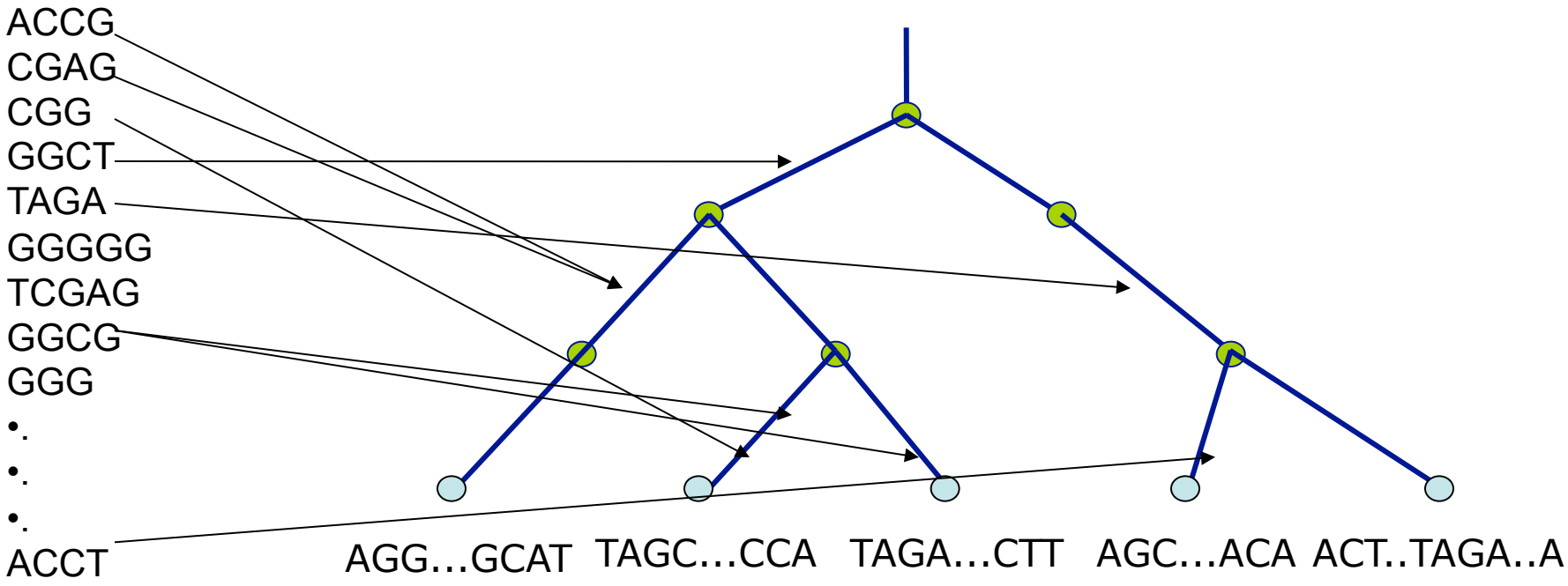




Phylogenetic Placement

Fragmentary sequences from some gene

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



SEPP

- **SEPP: SATé-enabled Phylogenetic Placement**, by Mirarab, Nguyen, and Warnow

- Pacific Symposium on Biocomputing, 2012 (special session on the Human Microbiome)

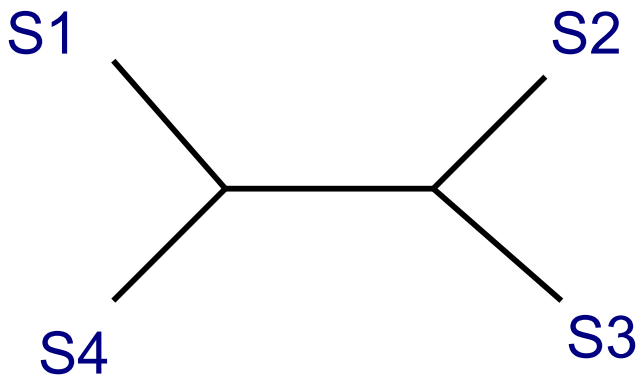
Phylogenetic Placement

Step 1: Align each query sequence to backbone alignment

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

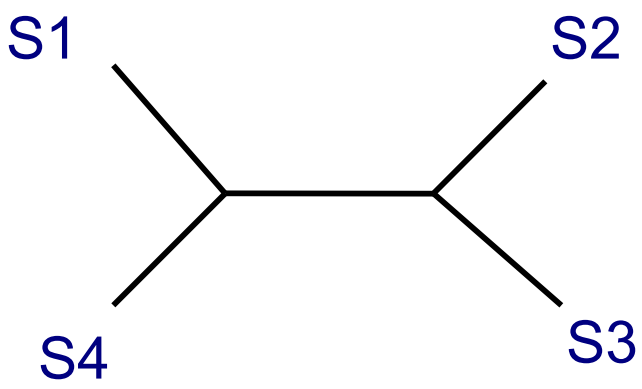
Align Sequence

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



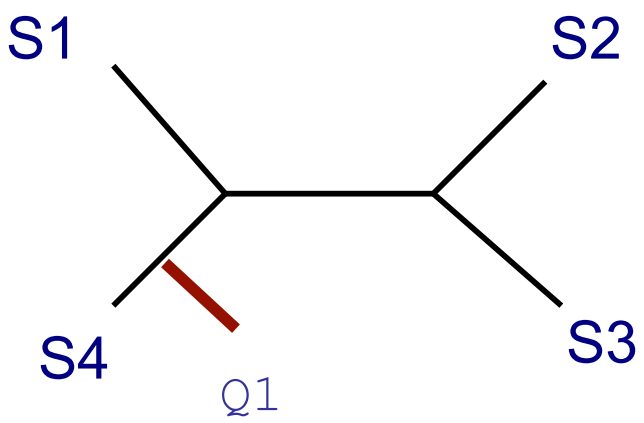
Align Sequence

```
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--------
```

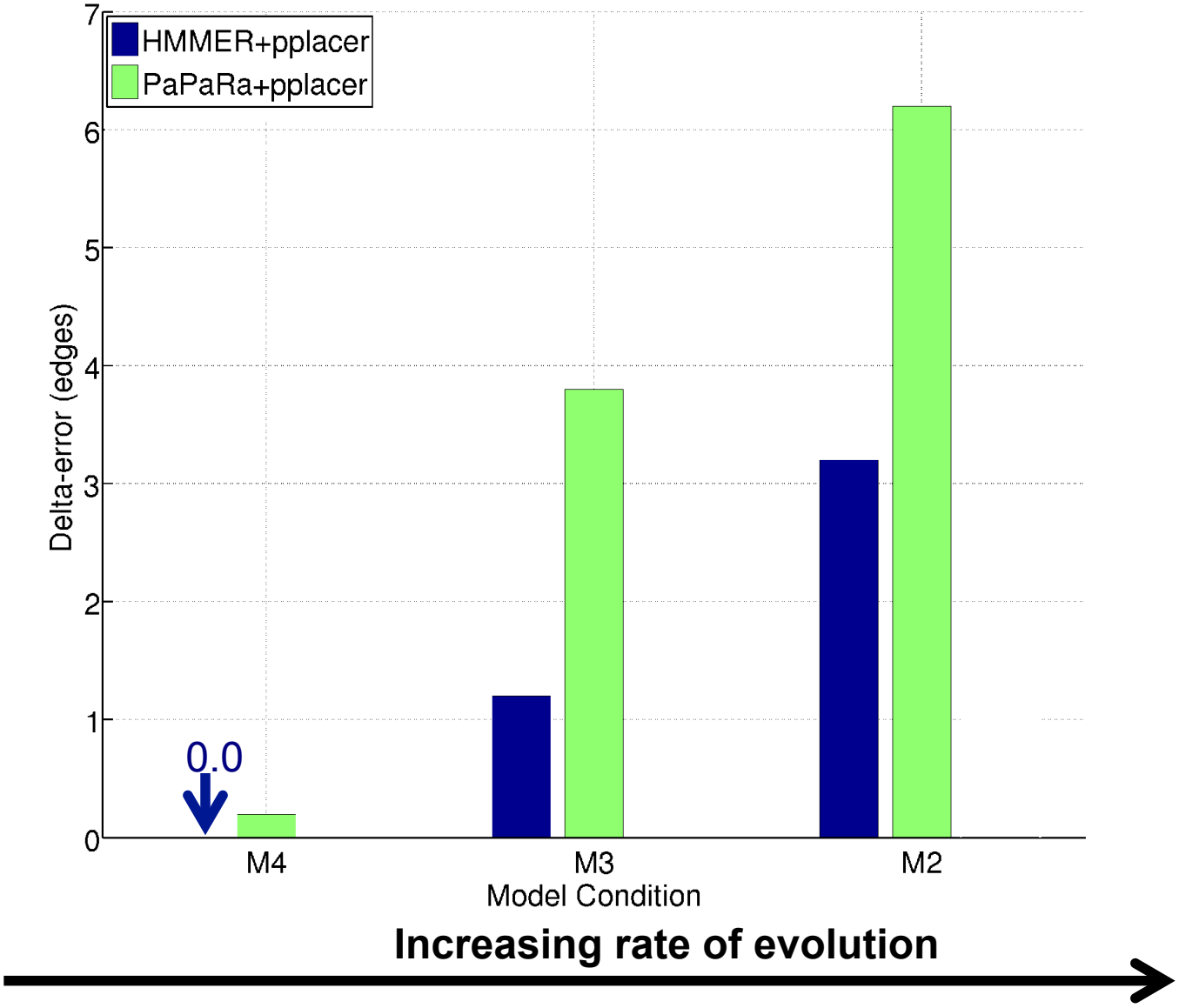


Phylogenetic Placement

- 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)

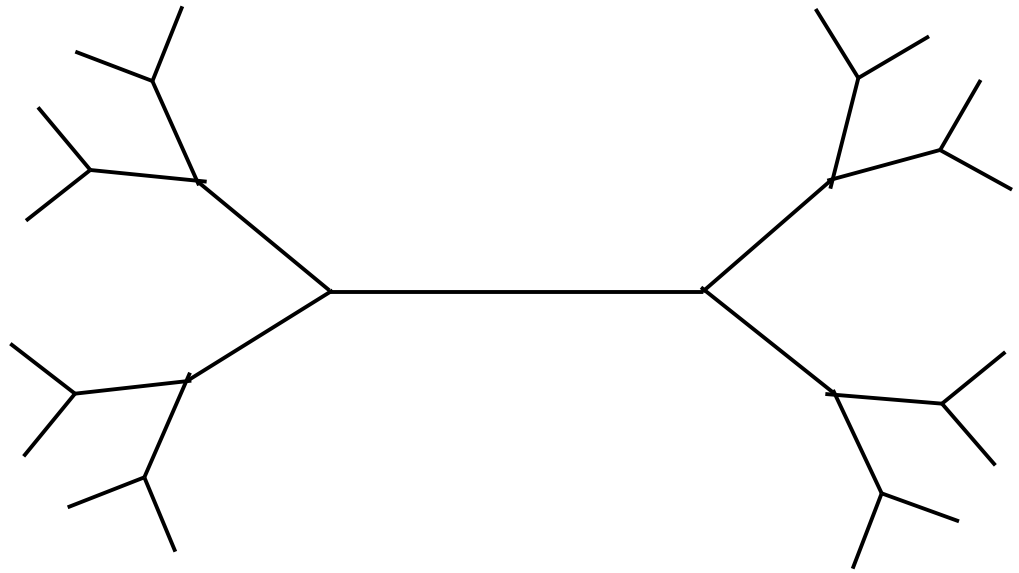
Note: pplacer and EPA use maximum likelihood, and are reported to have the same accuracy.

HMMER vs. PaPaRa

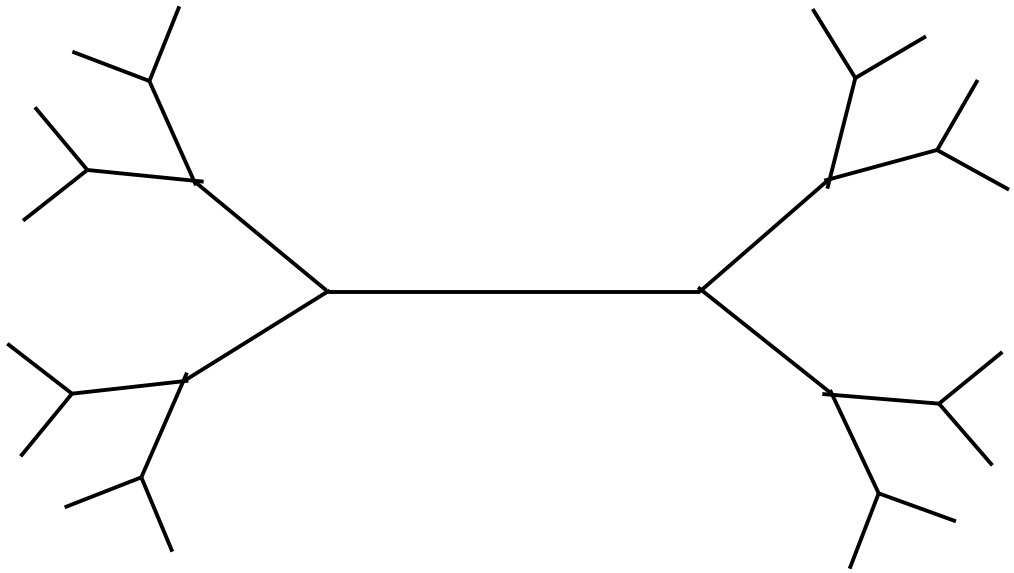


HMMER+pplacer:

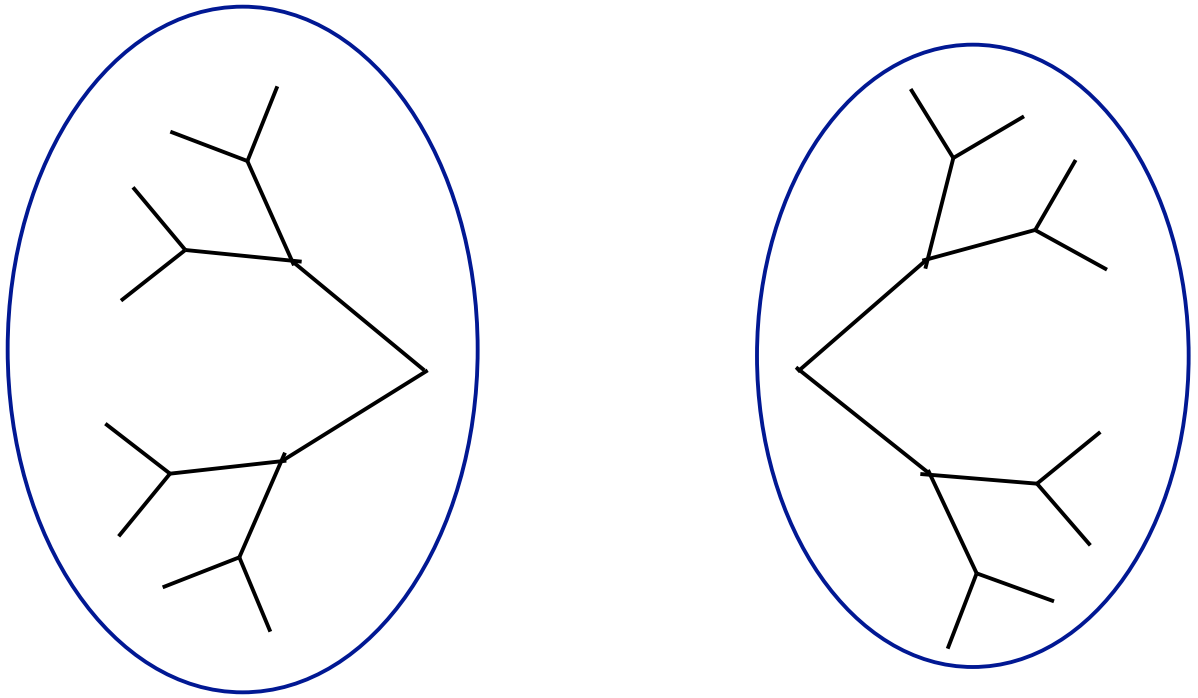
1) build one HMM for the entire alignment
2) Align fragment to the HMM, and insert into alignment
3) Insert fragment into tree to optimize likelihood



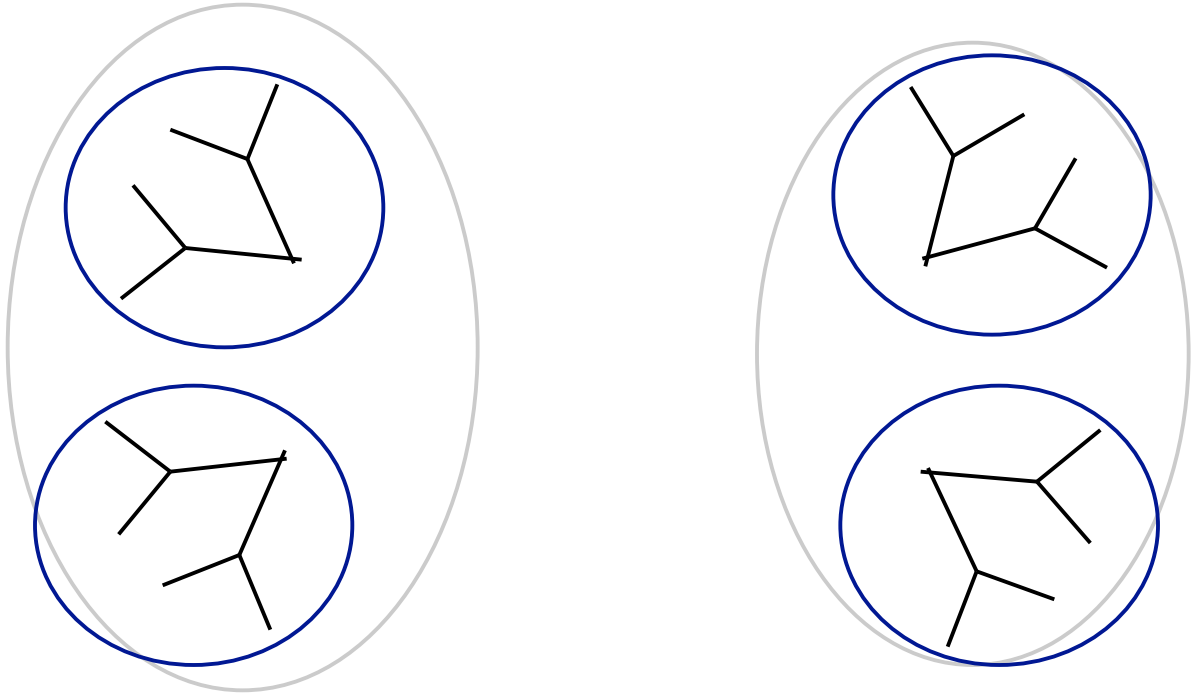
One Hidden Markov Model for the entire alignment?



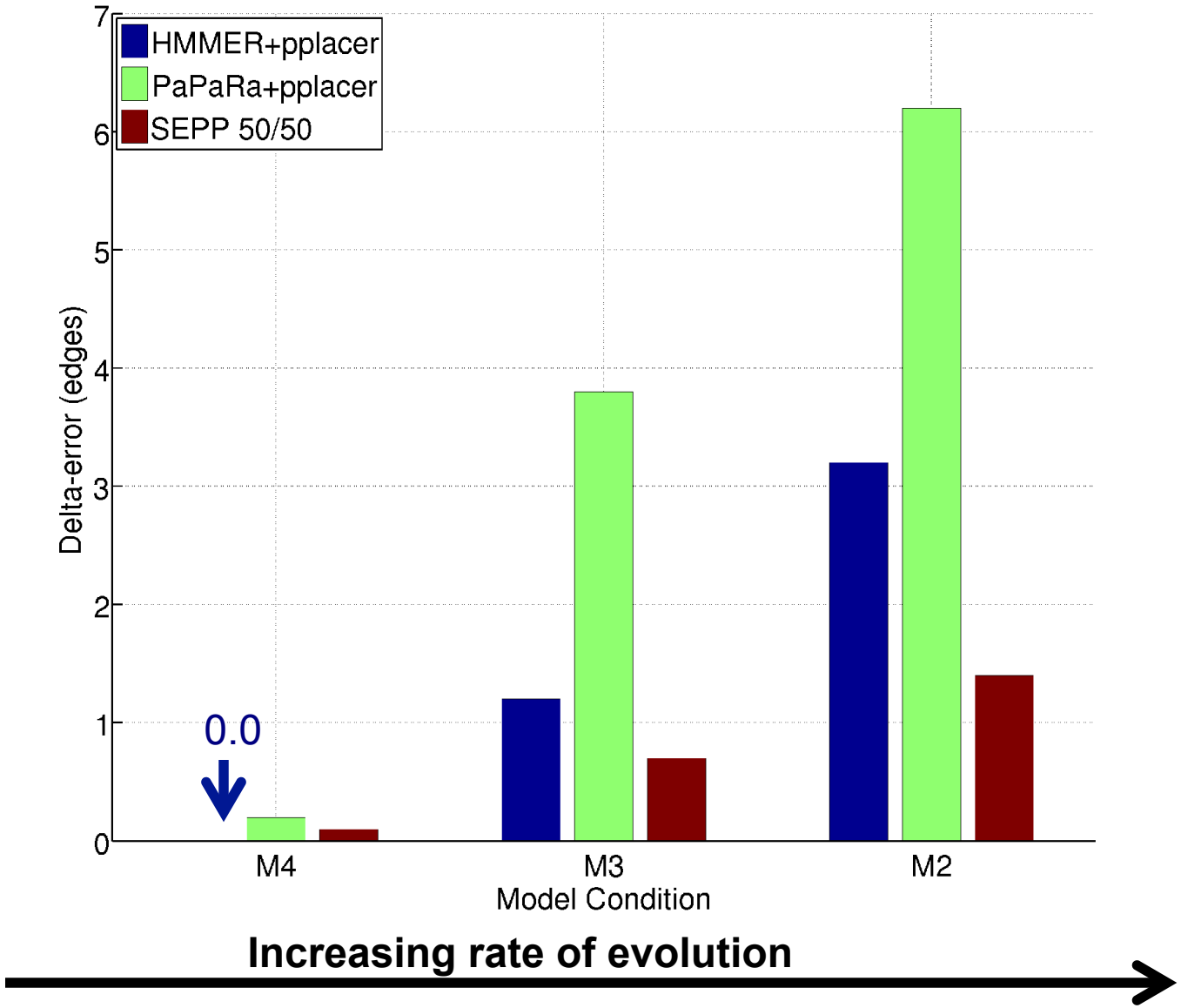
Or 2 HMMs?



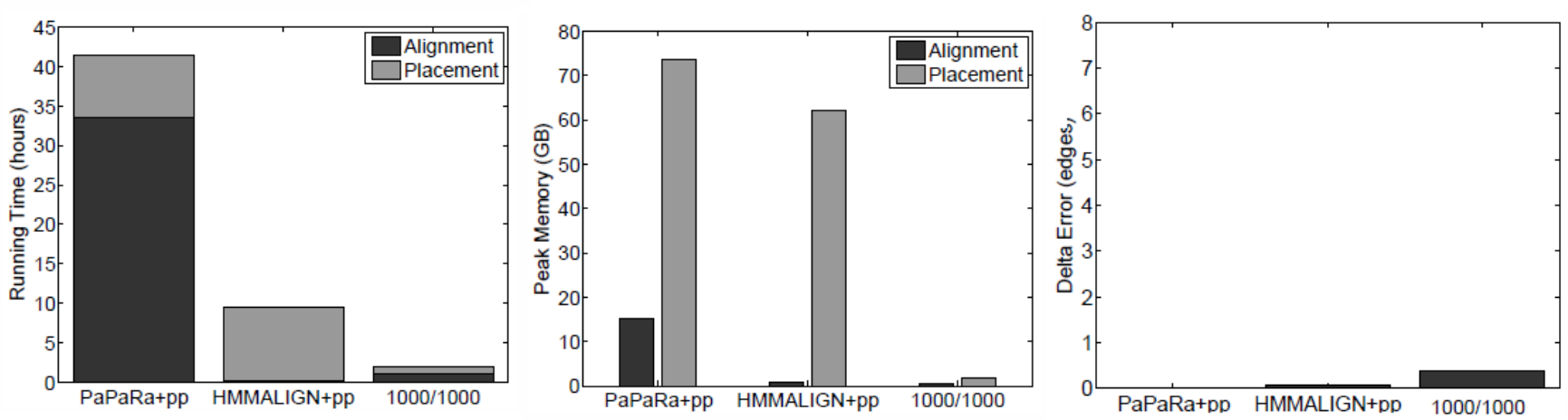
Or 4 HMMs?



SEPP(10%), based on ~10 HMMs



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

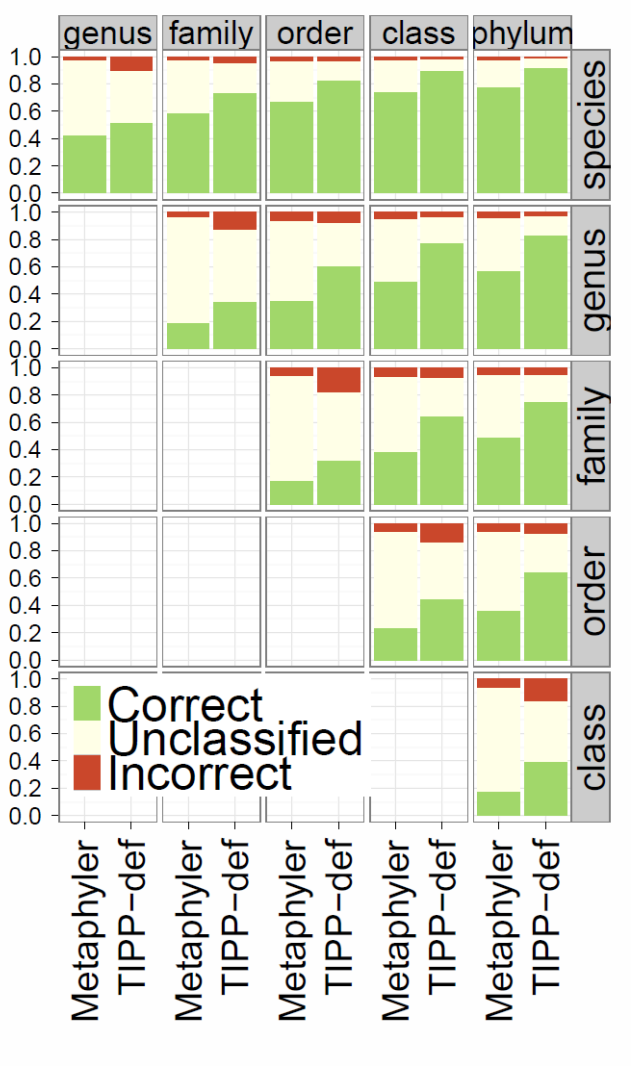
TIPP: SEPP + statistics

SEPP has high recall but low precision (classifies almost everything)

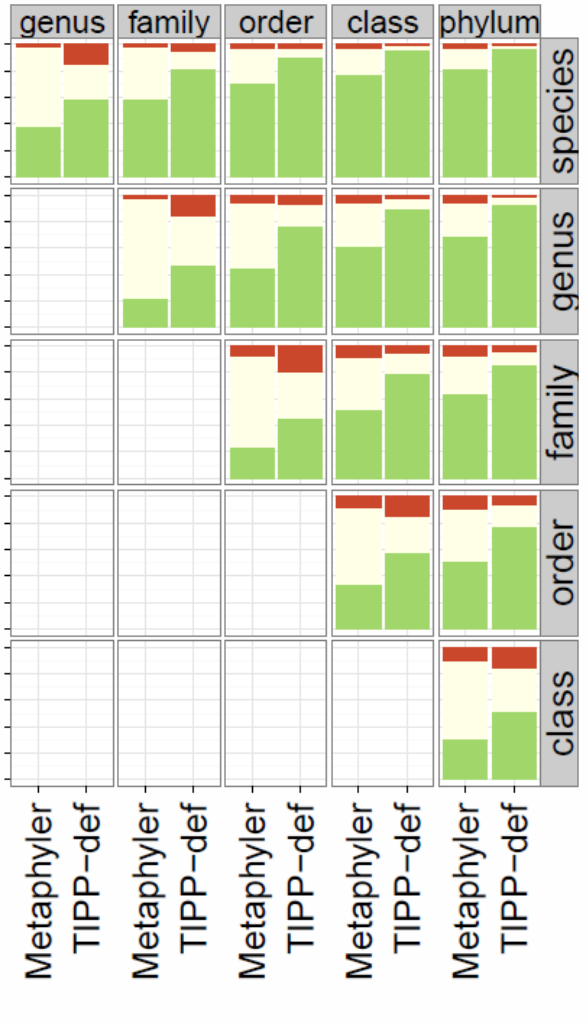
TIPP: dramatically reduces false positive rate with small reduction in true positive rate, by considering uncertainty in alignment (HMMER) and placement (pplacer)

Leave-one-out on 30 marker genes

Illumina Error model



454 Error model



Summary: 4 Phylogenetic "boosters"

- SATé: co-estimation of alignments and trees
- UPP: ultra-large multiple sequence alignment
- TIPP: taxonomic identification of short reads
- SEPP: phylogenetic placement

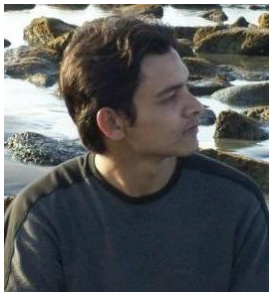
Conclusions

- Divide-and-conquer helps improve accuracy, speed, and scalability of phylogenetic estimation and alignment estimation methods.

- The use of multiple HMMs, based on tree-decompositions, may be generally useful for classification problems.

Warnow Laboratory





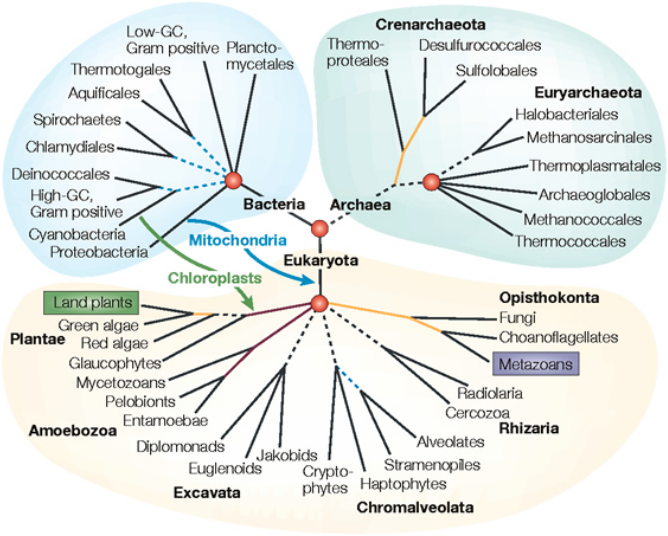
PhD students: Siavash Mirarab, Nam Nguyen, and Md. S. Bayzid

Undergrad: Keerthana Kumar

Lab Website: http://www.cs.utexas.edu/users/phylo

Funding: Guggenheim Foundation,Packard, HHMI, NSF, Microsoft Research New England, David Bruton Jr. Centennial Professorship, and TACC (Texas Advanced Computing Center)

The Tree of Life: *Multiple Challenges*



Nature Reviews | Genetics

Large datasets:

100,000+ sequences

10,000+ genes

"BigData" complexity

Challenges:

Ultra-large multiple-sequence alignment

Alignment-free phylogeny estimation

Supertree estimation

Estimating species trees from incongruent gene trees

Absolute fast converging methods

Genome rearrangement phylogeny

Reticulate evolution

Visualization of large trees and alignments

Data mining techniques to explore multiple optima

Phylogenetic "boosters"

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

Techniques: divide-and-conquer, iteration, chordal graph algorithms, 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 and 2012)
- SuperFine-boosting for supertree methods (2012)
- DACTAL: almost alignment-free phylogeny estimation methods (2012)
- SEPP-boosting for phylogenetic placement of short sequences (2012)
- UPP-boosting for alignment methods (in preparation)
- PASTA-boosting for alignment methods (submitted)
- TIPP-boosting for metagenomic taxon identification (in preparation)
- Bin-and-conquer for coalescent-based species tree estimation (2013)

Algorithmic Strategies

- Divide-and-conquer
- Chordal graph decompositions
- Iteration
- Multiple HMMs
- Bin-and-conquer (technique used for improving species tree estimation from multiple gene trees, Bayzid and Warnow, Bioinformatics 2013)

Other Current Research

- Large-scale alignment (PASTA)
- Coalescent-based species tree estimation
- Alignment and phylogeny estimation for fragmentary data
- Metagenomic analysis

PASTA (in preparation)

- **Practical Alignments using SATe and TrAnsitivity**
- Authors: Siavash Mirarab and Tandy Warnow
- Key idea: Use transitivity to extend overlapping alignments

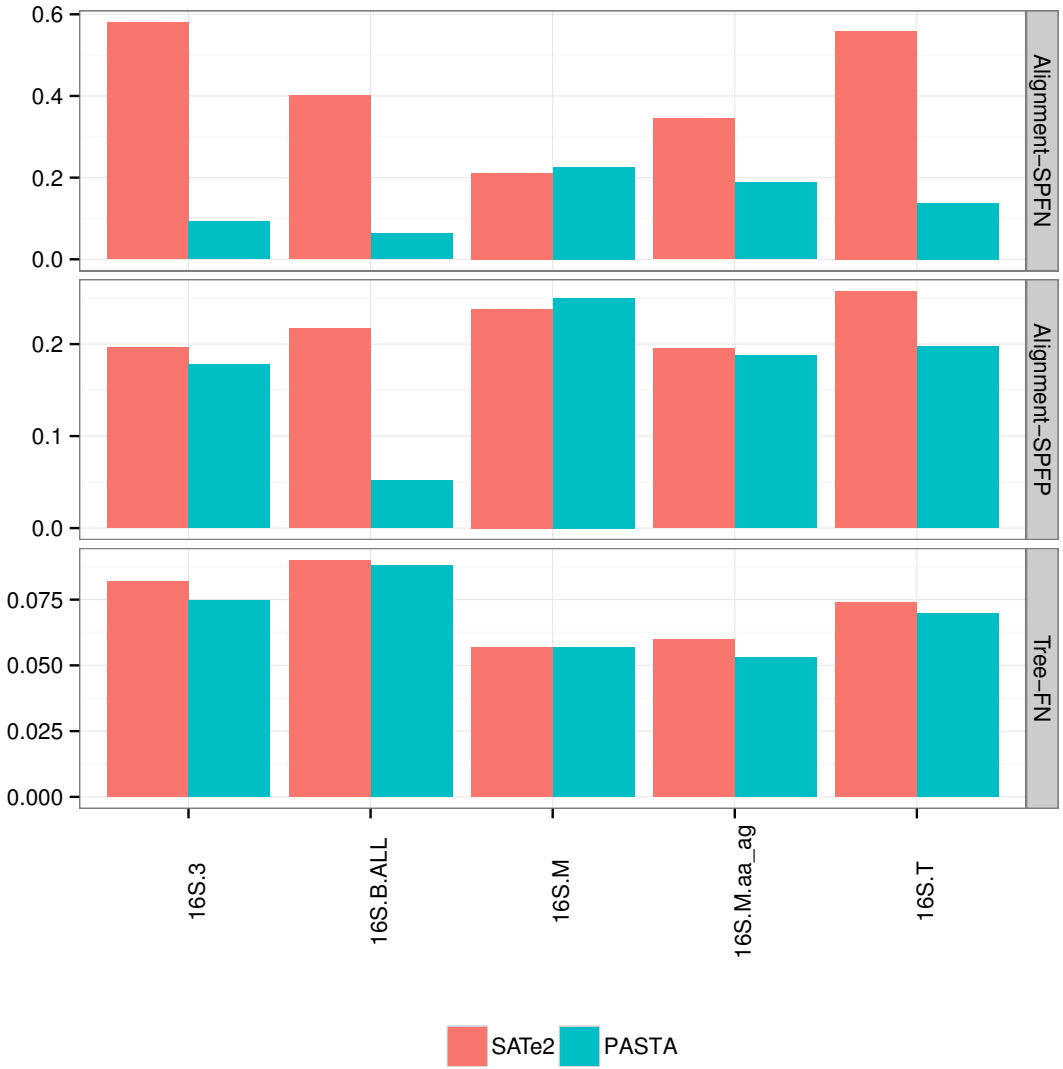
Part IV: UPP: Ultra-large alignment using SEPP

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

PASTA vs. SATe-2:
better alignments, comparable trees



Benchmark datasets:

Gutell's rRNA with structurally-based alignments, and trees estimated using maximum likelihood (FastTree-2).

Datasets range from 900 to 28,000 sequences.

Performance for PASTA

- Improved alignment and tree accuracy compared to SATé and UPP
- Faster than SATé but slower than UPP
- Highly scalable – up to 200,000 sequences
- Highly parallelizable

Submitted for publication

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.***

Alignment Error on 10K and 50K RNASim datasets