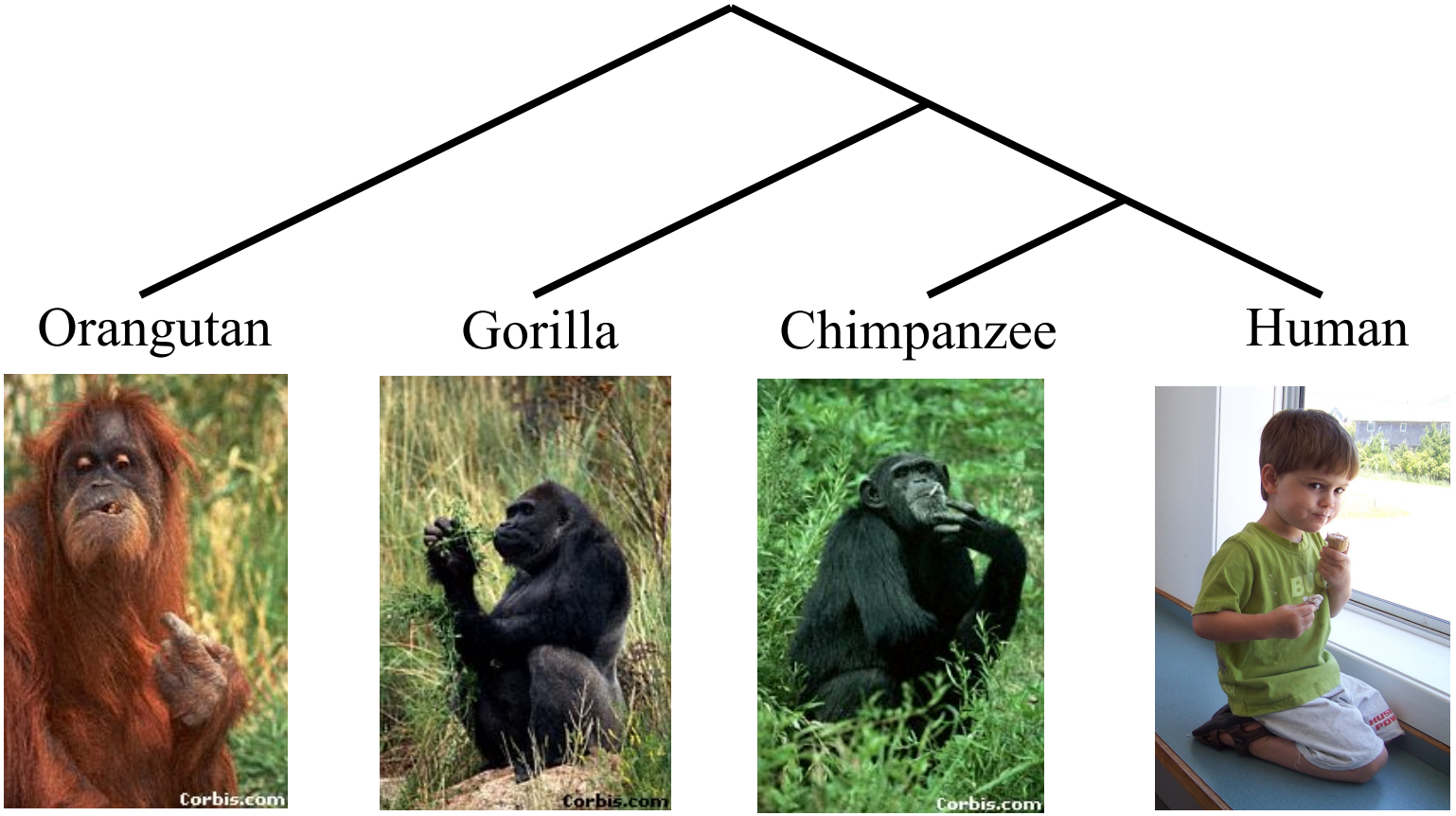
New techniques that "boost" methods for large-scale multiple sequence alignment and phylogenetic estimation

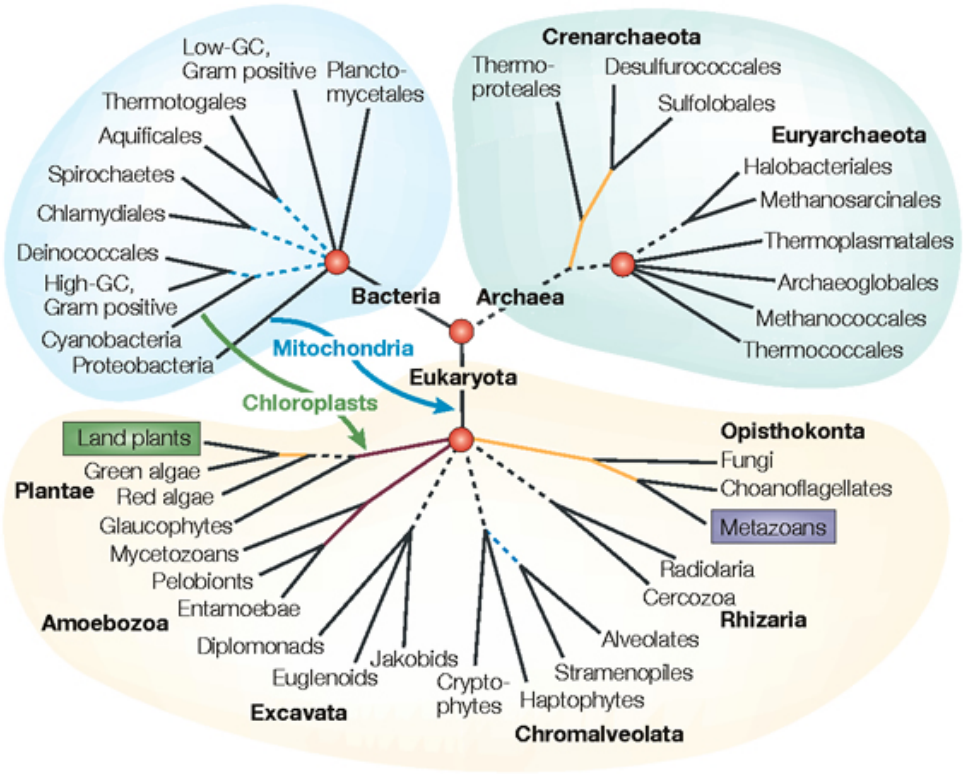
Tandy Warnow
Department of Computer Science
The University of Texas at Austin

Phylogeny (evolutionary tree)



From the Tree of the Life Website,
University of Arizona

Assembling the Tree of Life



Nature Reviews | Genetics

Evolution informs about everything in biology

- Big genome sequencing projects just produce data – so what?
- Evolutionary history relates all organisms and genes, and helps us understand and predict
  - interactions between genes (genetic networks)
  - drug design
  - predicting functions of genes
  - influenza vaccine development
  - origins and spread of disease
  - origins and migrations of humans

Challenges for Large-Scale Phylogeny and Alignment Estimation

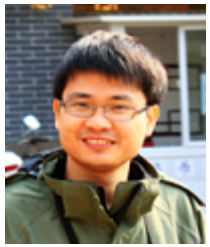
- NP-hard optimization problems and very large datasets (up to 500,000 taxa and tens of thousands of genes)
- Statistical estimation problems complicated by substantial error in the input data
- Much biological discovery enabled by accurate trees and alignments, but estimating highly accurate alignments and trees is difficult

Avian Phylogenomics Project

E.Jarvis, HHMI

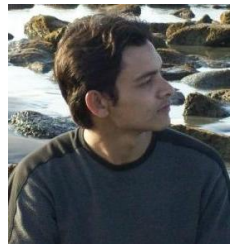
MTP Gilbert, Copenhagen

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 our new coalescent-based species tree method (and also “concatenation”)
- Multi-national team (20+ investigators)

Biggest challenges:

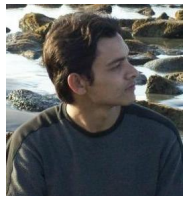
Estimating species tree from incongruent gene trees,
Poor phylogenetic signal in most genes

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, Nam Nguyen, and
UT-Austin

Md. S. Bayzid

- Transcriptomes of approx. 1200 species
- More than 13,000 gene families (most not single copy)
- Multi-institutional project (10+ universities)
- Gene trees and alignments estimated using SATé, UPP, and PASTA

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

Research Agenda

Major scientific goals:

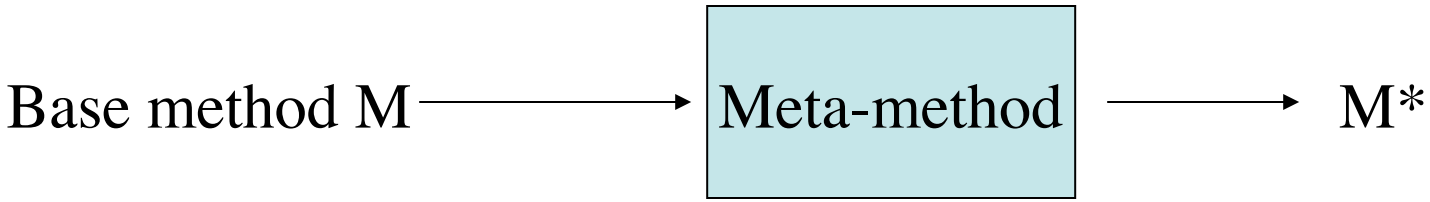
- Develop methods that produce more accurate alignments and phylogenetic estimations for *difficult-to-analyze datasets*
- Produce mathematical theory for statistical inference under complex models of evolution
- Develop novel machine learning techniques to boost the performance of classification methods

Software that:

- Can run efficiently on *desktop* computers on large datasets
- Can analyze ultra-large datasets (100,000+) using multiple processors
- Is freely available in *open source* form, with biologist-friendly GUIs

Meta-Methods

- Meta-methods "boost" the performance of base methods (e.g., for phylogeny or alignment estimation).



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 (unpublished)
- PASTA-boosting for alignment methods (unpublished)
- TIPP-boosting for metagenomic taxon identification (unpublished)
- Bin-and-conquer for coalescent-based species tree estimation (2013)

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 (unpublished)
- PASTA-boosting for alignment methods (unpublished)
- TIPP-boosting for metagenomic taxon identification (unpublished)
- Bin-and-conquer for coalescent-based species tree estimation (2013)

This Talk

Fast Converging Methods – estimating trees from polynomial length sequences (several papers, 1997-2001)

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

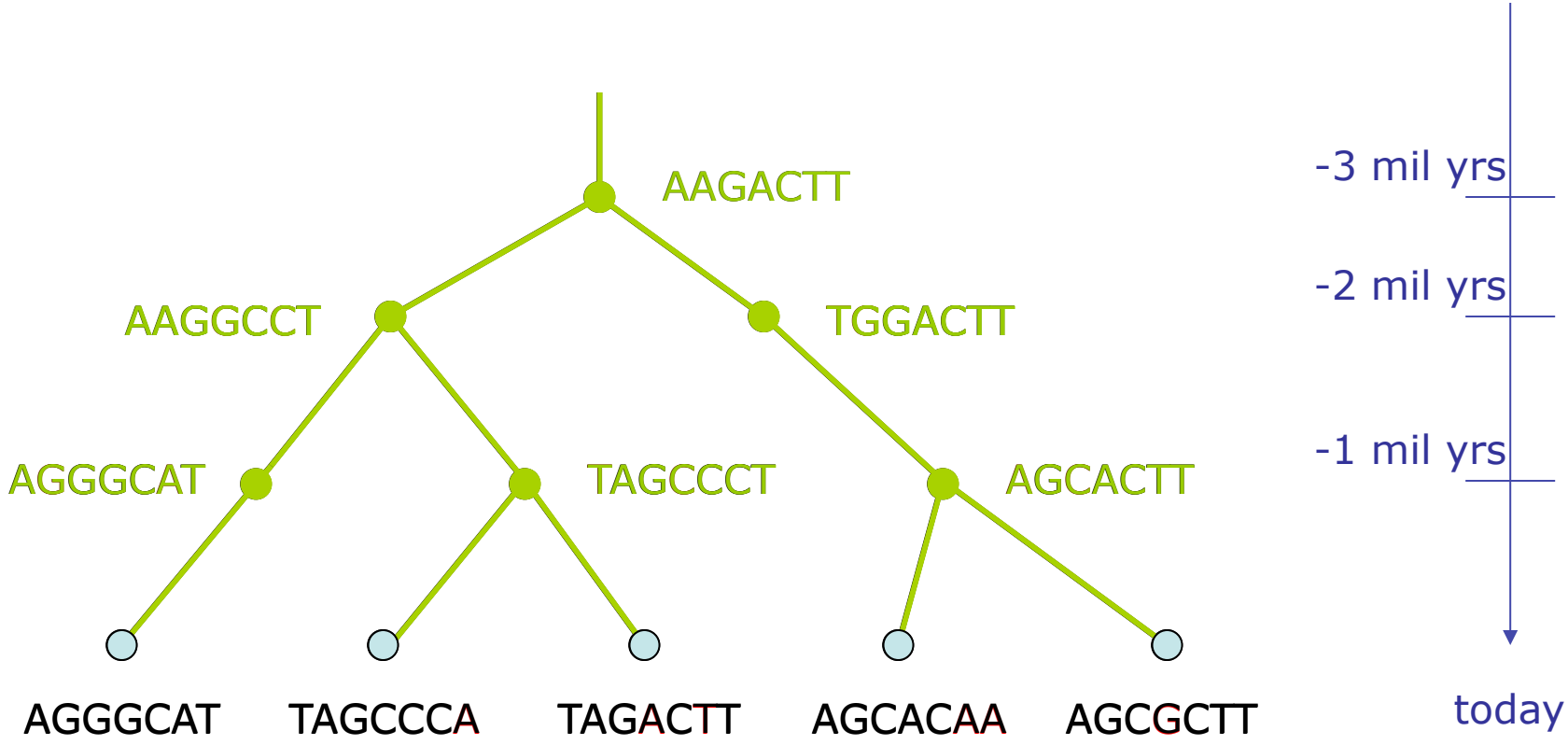
UPP - Ultra-large alignment estimation (unpublished)

Part 1: Absolute Fast Convergence

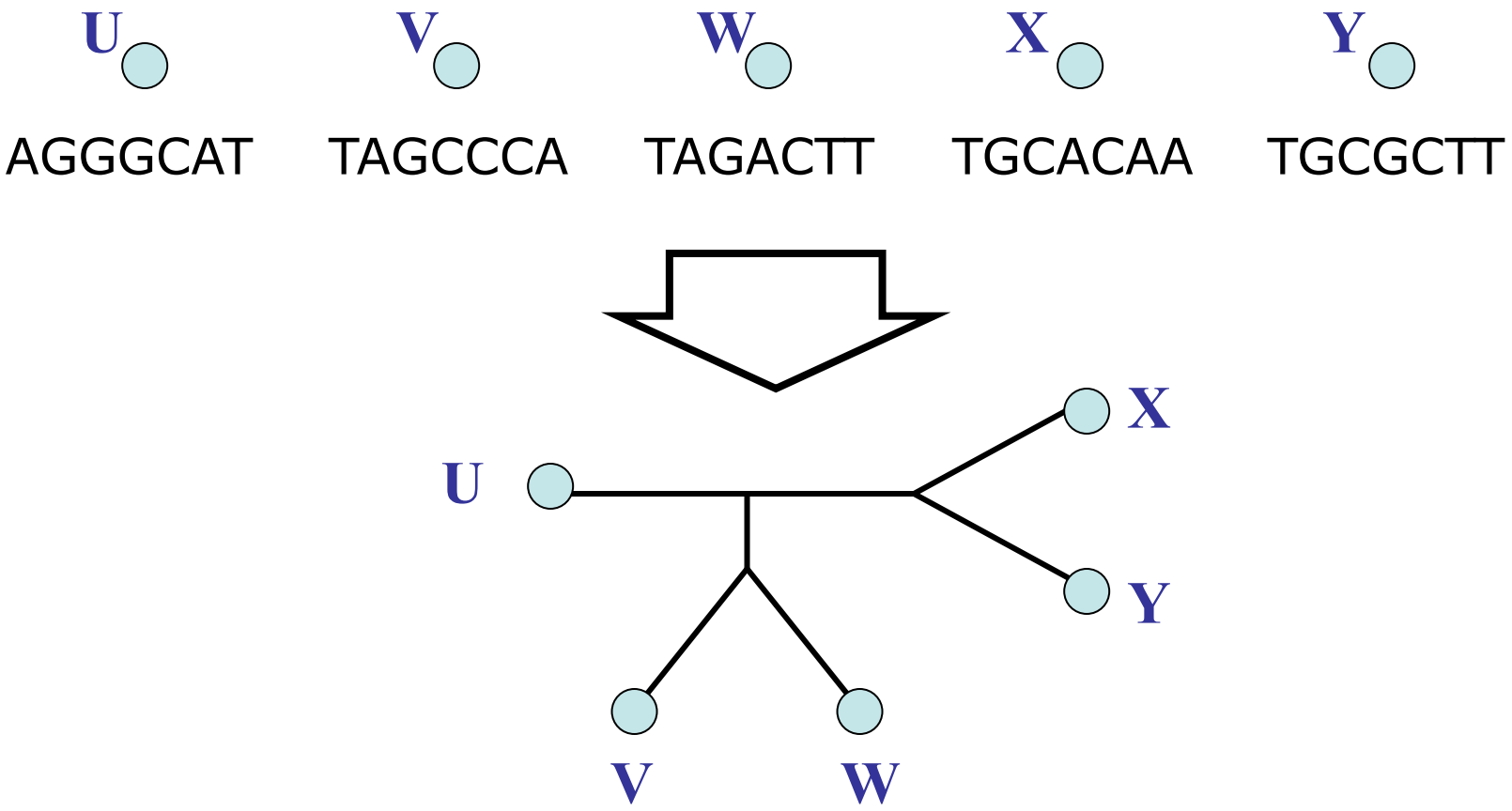
Performance criteria

- Running time.
- Space.
- Statistical performance issues (e.g., statistical consistency) with respect to a Markov model of evolution.
- “Topological accuracy” with respect to the underlying ***true tree or true alignment.*** Typically studied in simulation.
- Accuracy with respect to a particular criterion (e.g. maximum likelihood score), on real data.

DNA Sequence Evolution



Phylogeny Problem



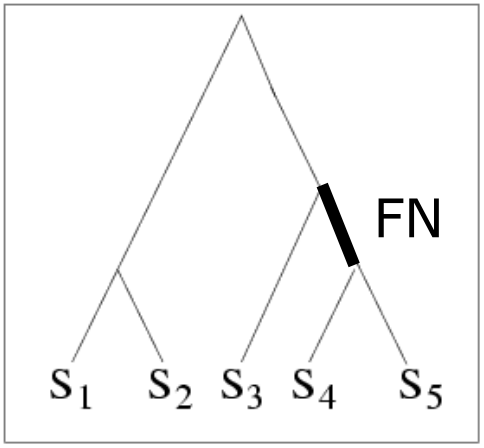
Markov models of site evolution

Simplest (Jukes-Cantor):

- The model tree is a pair (T,{e,p(e)}), where T is a rooted binary tree, and p(e) is the probability of a substitution on the edge e
- The state at the root is random
- If a site changes on an edge, it changes with equal probability to each of the remaining states
- The evolutionary process is Markovian

More complex models (such as the General Markov model) are also considered, with little change to the theory.

Quantifying Error



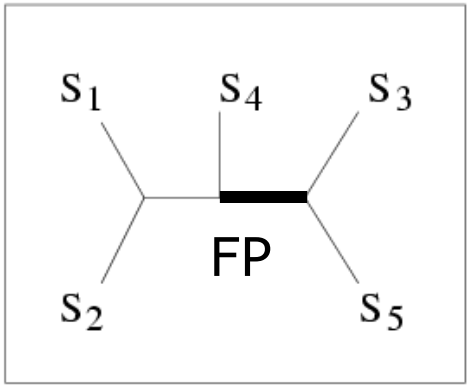
TRUE TREE

S_1	ACAATTAGAAC
S_2	ACCCTTAGAAC
S_3	ACCATTCCAAC
S_4	ACCAGACCAAC
S_5	ACCAGACCGGA

DNA SEQUENCES

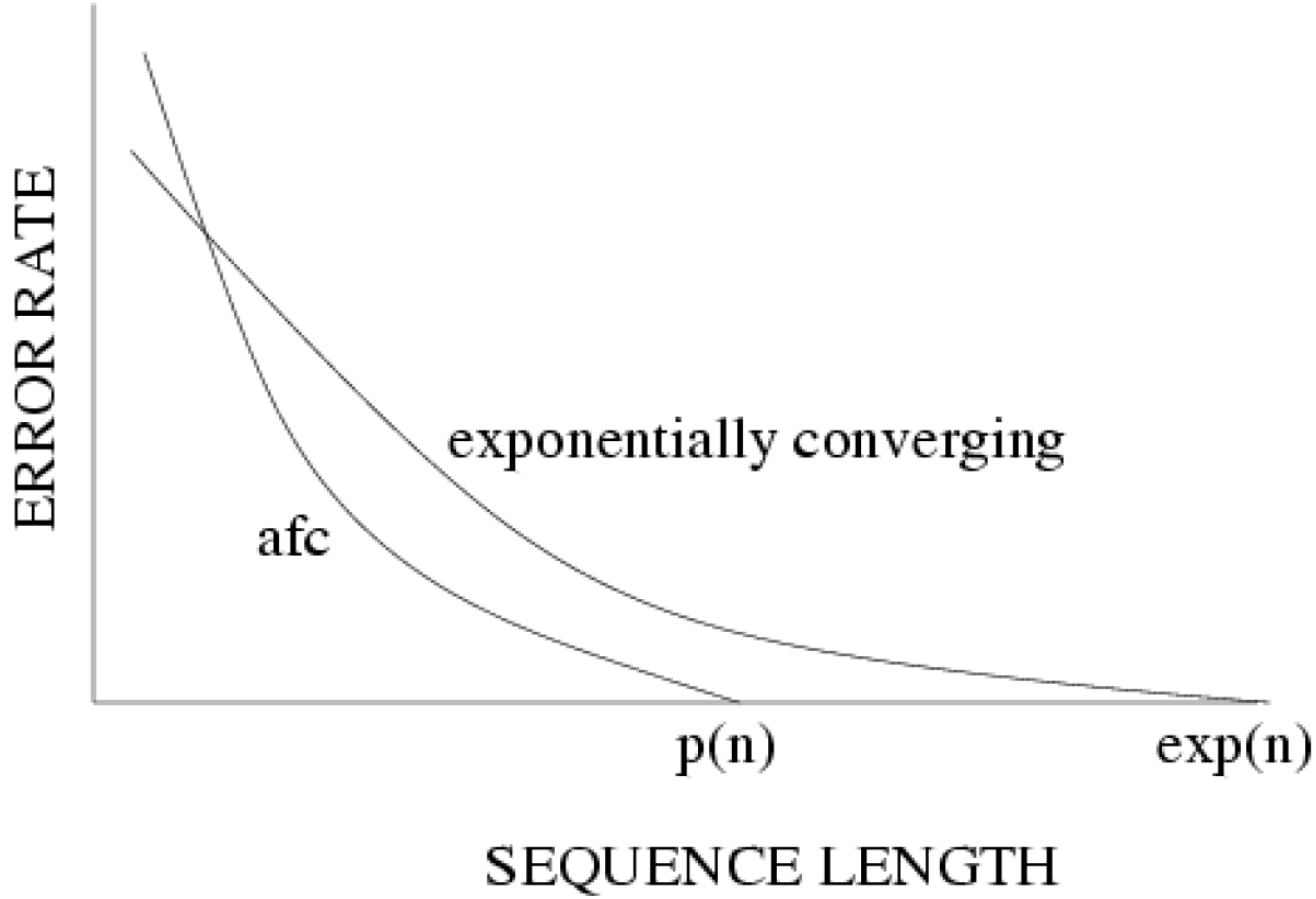
FN: false negative
(missing edge)
FP: false positive
(incorrect edge)

50% error rate

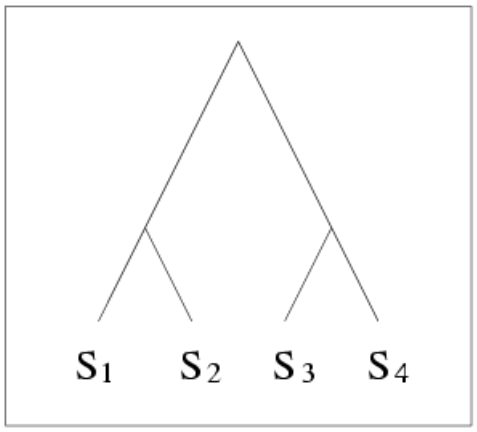


INFERRED TREE

Absolute fast convergence vs. exponential convergence



Distance-based estimation



TRUE TREE

S_1	ACAATTAGAAC
S_2	ACCCTTAGAAC
S_3	ACCATTCCAAC
S_4	ACCAGACCAAC

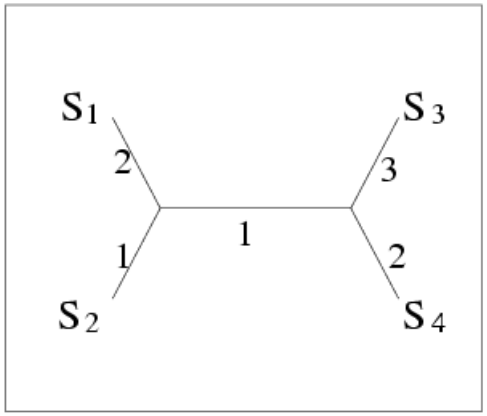
DNA SEQUENCES

STATISTICAL ESTIMATION OF PAIRWISE DISTANCES

	S_1	S_2	S_3	S_4
S_1	0	3	6	5
S_2		0	5	4
S_3			0	5
S_4				0

DISTANCE MATRIX

METHODS SUCH AS NEIGHBOR JOINING

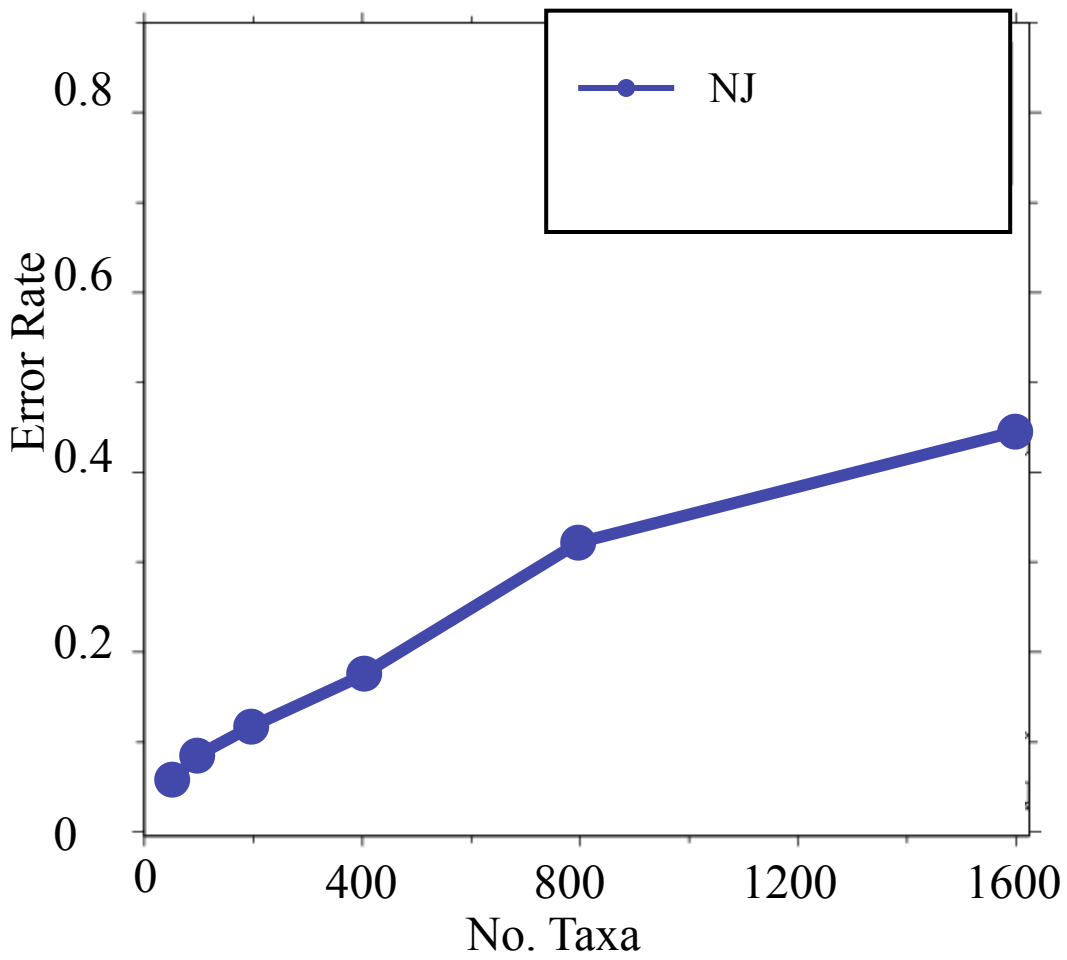


INFERRED TREE

Theorem (Erdos et al., Atteson):

Neighbor joining (and some other methods) will return the true tree with high probability, provided sequence lengths are exponential in the evolutionary diameter of the tree.

Performance on large diameter trees

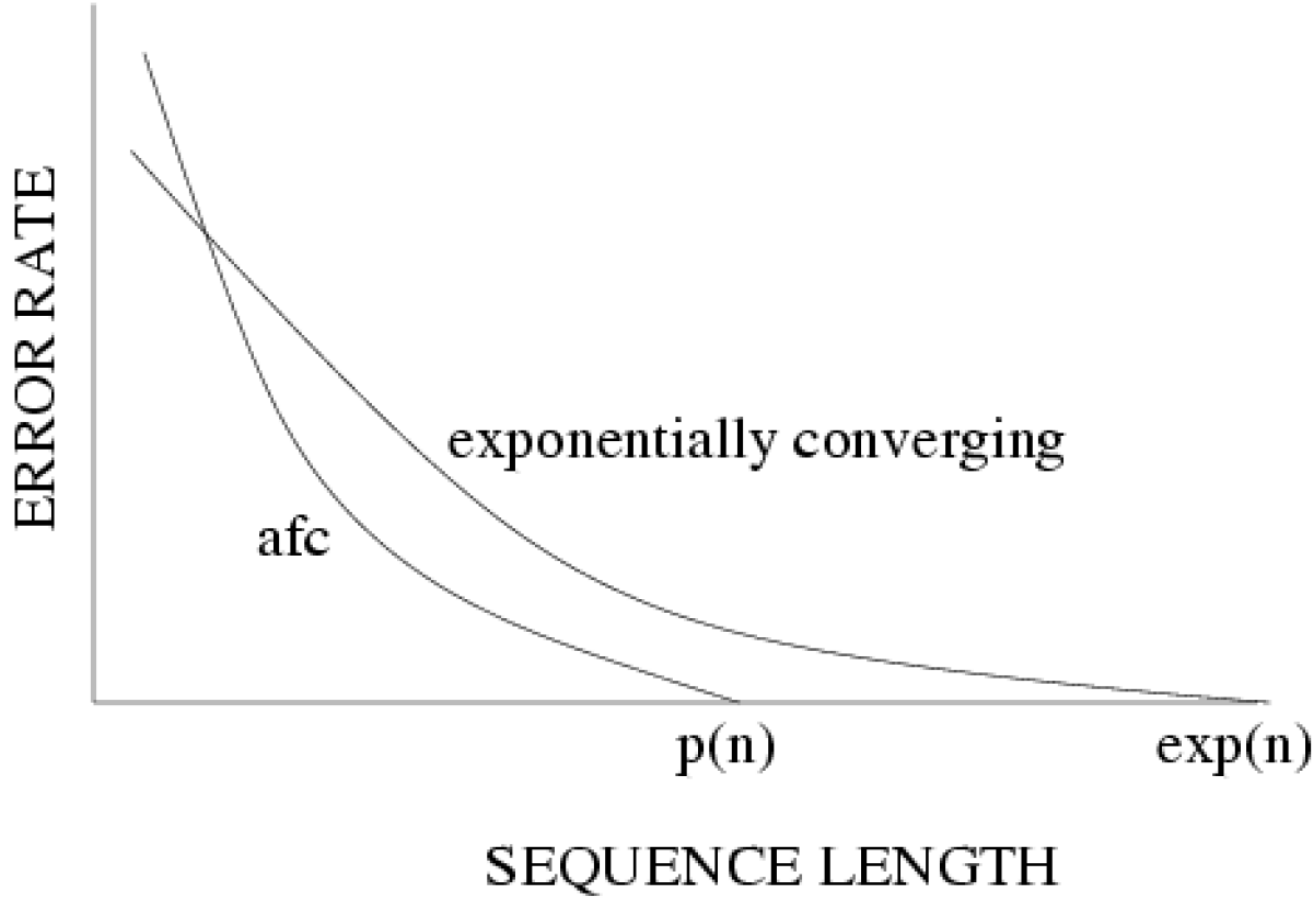


Simulation study based upon fixed edge lengths, K2P model of evolution, sequence lengths fixed to 1000 nucleotides.

Error rates reflect proportion of incorrect edges in inferred trees.

[Nakhleh et al. ISMB 2001]

Absolute fast convergence vs. exponential convergence



Absolute fast-converging methods

1997: Erdos, Steel, Szekely, and Warnow (ICALP).

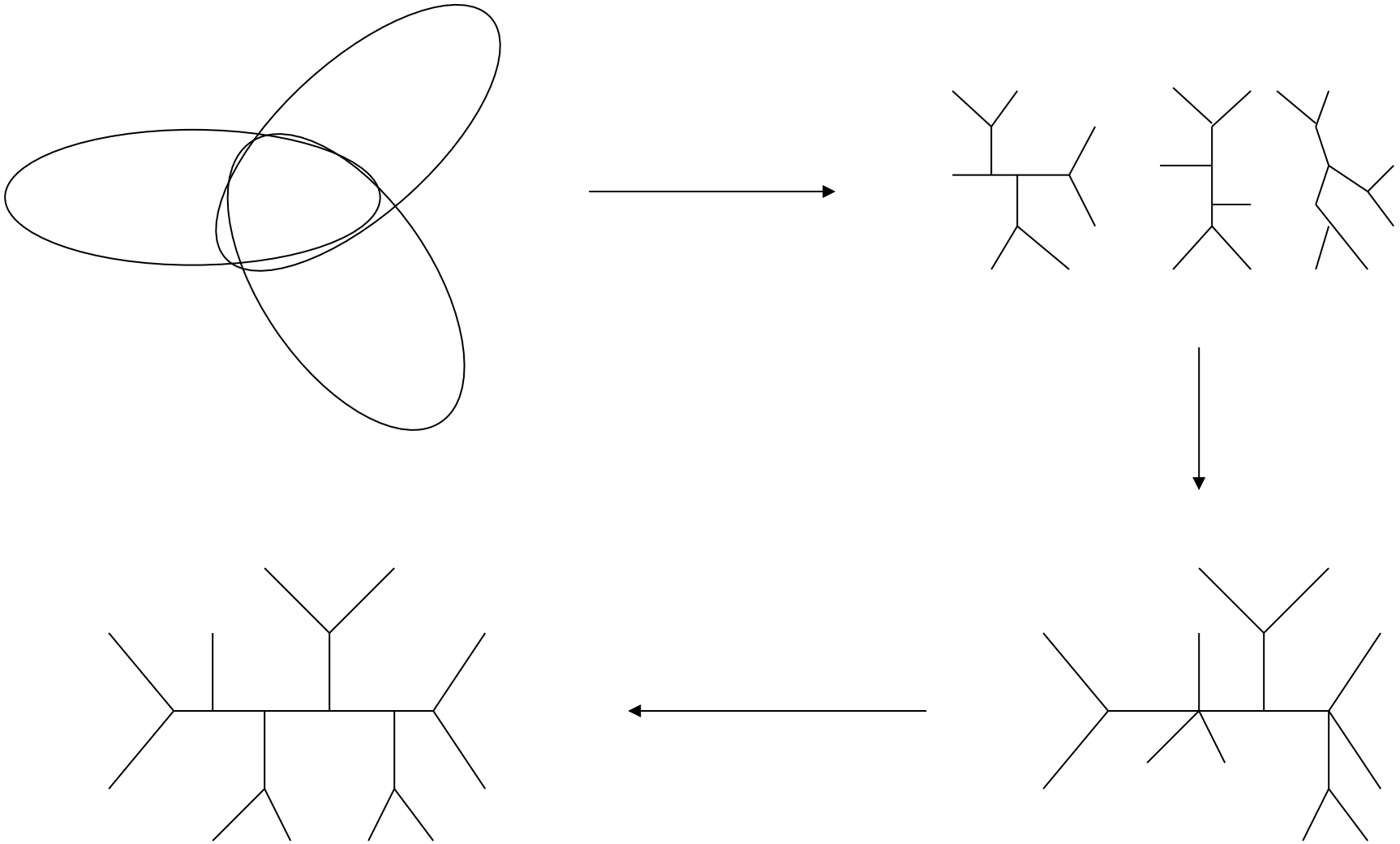
1999: Erdos, Steel, Szekely, and Warnow (RSA and TCS);
Huson, Nettles and Warnow (J. Computational Biology)

2001: Warnow, St. John, and Moret (SODA);
Nakhleh, St. John, Roshan, Sun, and Warnow (ISMB)

Using divide-and-conquer

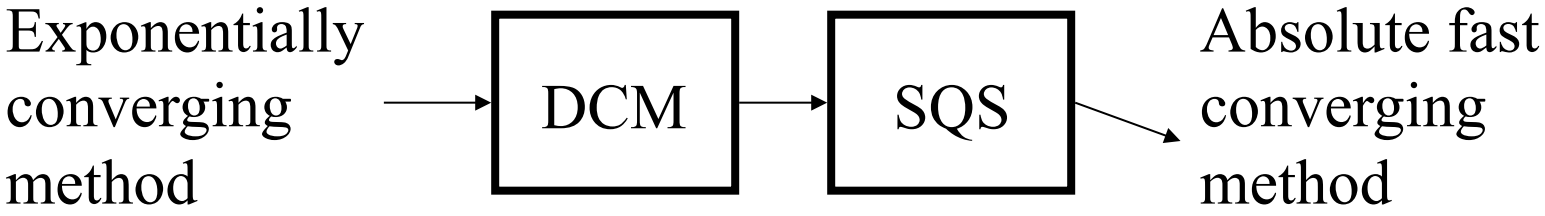
- Idea: better (more accurate) trees will be found if we compute trees on subsets with smaller diameters, and then combine trees on these subsets

DCMs: Divide-and-conquer for improving phylogeny reconstruction



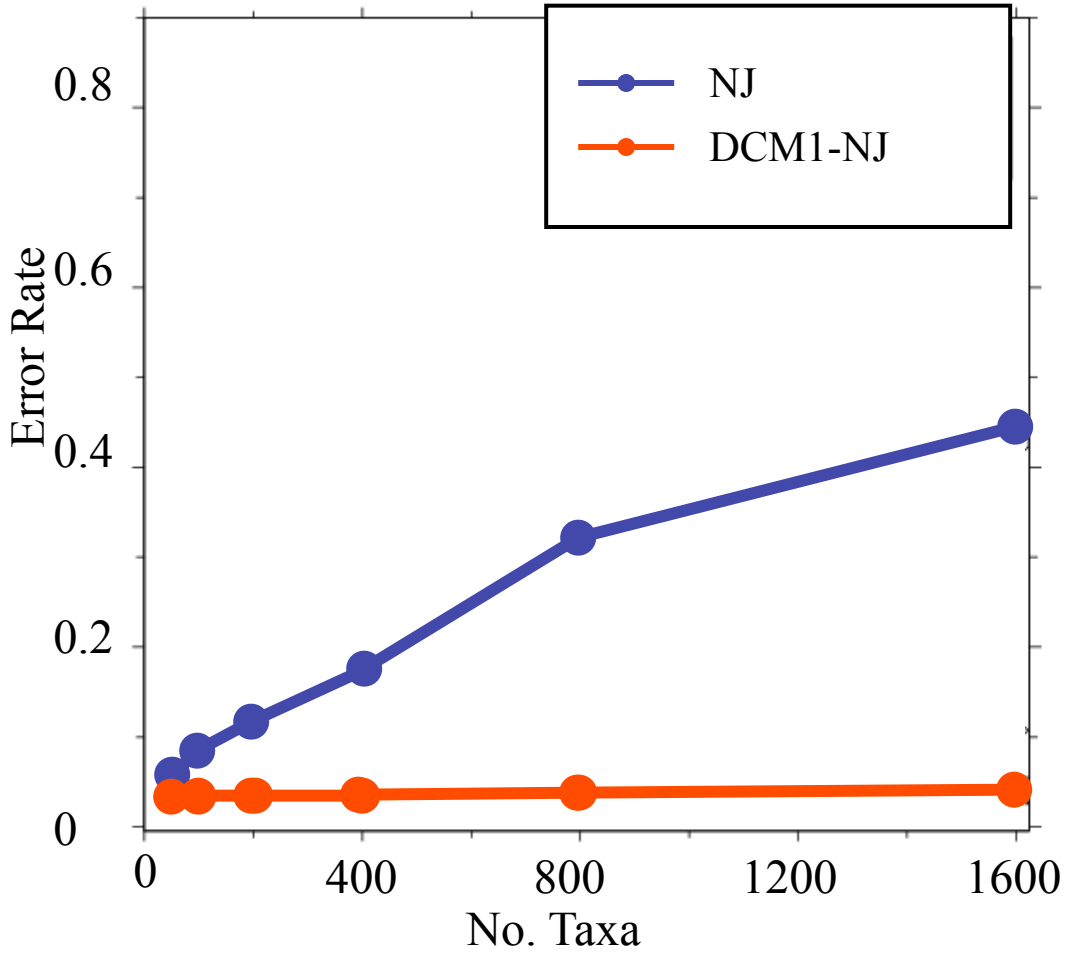
DCM-Boosting *[Warnow et al. 2001]*

- DCM+SQS is a two-phase procedure which reduces the sequence length requirement of methods.



DCM1-boosting distance-based methods

[Nakhleh et al. ISMB 2001]



Theorem (Warnow et al., SODA 2001):

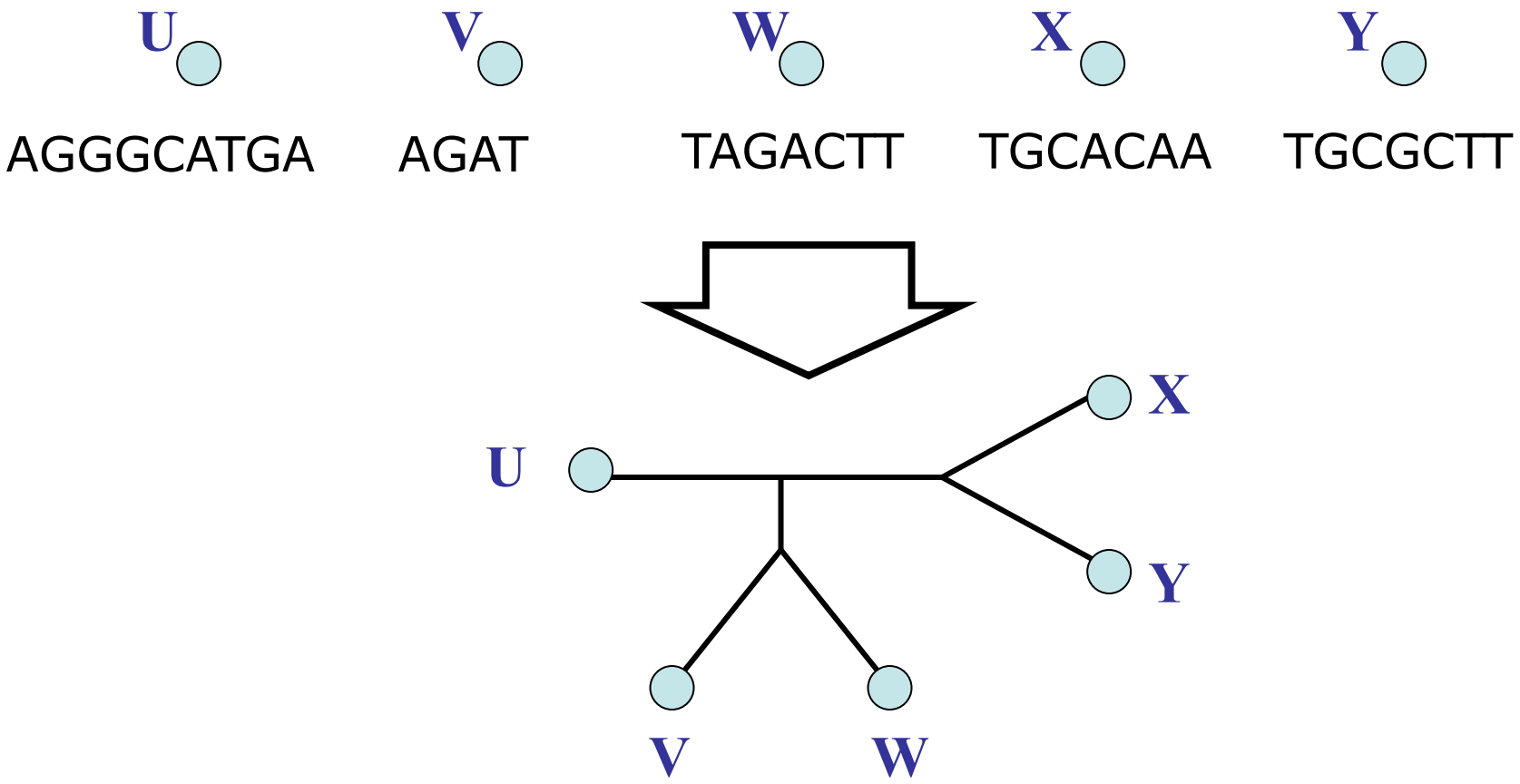
DCM1-NJ converges to the true tree from polynomial length sequences

Fast-converging methods (and related work)

- 1997: Erdos, Steel, Szekely, and Warnow (ICALP).
- 1999: Erdos, Steel, Szekely, and Warnow (RSA, TCS);
  Huson, Nettles and Warnow (J. Comp Bio.)
- 2001: Warnow, St. John, and Moret (SODA);
  Nakhleh, St. John, Roshan, Sun, and Warnow (ISMB)
  Cryan, Goldberg, and Goldberg (SICOMP);
  Csuros and Kao (SODA);
- 2002: Csuros (J. Comp. Bio.)
- 2006: Daskalakis, Mossel, Roch (STOC),
  Daskalakis, Hill, Jaffe, Mihaescu, Mossel, and Rao (RECOMB)
- 2007: Mossel (IEEE TCBB)
- 2008: Gronau, Moran and Snir (SODA)
- 2010: Roch (Science)
- 2013: Roch (in preparation)

and others

What about indels?



Part II: SATé

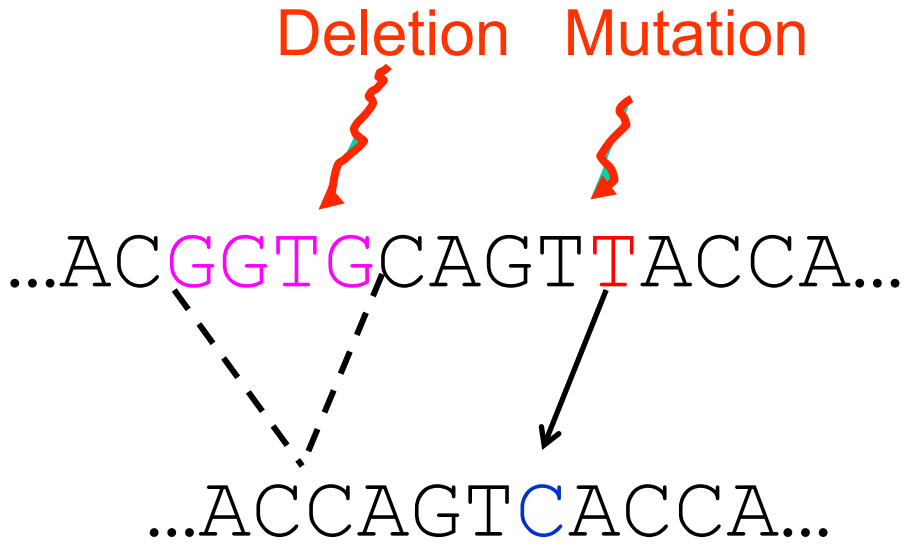
Simultaneous Alignment and Tree Estimation

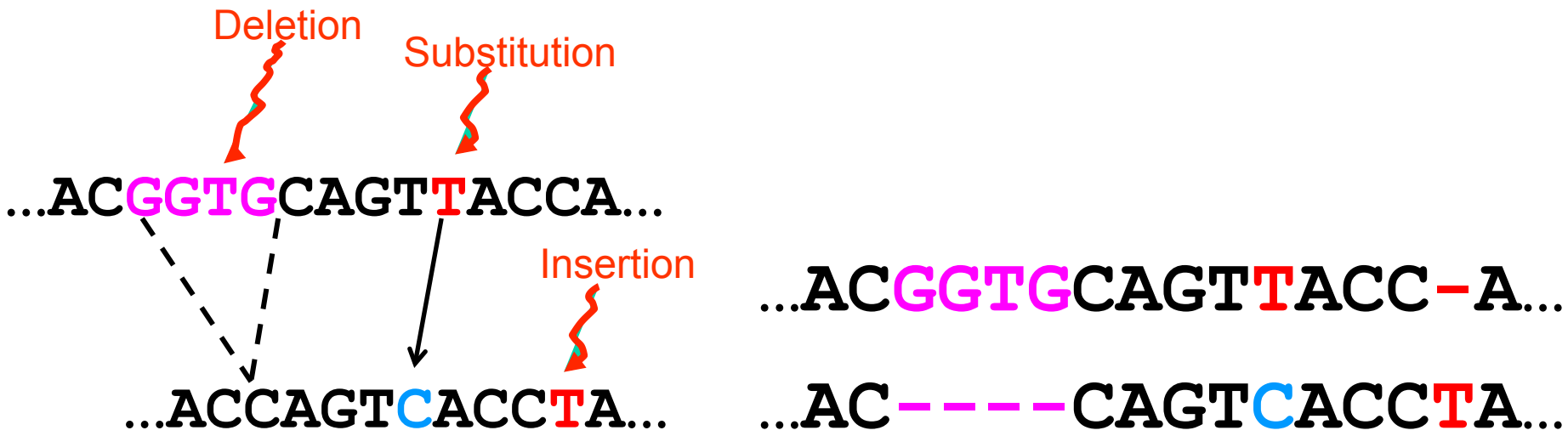
Liu, Nelesen, Raghavan, Linder, and Warnow, *Science*, 19 June 2009, pp. 1561-1564.

Liu et al., Systematic Biology 2012

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

Indels (insertions and deletions) also occur!





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
```



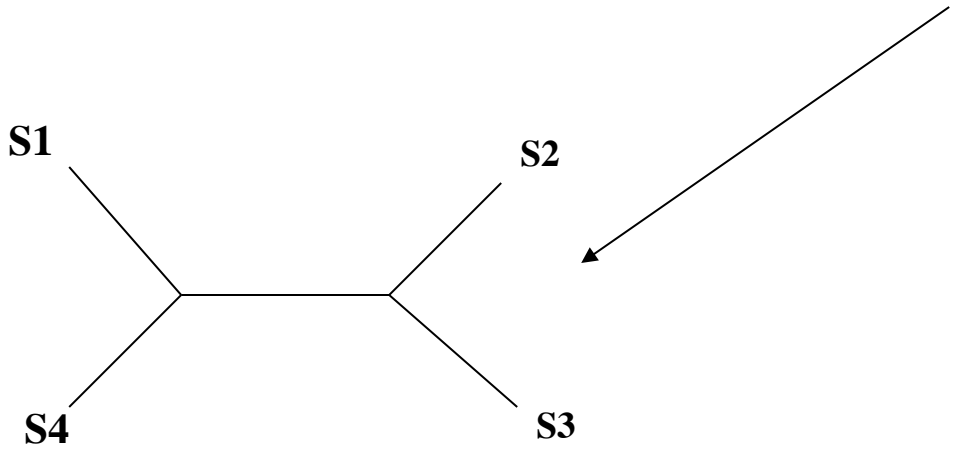
```
S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAC--GACCGC--
S3 = TAG-CT-------GACCGC--
S4 = -------TCAC--GACCGACA
```

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

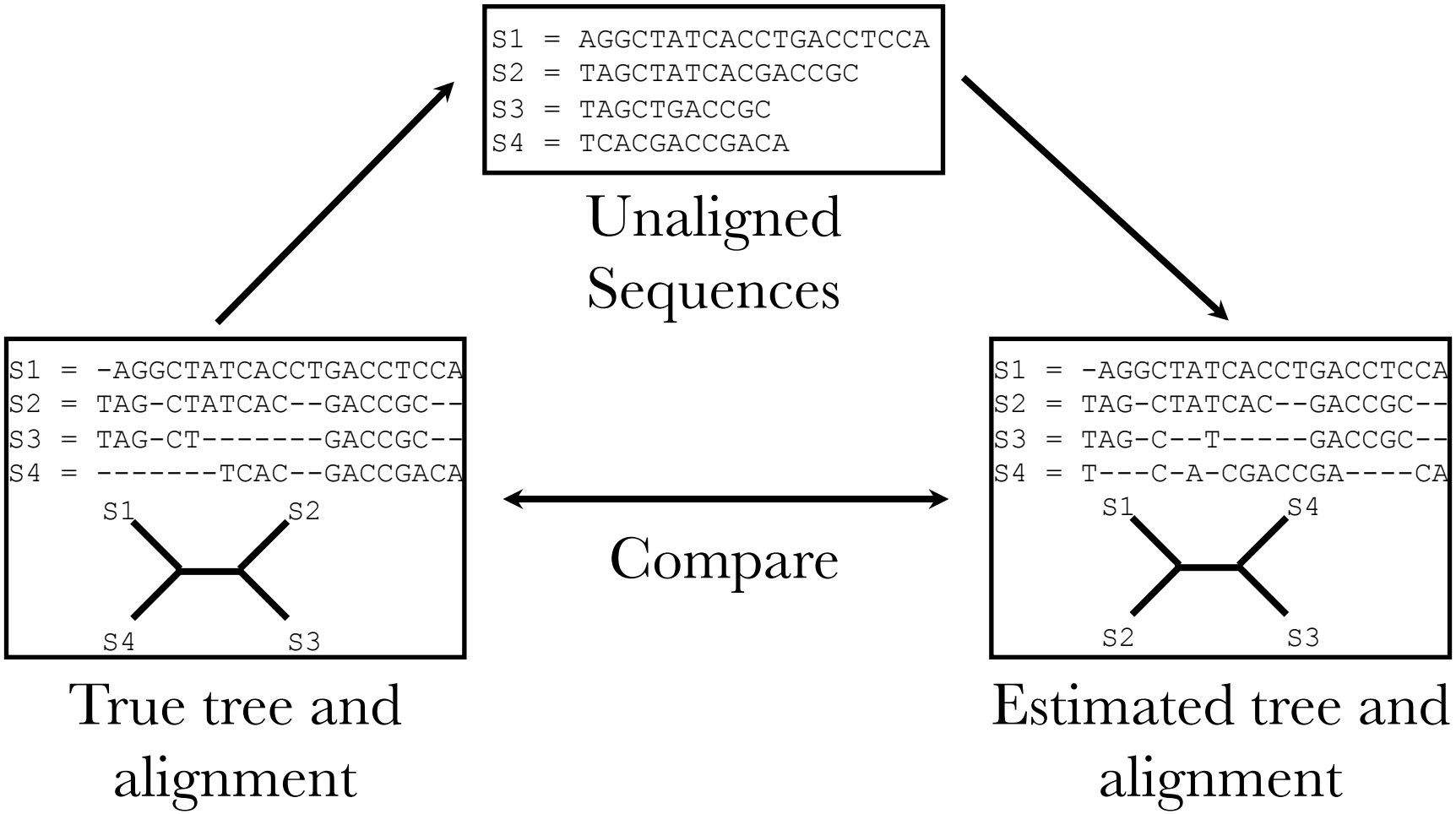
- Clustal
- POY (and POY*)
- Probcons (and Probtree)
- Probalign
- MAFFT
- Muscle
- Di-align
- T-Coffee
- Prank (PNAS 2005, Science 2008)
- Opal (ISMB and Bioinf. 2007)
- *FSA (PLoS Comp. Bio. 2009)*
- *Infernal (Bioinf. 2009)*
- Etc.

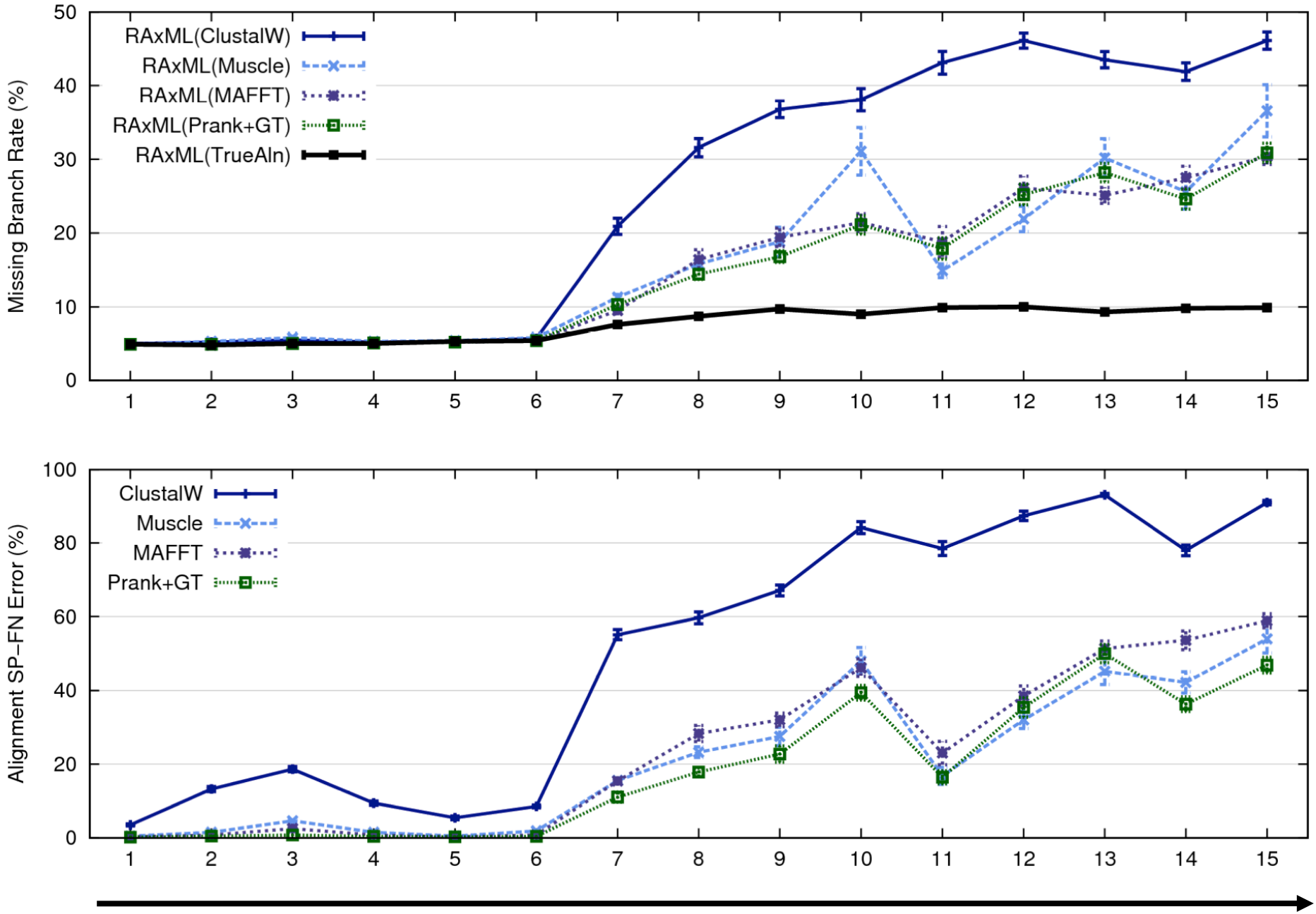
Phylogeny methods

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

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

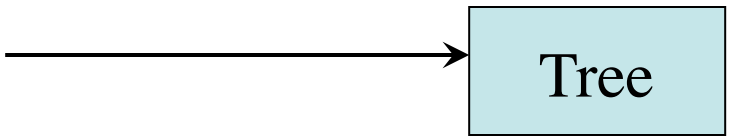
Simulation Studies




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)

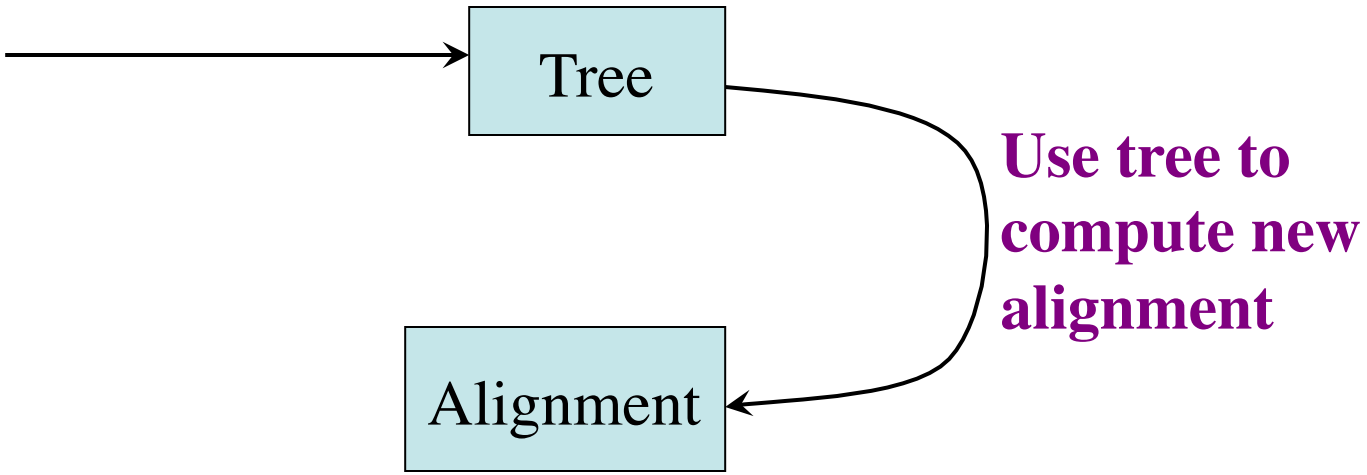
SATé Algorithm

Obtain initial alignment
and estimated ML tree



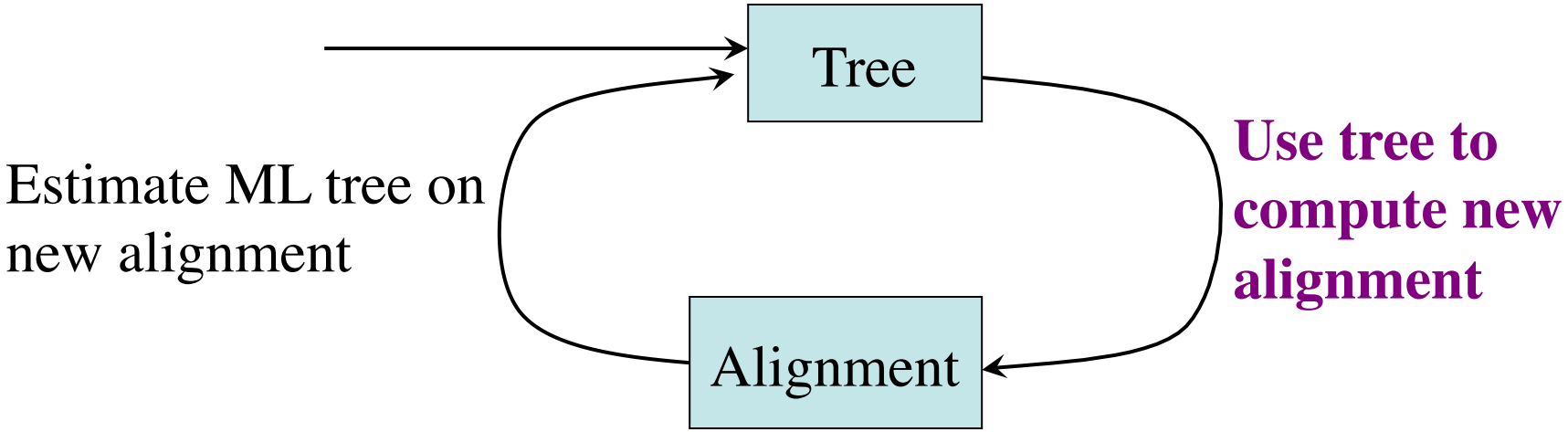
SATé Algorithm

Obtain initial alignment
and estimated ML tree



SATé Algorithm

Obtain initial alignment
and estimated ML tree

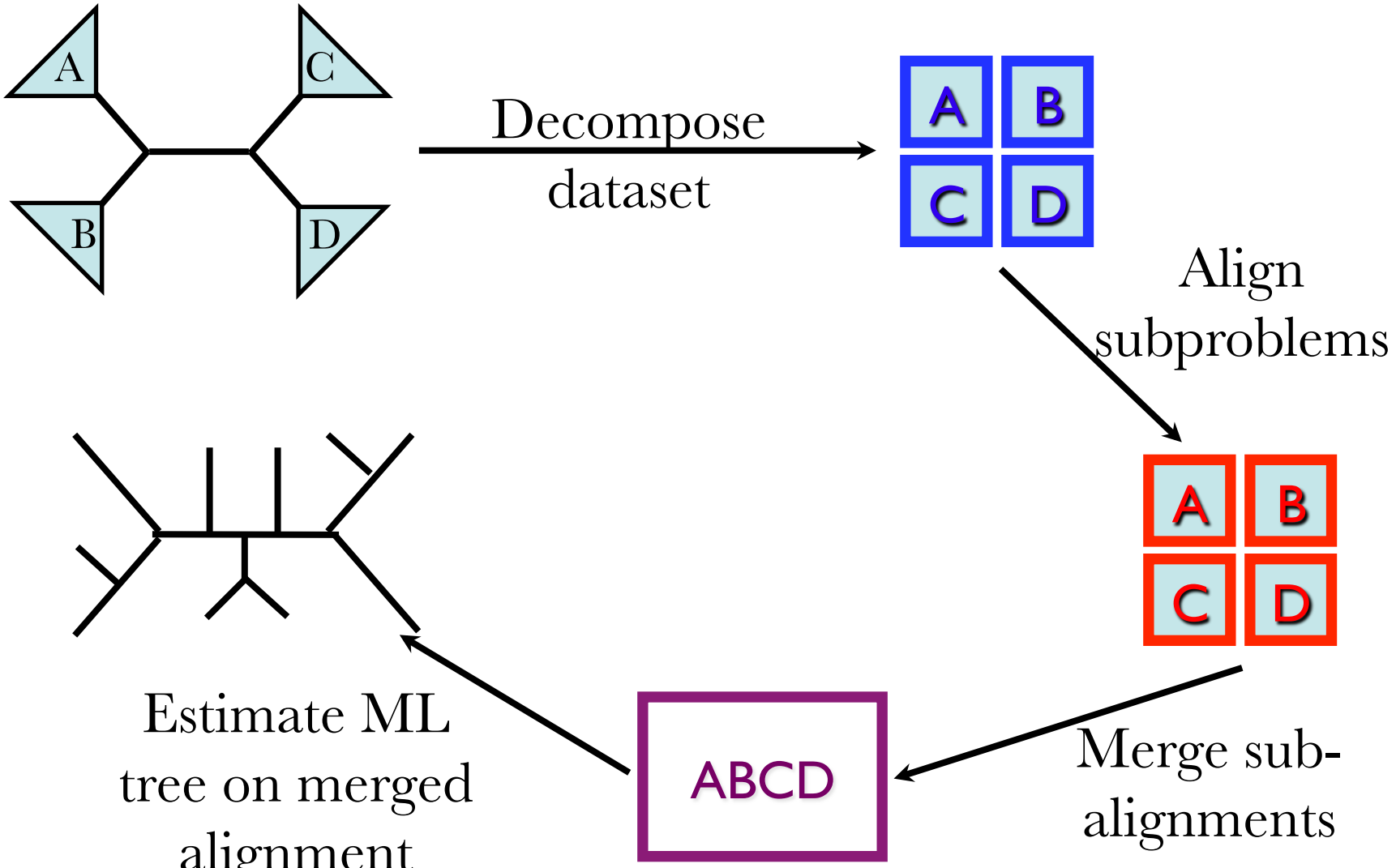


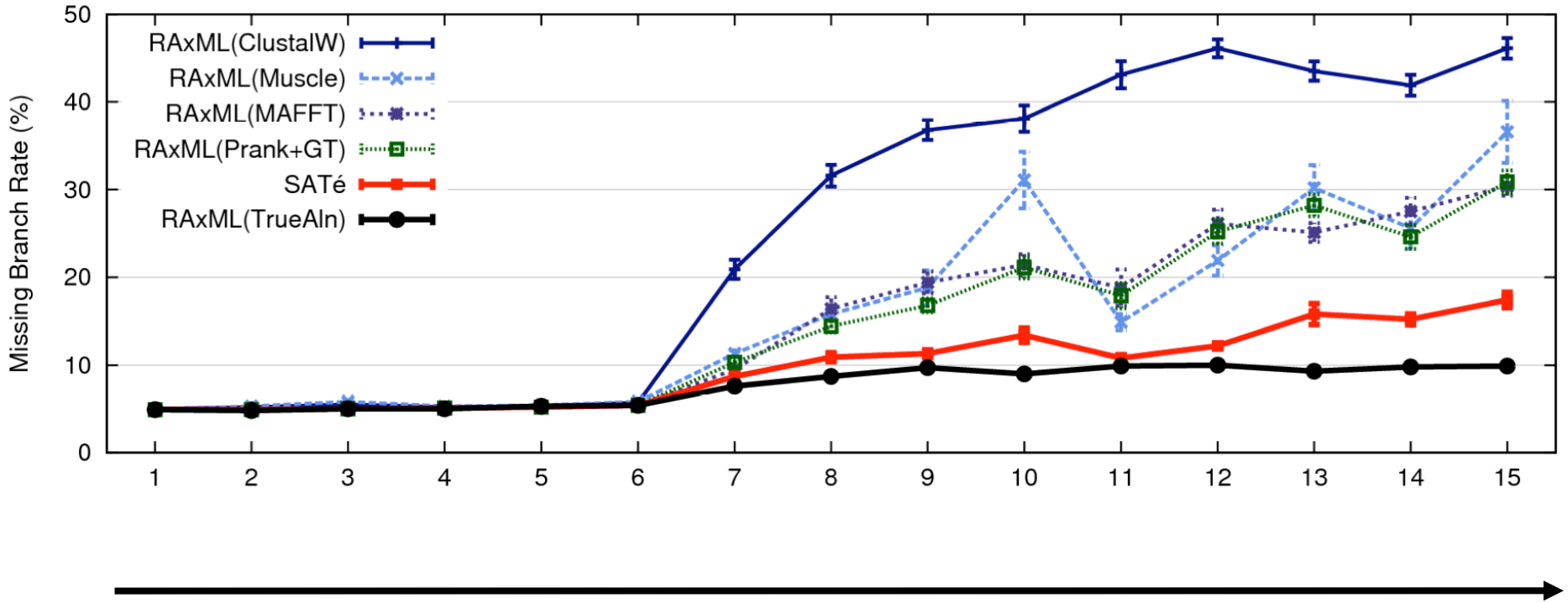
Re-alignment on the tree

- 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

Re-aligning on a tree

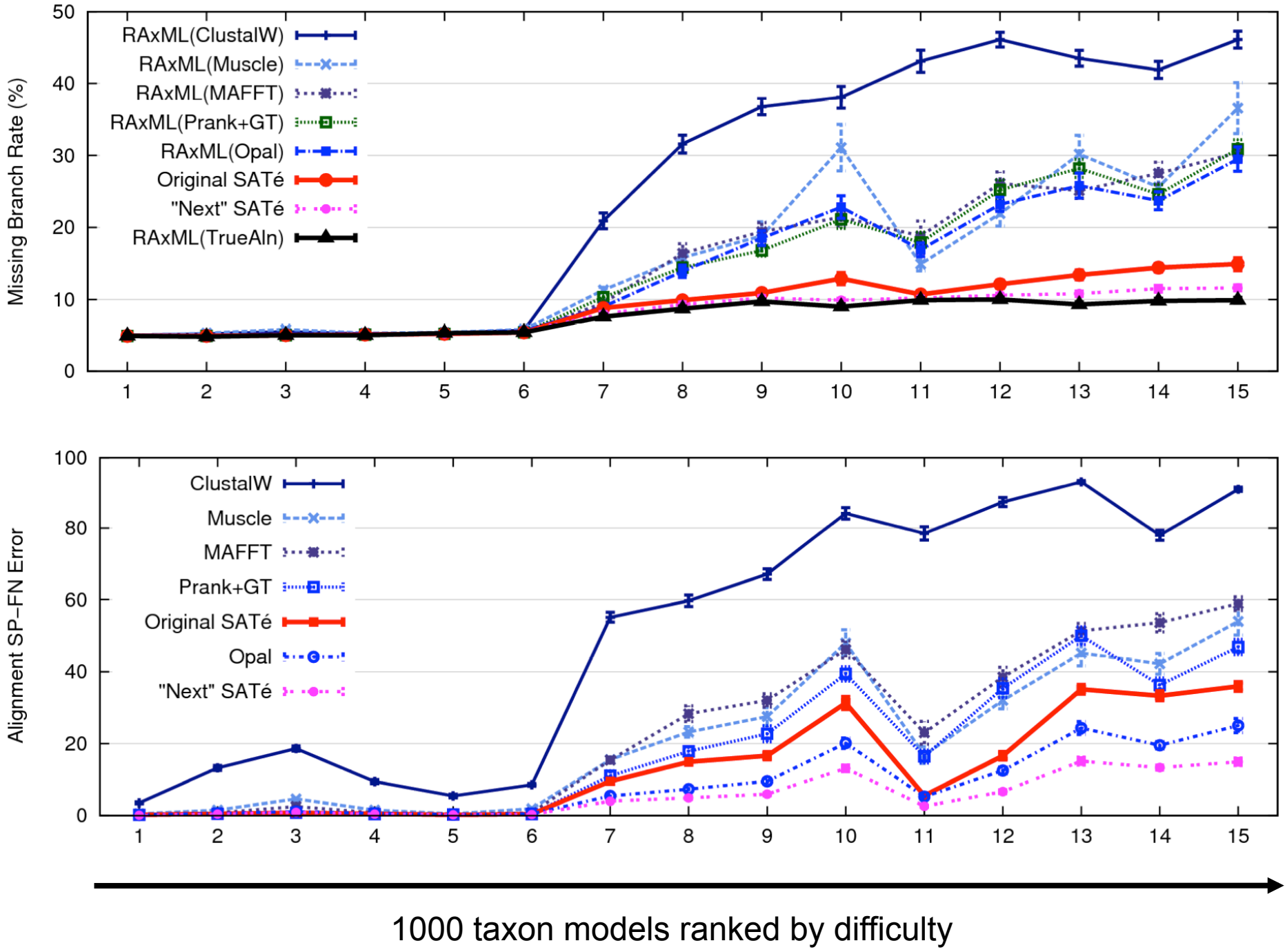




1000 taxon models, ordered by difficulty

24 hour SATé analysis, on desktop machines

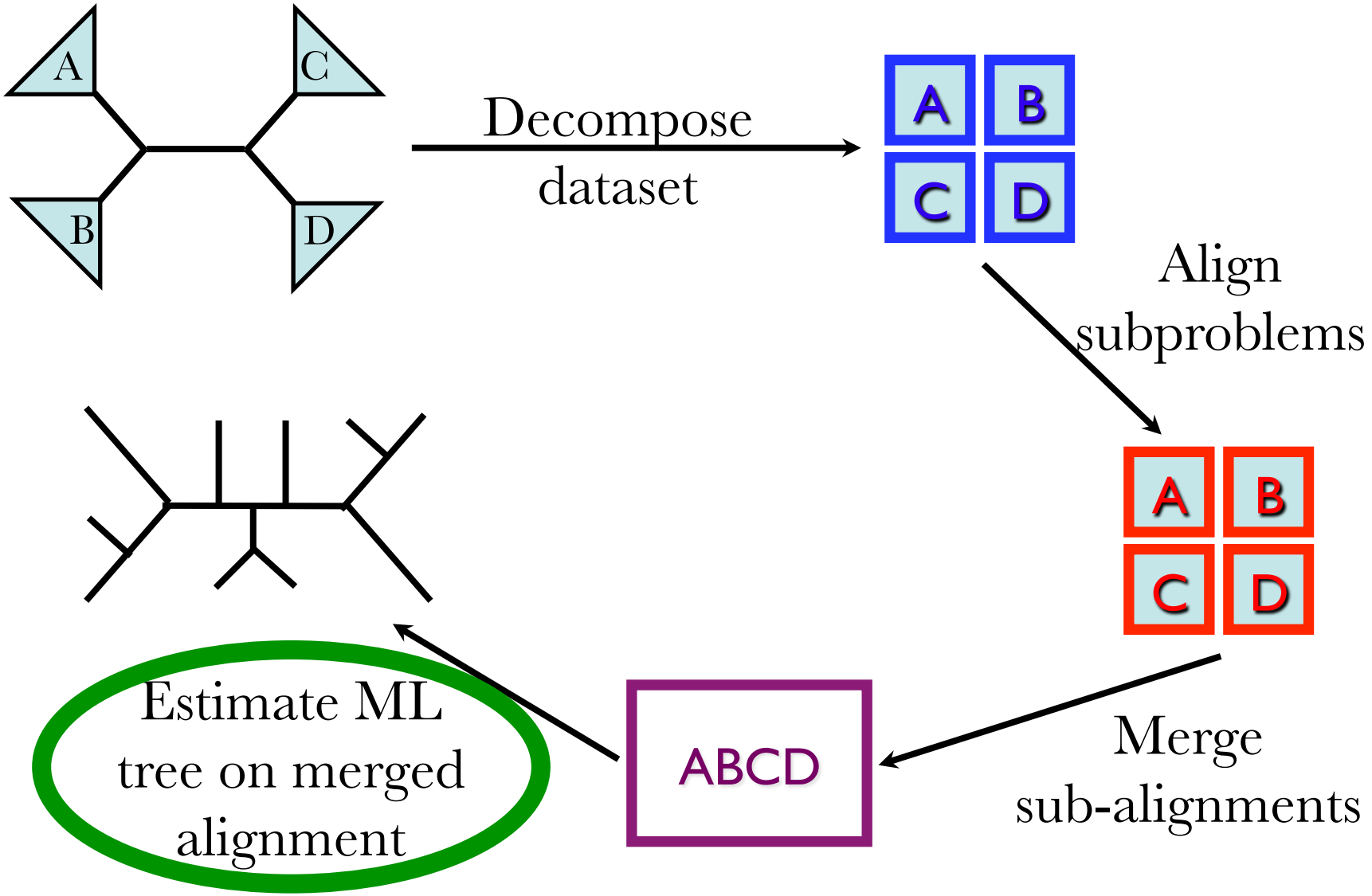
(Similar improvements for biological datasets)


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

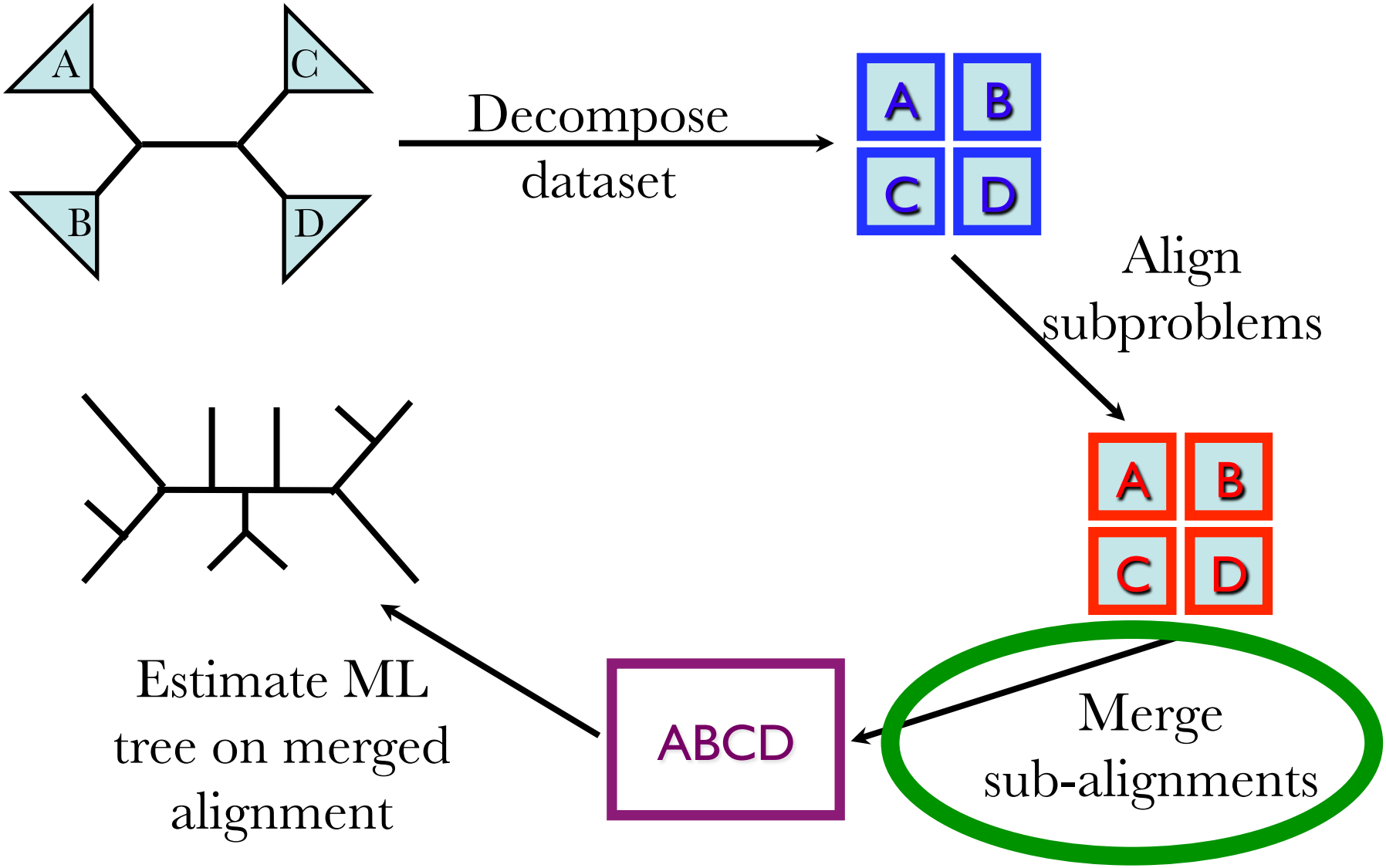
Brief discussion

- **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)
- **Performance in practice** results from use of base methods (and ability to use best versions of base methods).

Limitations



Limitations



UPP: Ultra-large alignment using SEPP

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

Authors: Nam Nguyen, Siavash Mirarab, and Tandy Warnow

In preparation – expected submission Fall 2013

UPP: Ultra-large alignment using SEPP

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: Ultra-large alignment using SEPP

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: Ultra-large alignment using SEPP

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?

Obvious approach: build HMM (Hidden Markov Model) for backbone alignment, and use it to align remaining sequences

HMMER (Sean Eddy, HHMI) leading software for this purpose

Using HMMER

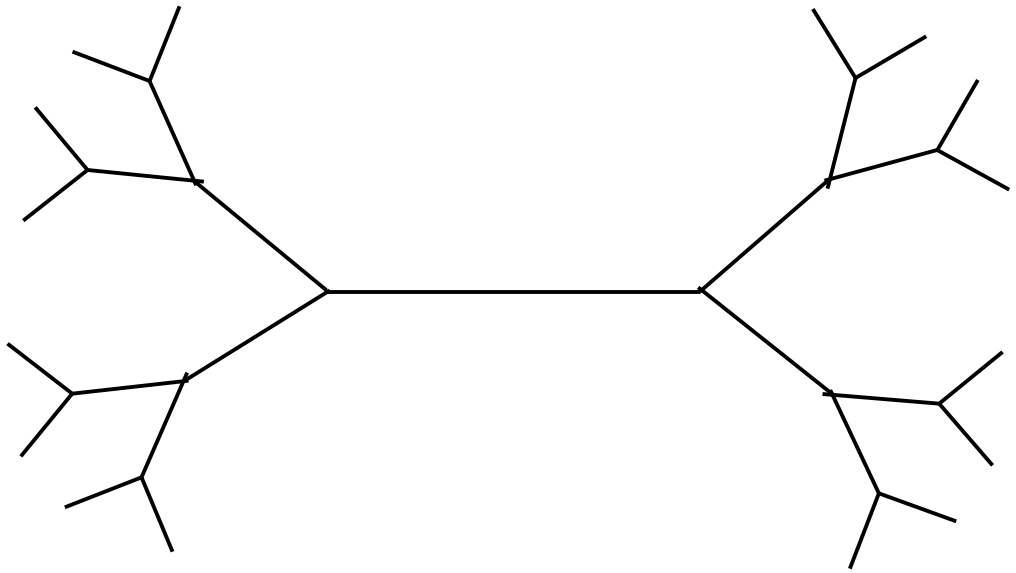
Using HMMER works well…

Using HMMER

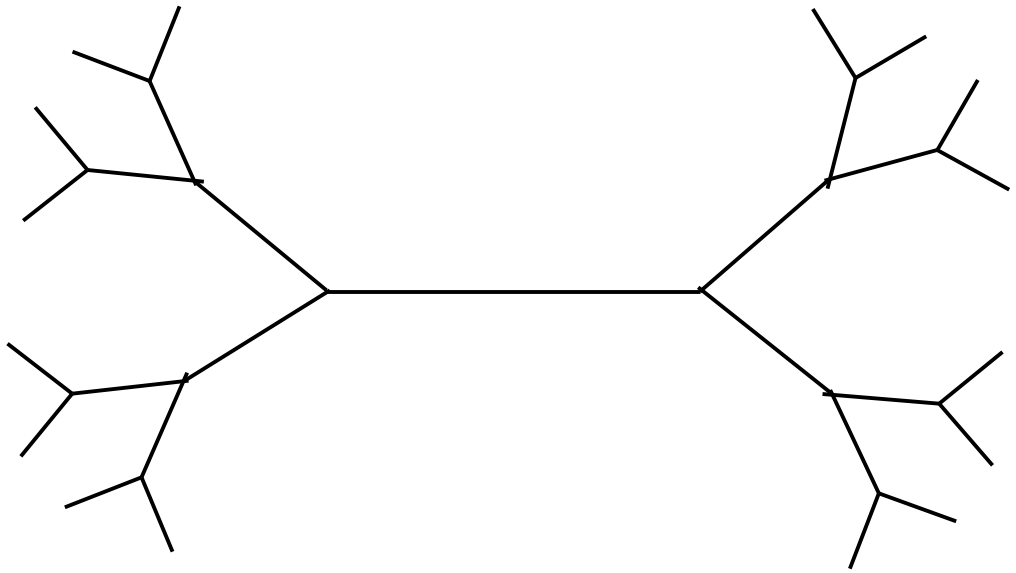
Using HMMER works well…except when the dataset is big!

Using HMMER to add sequences to an existing alignment

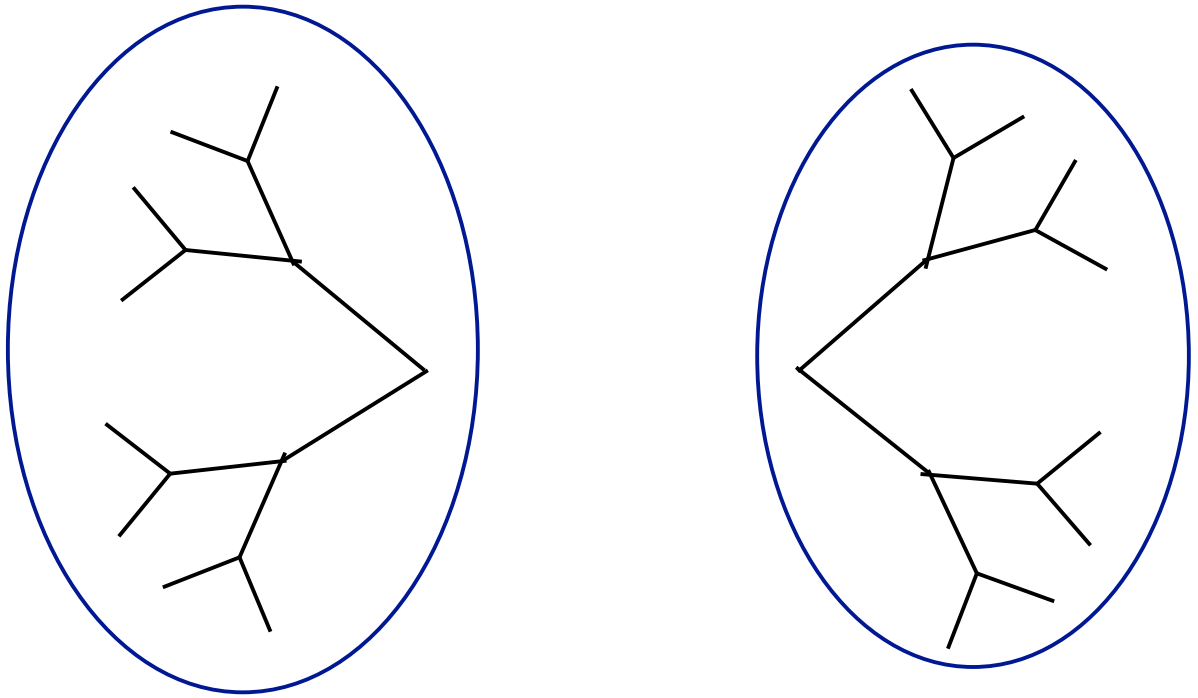
1) build one HMM for the backbone alignment
2) Align sequences to the HMM, and insert into backbone alignment



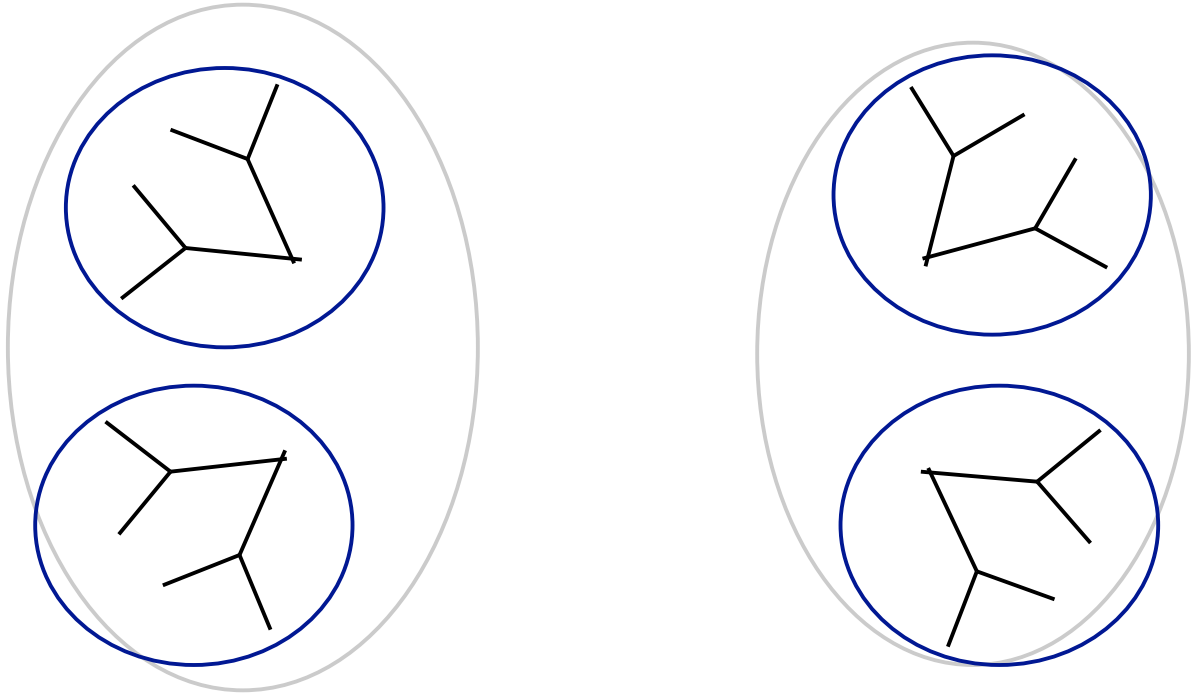
One Hidden Markov Model for the entire alignment?



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.

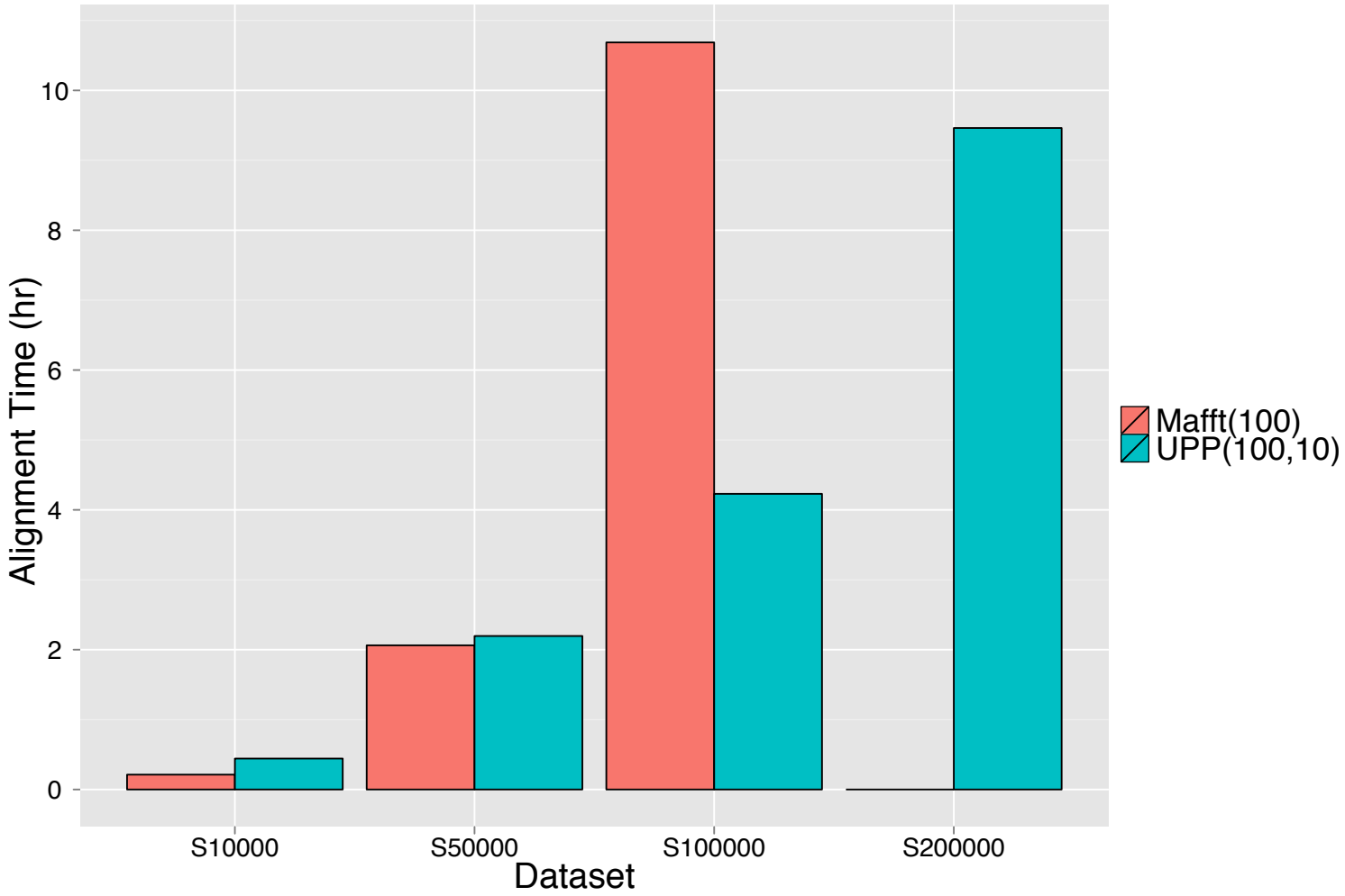
UPP design

- **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 matters;** we use SATé.

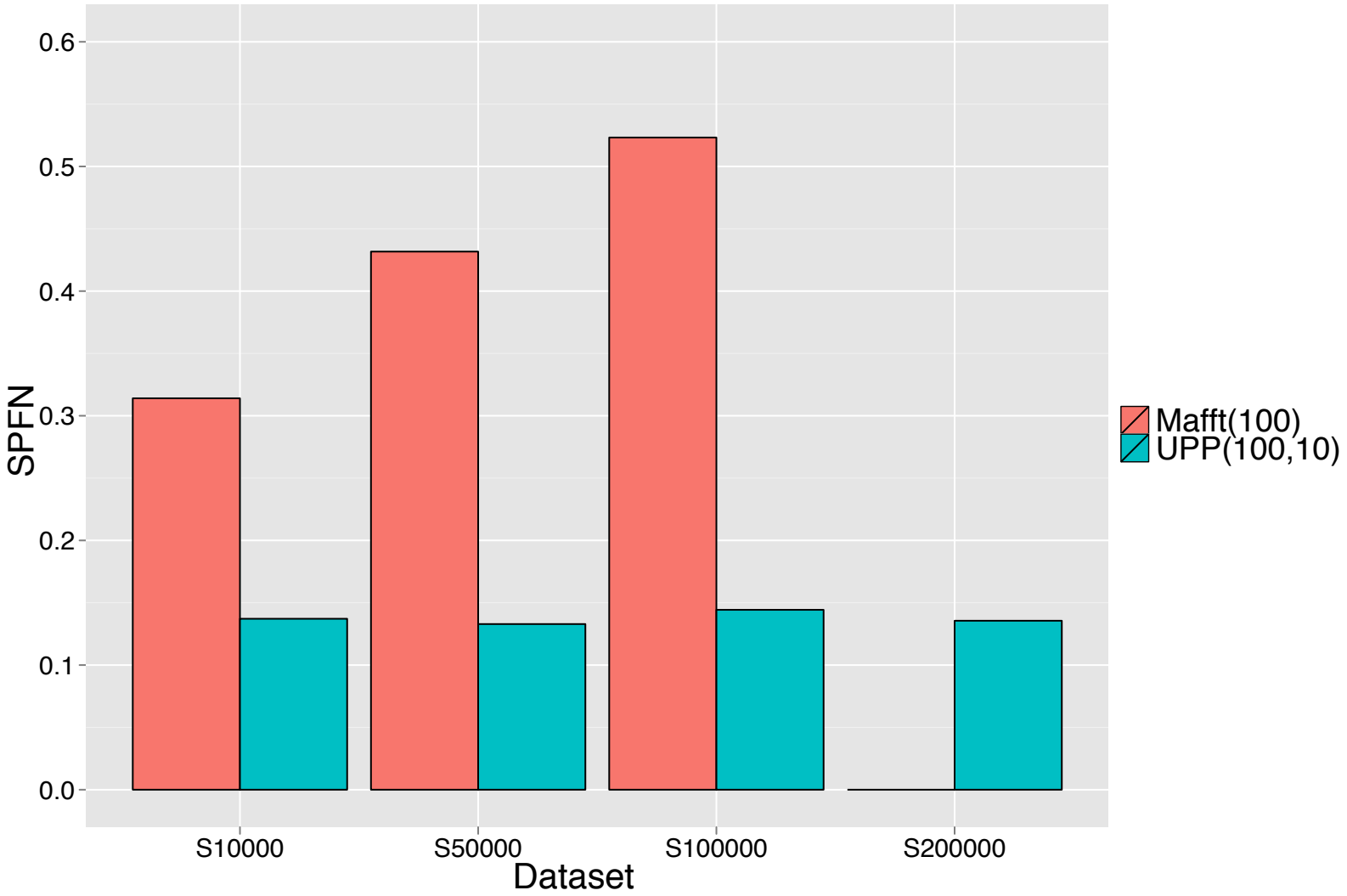
Evaluation of UPP

- **Simulated Datasets**: 1,000 to 1,000,000 sequences (RNASim, Junhyong Kim 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

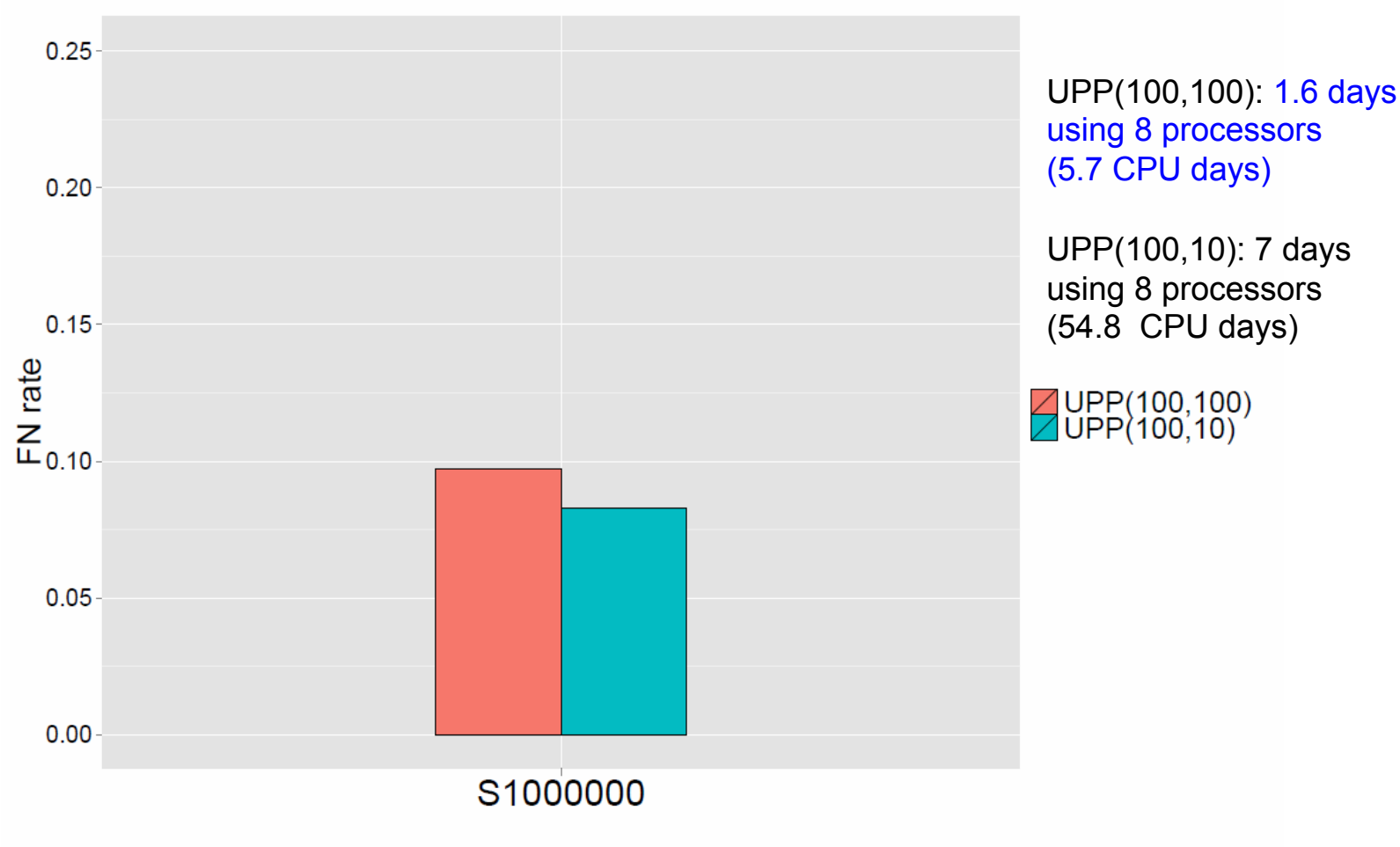
UPP vs. MAFFT Running Time



UPP vs. MAFFT Alignment Error



One Million Sequences: Tree Error



Note improvement obtained by using UPP decomposition

UPP performance

- UPP is very fast, parallelizable, and scalable.
- UPP vs. standard MSA methods: UPP is more accurate on large datasets (with 1000+ taxa), and trees on UPP alignments are more accurate than trees on standard alignments.
- UPP vs. SATé: UPP is much faster and can analyze much larger datasets; UPP has about the same alignment accuracy, but produces slightly less accurate trees.

Other uses of multiple HMMs

- SEPP: Phylogenetic Placement of short reads into existing tree (Nguyen, Mirarab, and Warnow, PSB 2012)

- TIPP: taxon identification of metagenomic sequences (in preparation, Nguyen et al. 2013)

Summary

3 Phylogenetic "Boosters"

- DCM1: reducing sequence length requirements
- SATé: co-estimation of alignments and trees
- UPP: ultra-large multiple sequence alignment

Algorithmic Strategies

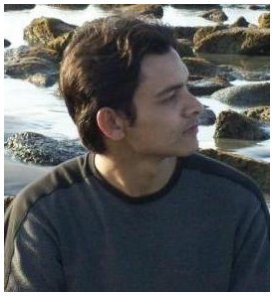
- Divide-and-conquer
- Chordal graph decompositions
- Iteration
- Multiple HMMs
- Bin-and-conquer (technique used for improving species tree estimation from multiple gene trees, Bayzid and Warnow, Bioinformatics 2013)

Other Current Research

- Large-scale alignment (PASTA and UPP)
- Coalescent-based species tree estimation (bin-and-conquer)
- Alignment and phylogeny estimation using NGS (next generation sequencing) data
- Metagenomic analysis

Warnow Laboratory





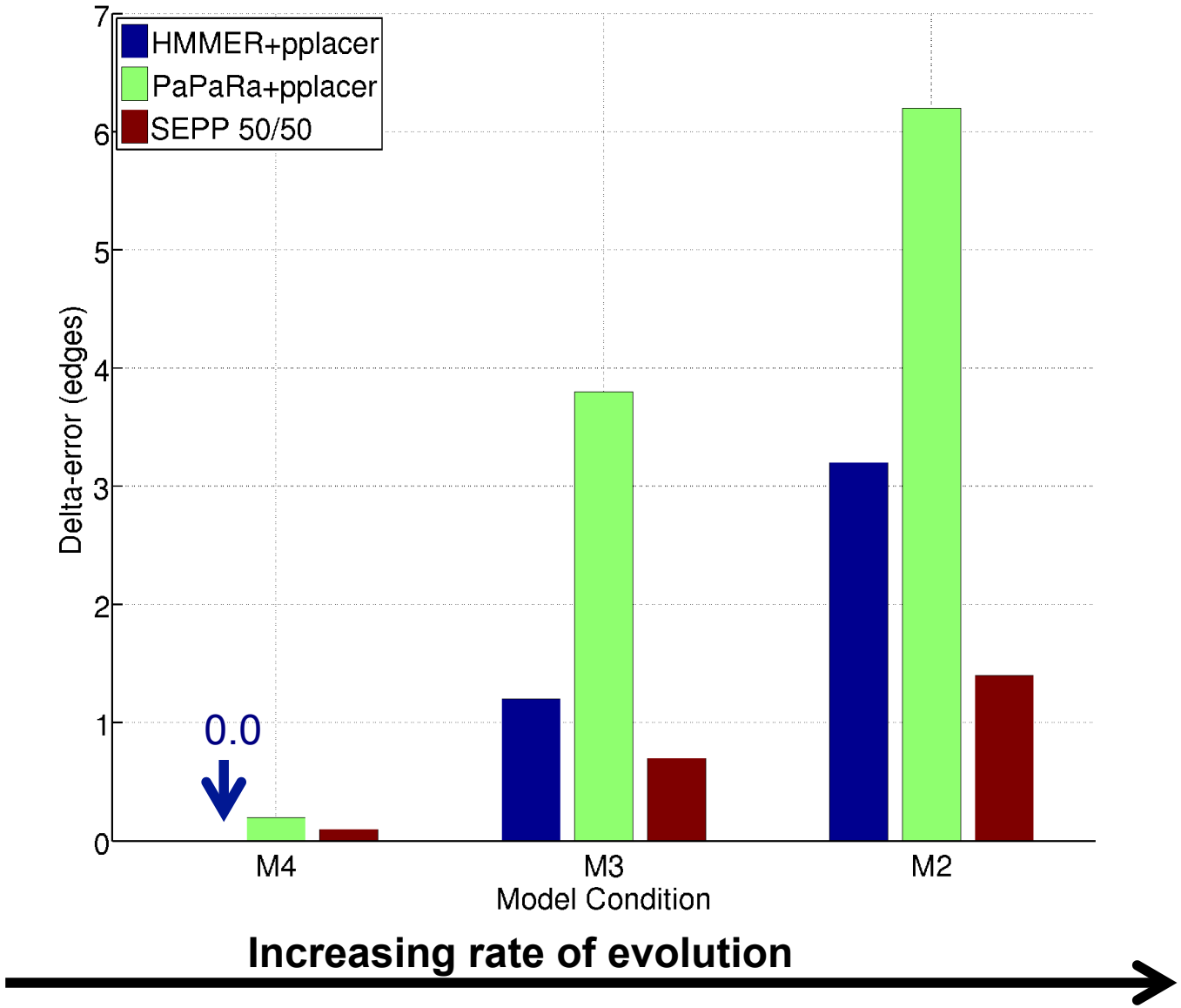
PhD students: Siavash Mirarab, Nam Nguyen, and Md. S. Bayzid

Undergrad: Keerthana Kumar

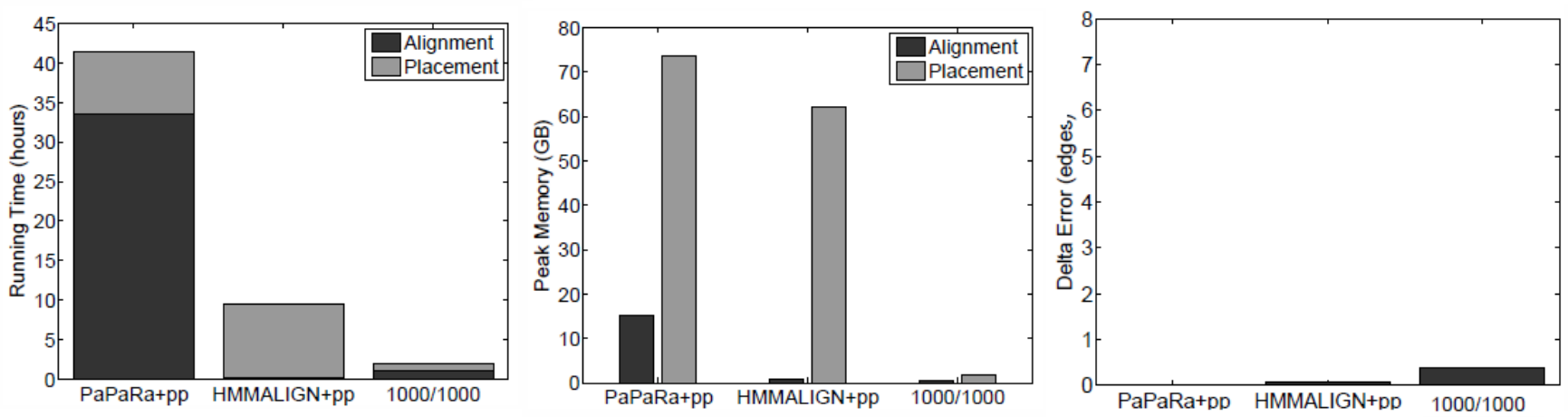
Lab Website: http://www.cs.utexas.edu/users/phylo

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

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

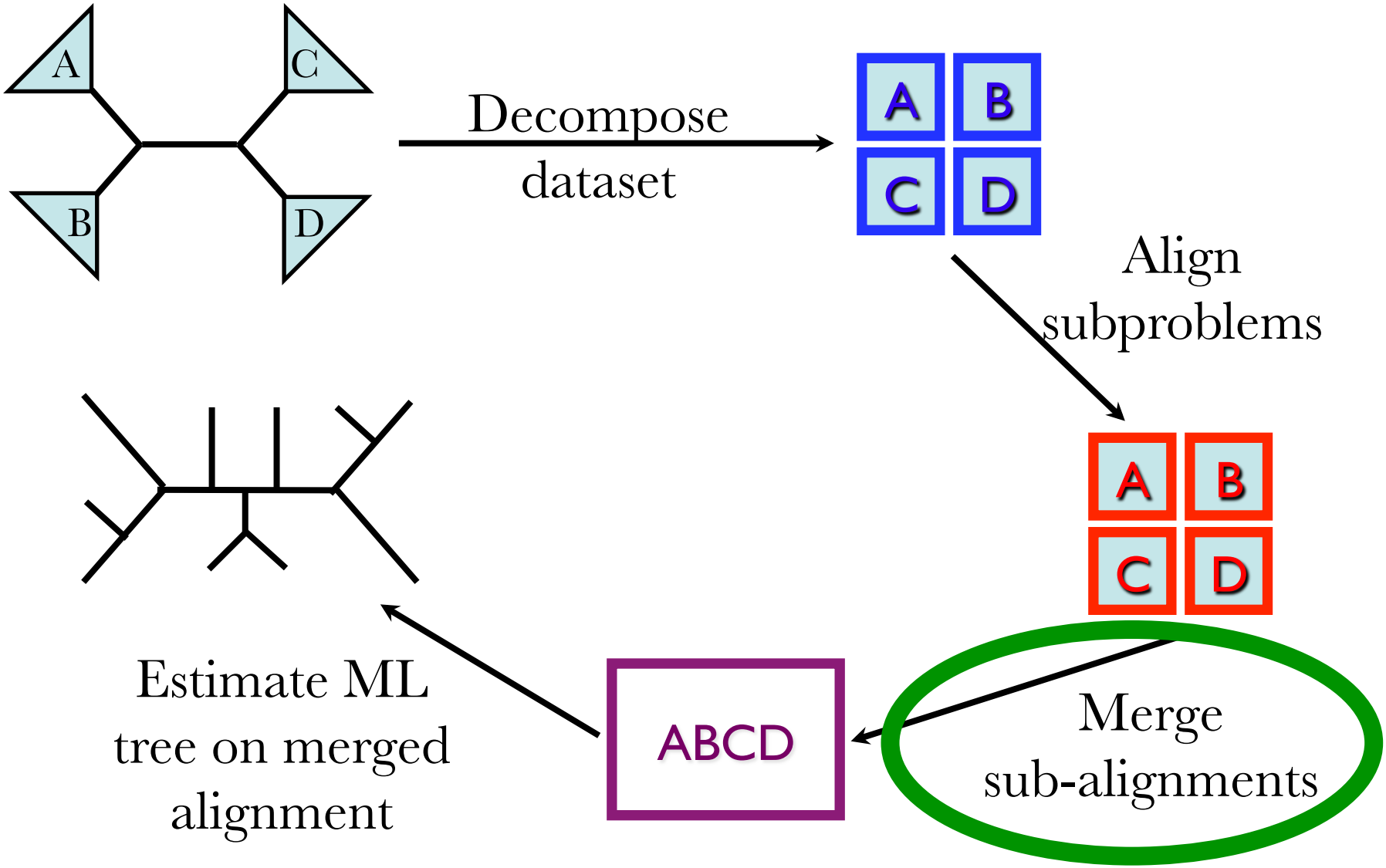
HMMALIGN+pplacer: ~30 days

SEPP 1000/1000: ~6 days

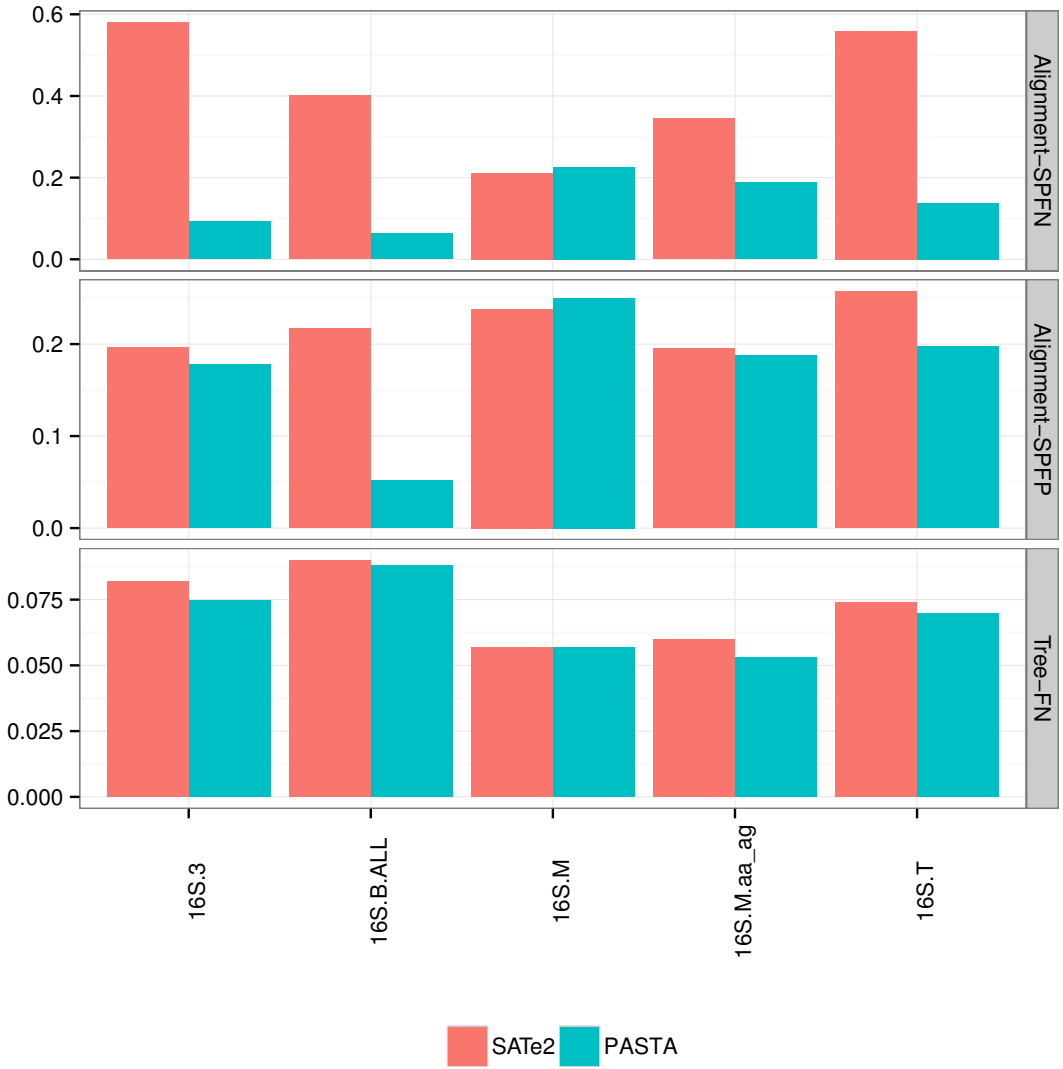
PASTA (in preparation)

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

Limitations



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 accuracy compared to SATé and UPP on large datasets
- Comparable tree accuracy to SATé
- Faster than SATé but slower than UPP
- Highly scalable – can analyze same datasets as UPP (1 million taxa)
- Highly parallelizable

In preparation – submission planned for Fall 2013

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.***