Estimating Ultra-large Phylogenies and Alignments

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

How did life evolve on earth?

- Courtesy of the Tree of Life project

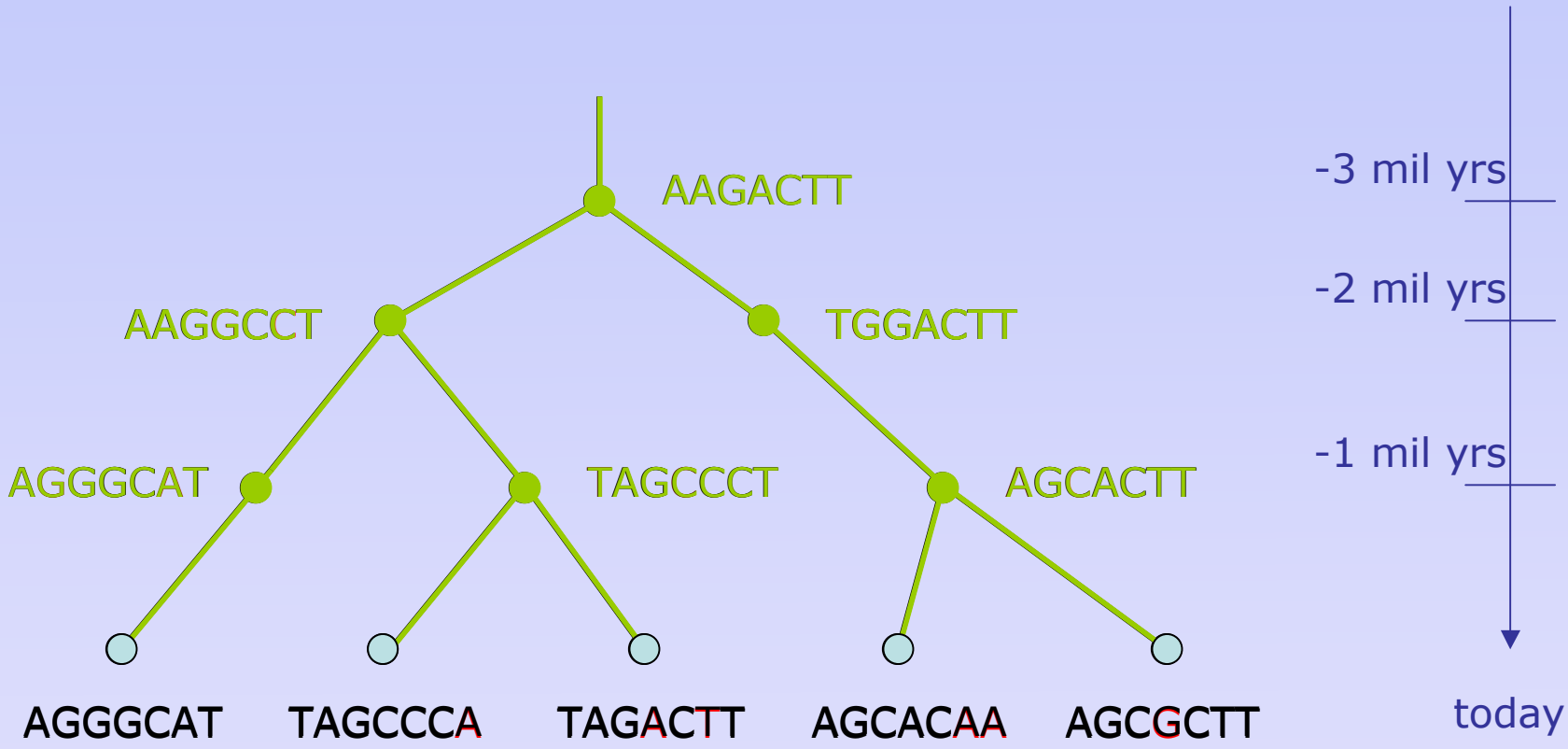
NP-hard optimization problems
Stochastic models of evolution
Statistical methods
Statistical performance issues

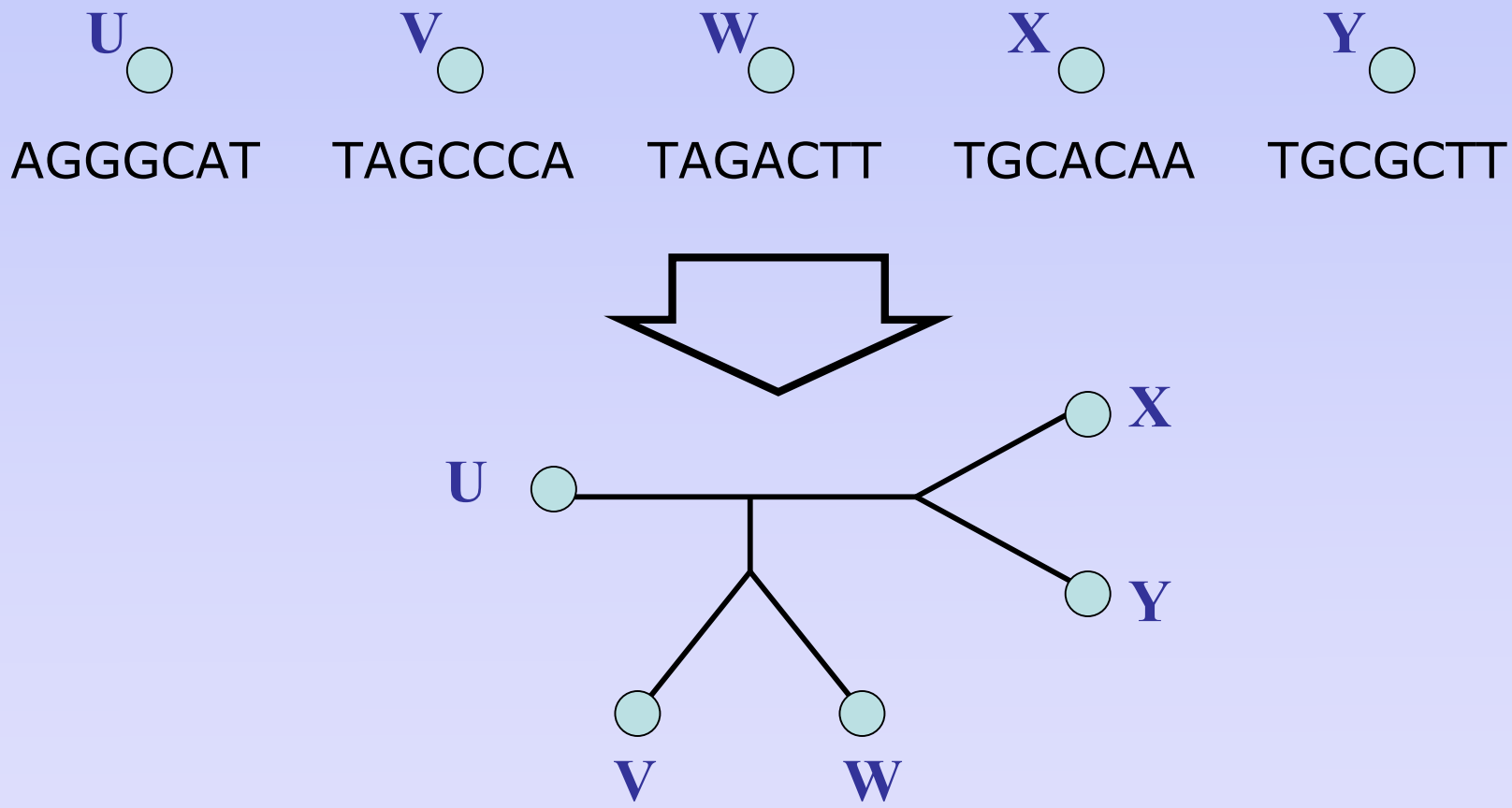
Millions of taxa

Important applications

Current projects (e.g., iPlant) will attempt to estimate phylogenies with upwards of 500,000 species

DNA Sequence Evolution




U
V
W
X
Y
AGGGCAT
TAGCCCA
TAGACTT
TGCACAA
TGCGCTT
X
U
Y
V
W

Current Research Projects

Method development:

- Large-scale multiple sequence alignment and phylogeny estimation
- Metagenomics
- Comparative genomics
- Estimating species trees from gene trees
- Supertree methods
- Phylogenetic estimation under statistical models

Dataset analyses (multi-institutional collaborations):

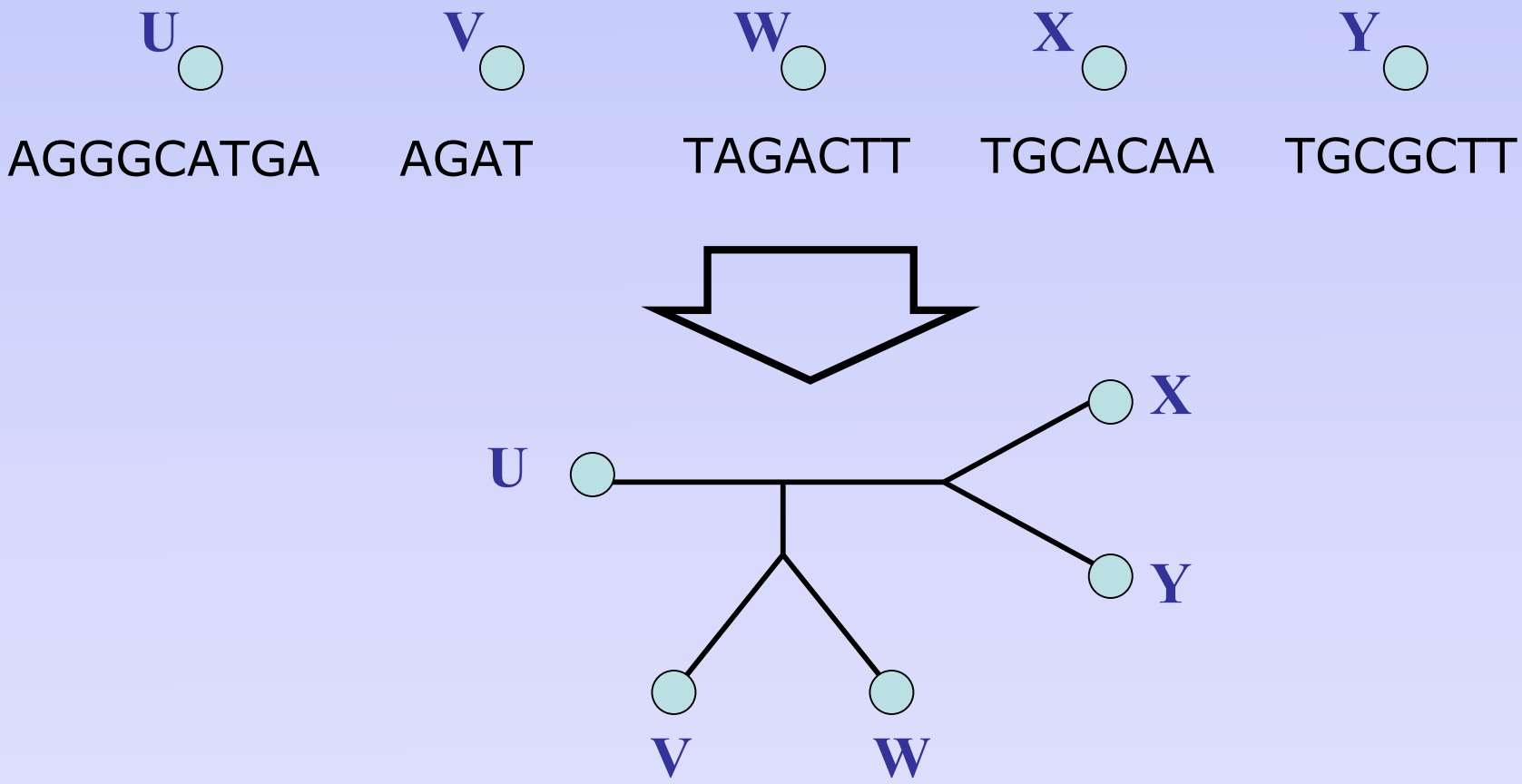
- Avian Phylogeny (and brain evolution)
- Human Microbiome
- Thousand Transcriptome (1KP) Project
- Conifer evolution

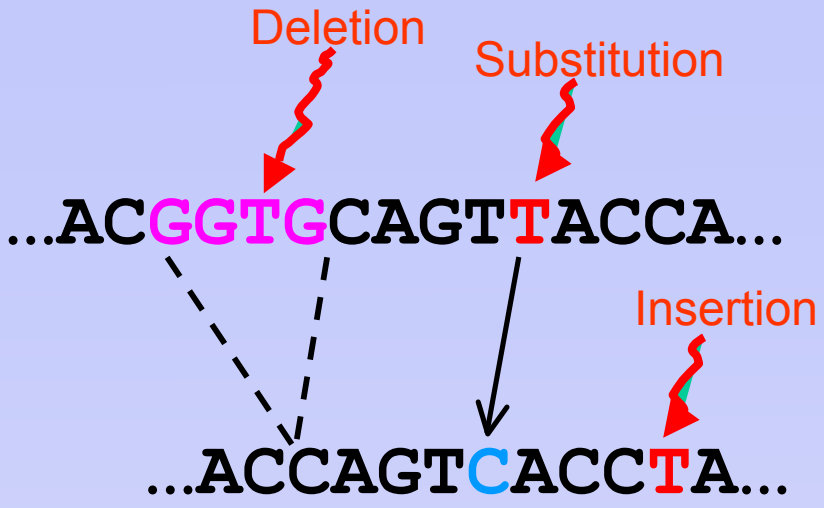
- **SATé**: Simultaneous Alignment and Tree Estimation (Liu et al., Science 2009, and Liu et al. Systematic Biology, in press)
- **DACTAL**: Divide-and-Conquer Trees without alignments (Nelesen et al., in preparation)
- **SEPP**: SATé-enabled Phylogenetic Placement (Mirarab, Nguyen, and Warnow, PSB 2012, in press)

Part 1: SATé

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

Liu et al., Systematic Biology (in press)


U
V
W
X
Y
AGGGCATGA
AGAT
TAGACTT
TGCACAA
TGCGCTT
X
U
Y
V
W



...ACGGTGCAGTTACC-A...

...AC----CAGTCACCTA...

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: Multiple Sequence Alignment

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



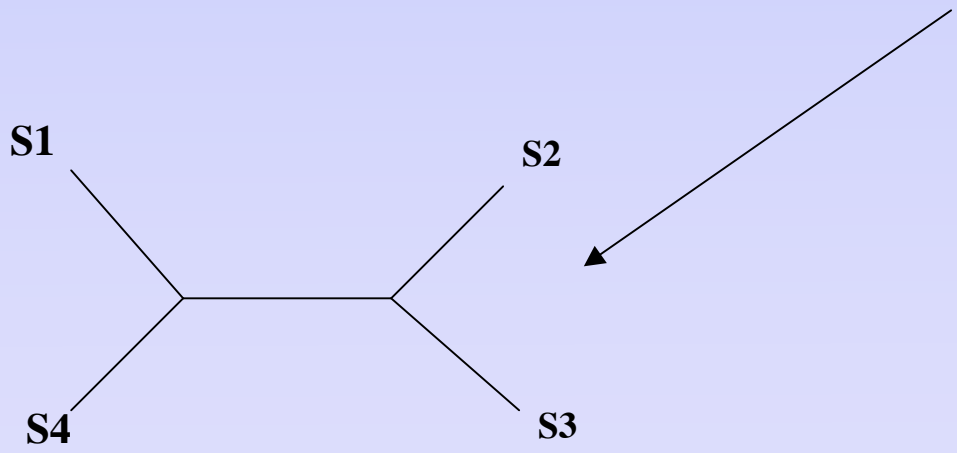
```
S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAC--GACCGC--
S3 = TAG-CT-------GACCGC--
S4 = -------TCAC--GACCGACA
```

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

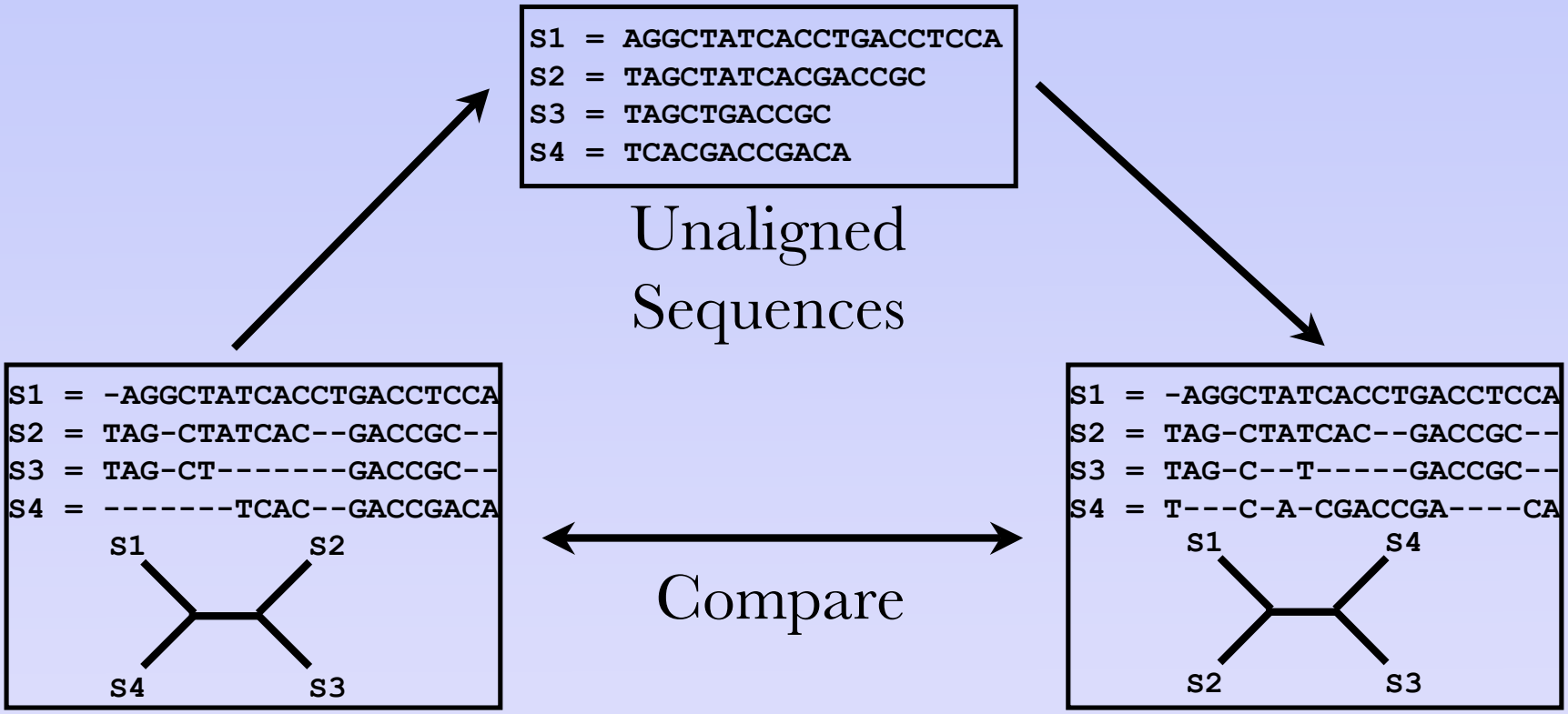
- Clustal
- POY (and POY*)
- Probcons (and Probtree)
- Probalign
- MAFFT
- Muscle
- Di-align
- T-Coffee
- Prank (PNAS 2005, Science 2008)
- Opal (ISMB and Bioinf. 2007)
- *FSA (PLoS Comp. Bio. 2009)*
- *Infernal (Bioinf. 2009)*
- Etc.

Phylogeny methods

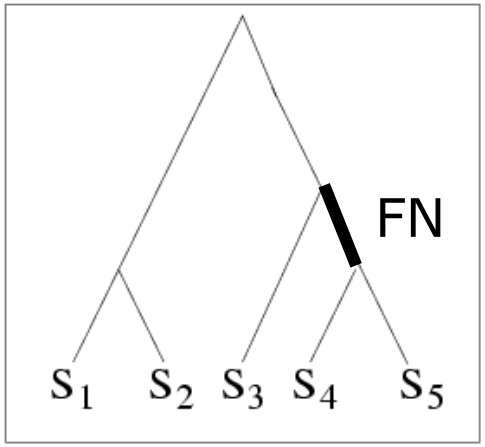
- Bayesian MCMC
- Maximum parsimony
- Maximum likelihood
- Neighbor joining
- FastME
- UPGMA
- Quartet puzzling
- Etc.

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

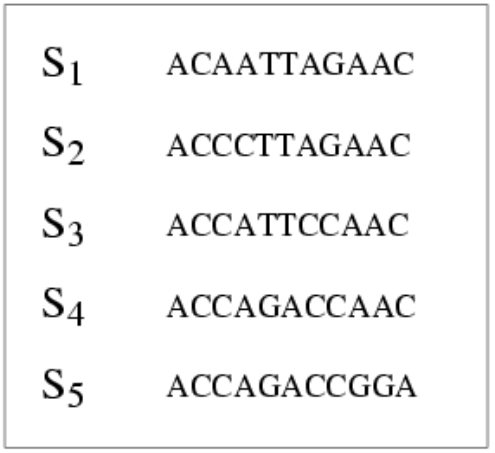
Simulation Studies



Quantifying Error



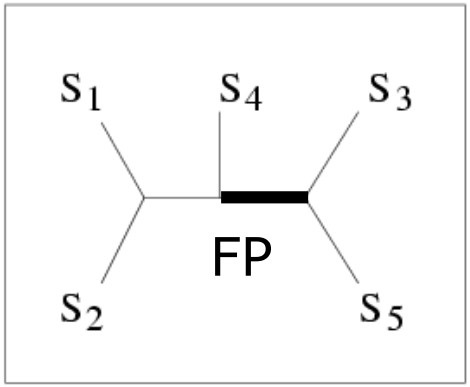
TRUE TREE



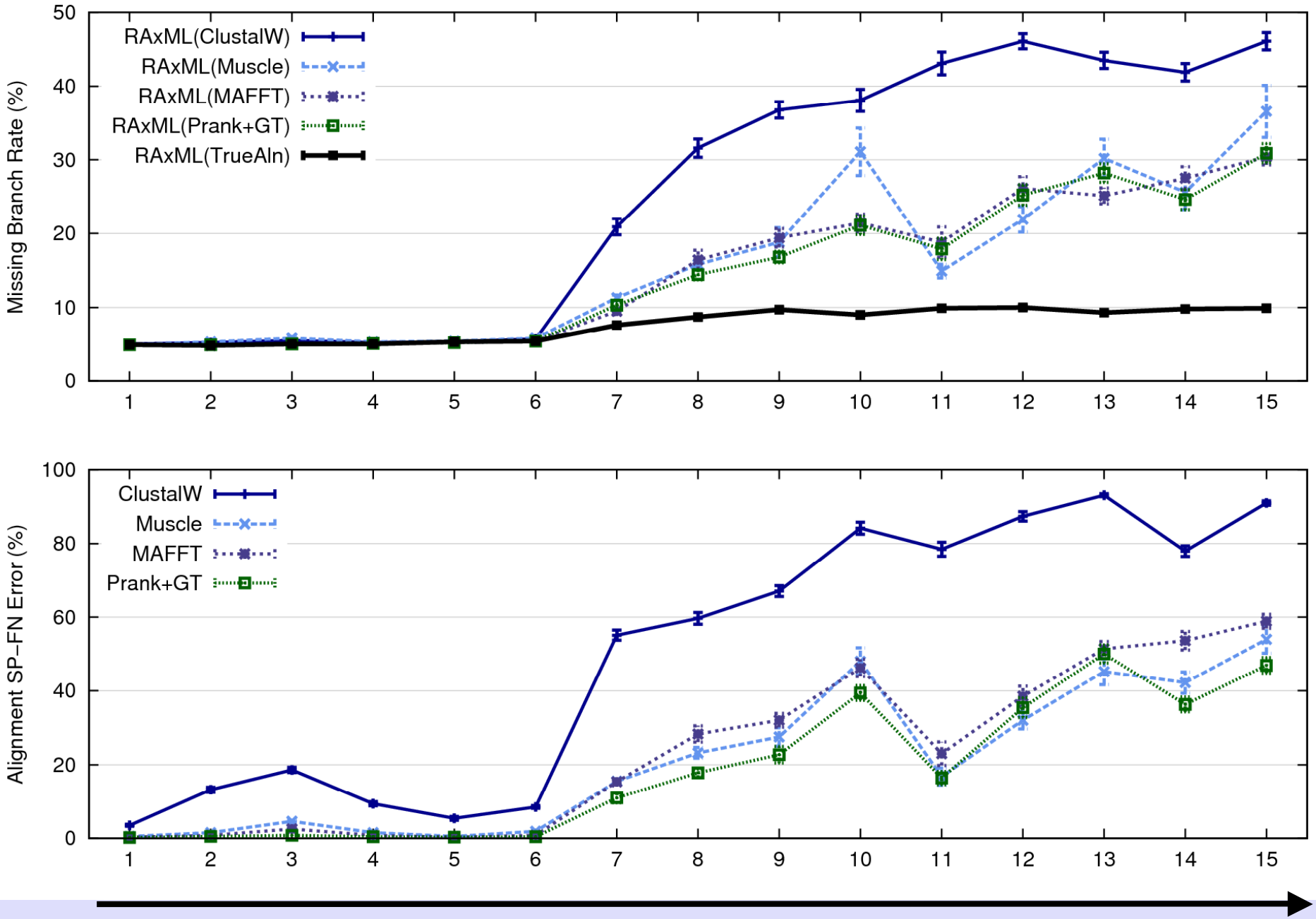
DNA SEQUENCES

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

50% error rate



INFERRED TREE


Missing Branch Rate (%)
RAxML(ClustalW)
RAxML(Muscle)
RAxML(MAFFT)
RAxML(Prank+GT)
RAxML(TrueAln)
Alignment SP–FN Error (%)
ClustalW
Muscle
MAFFT
Prank+GT
1000 taxon models, ordered by difficulty (Liu et al., 2009)

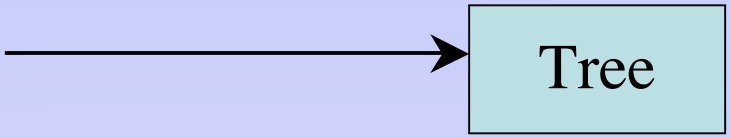
Problems with the two-phase approach

- 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.
- *Potentially useful markers are often discarded* if they are difficult to align.

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

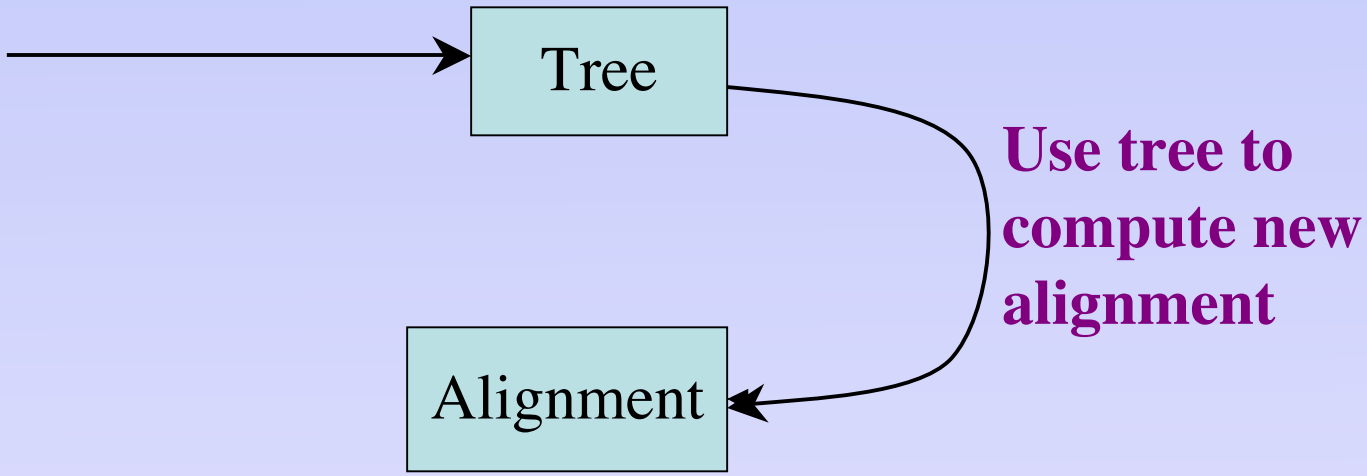
SATé Algorithm

Obtain initial alignment
and estimated ML tree



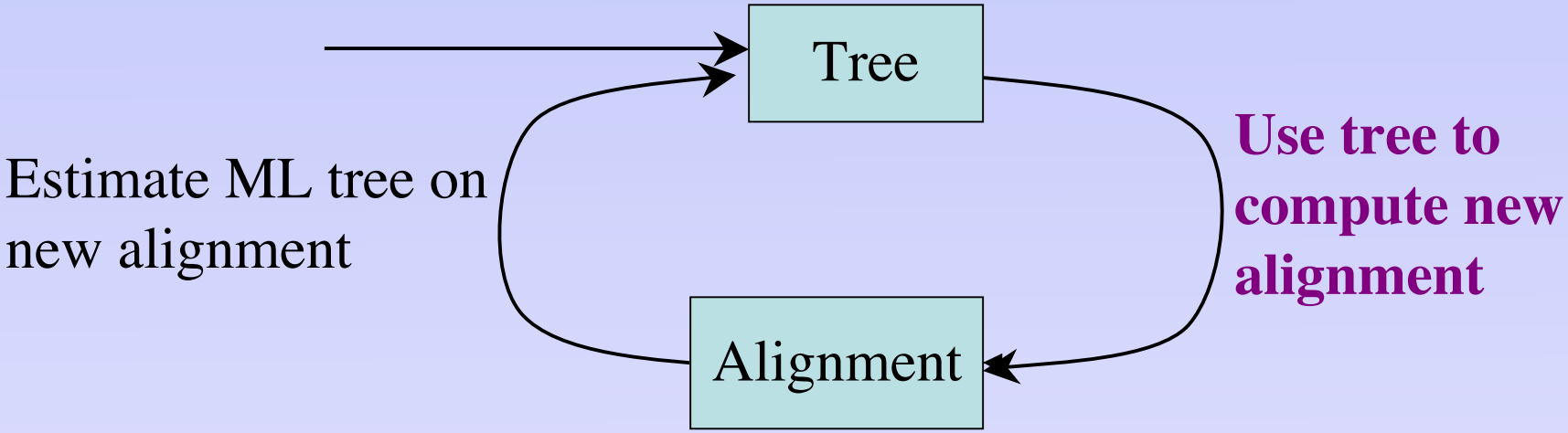
SATé Algorithm

Obtain initial alignment
and estimated ML tree



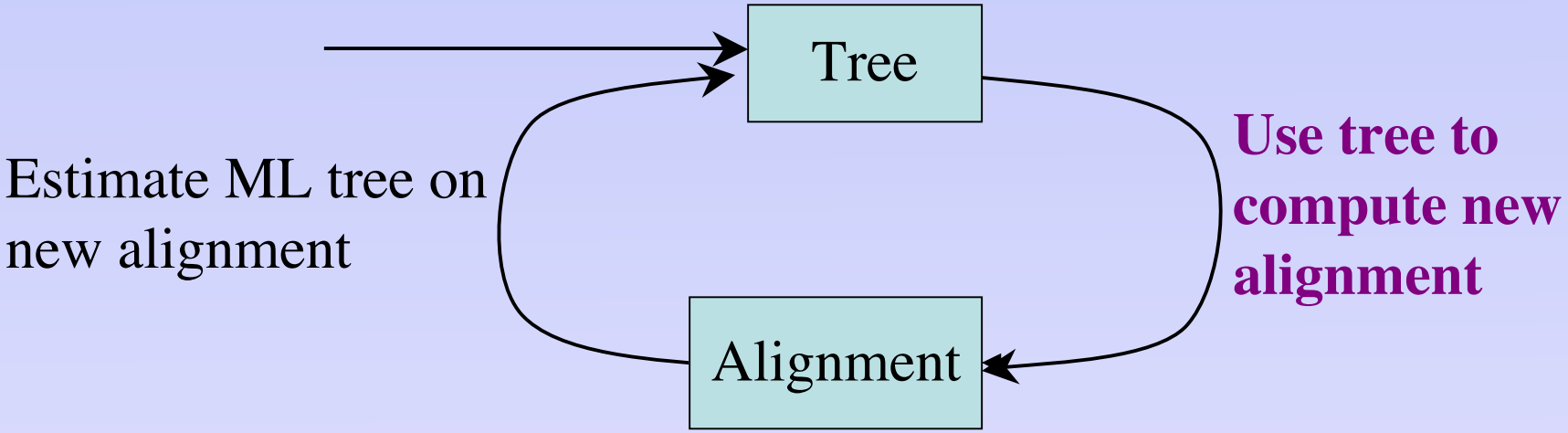
SATé Algorithm

Obtain initial alignment
and estimated ML tree



SATé Algorithm

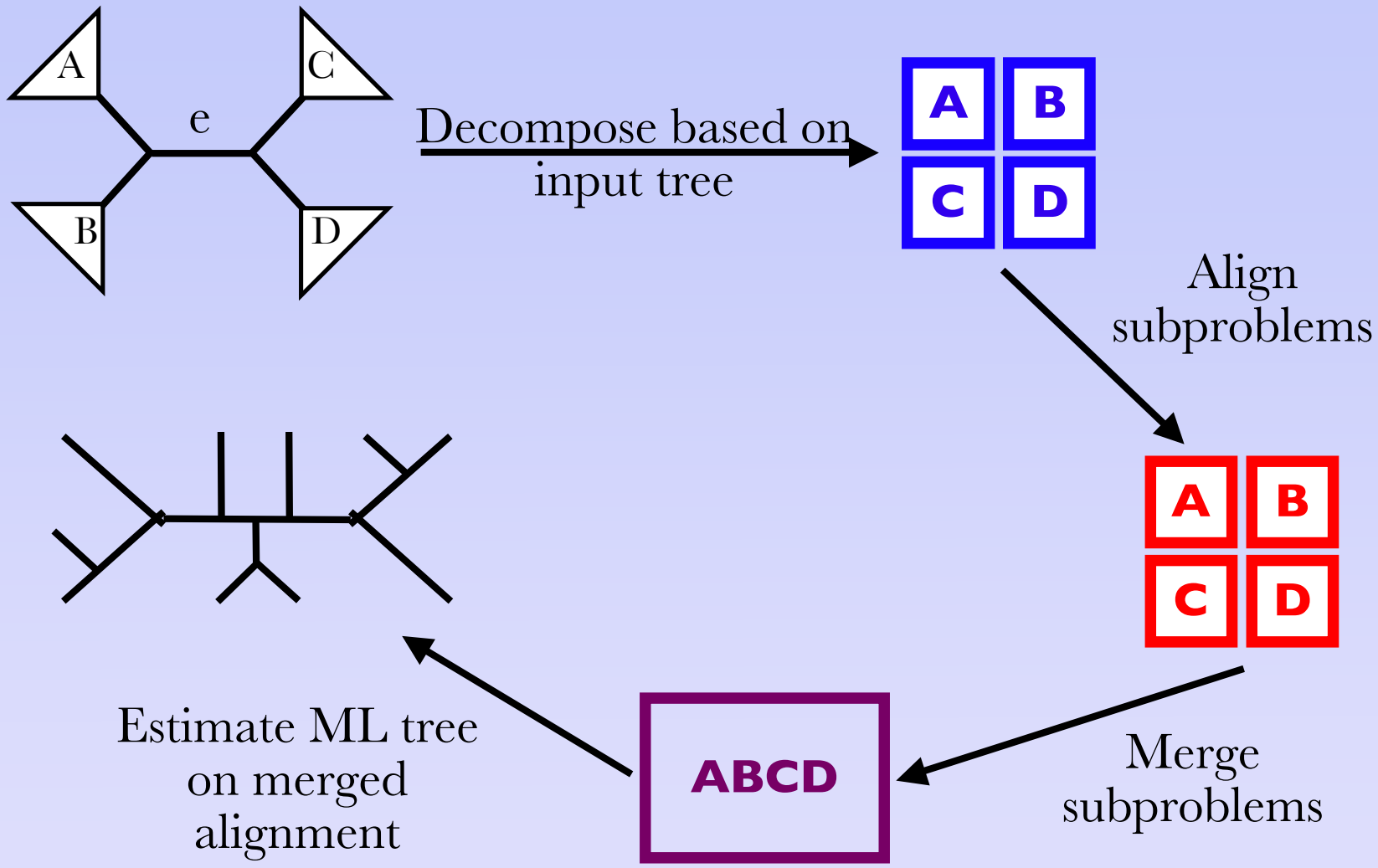
Obtain initial alignment
and estimated ML tree

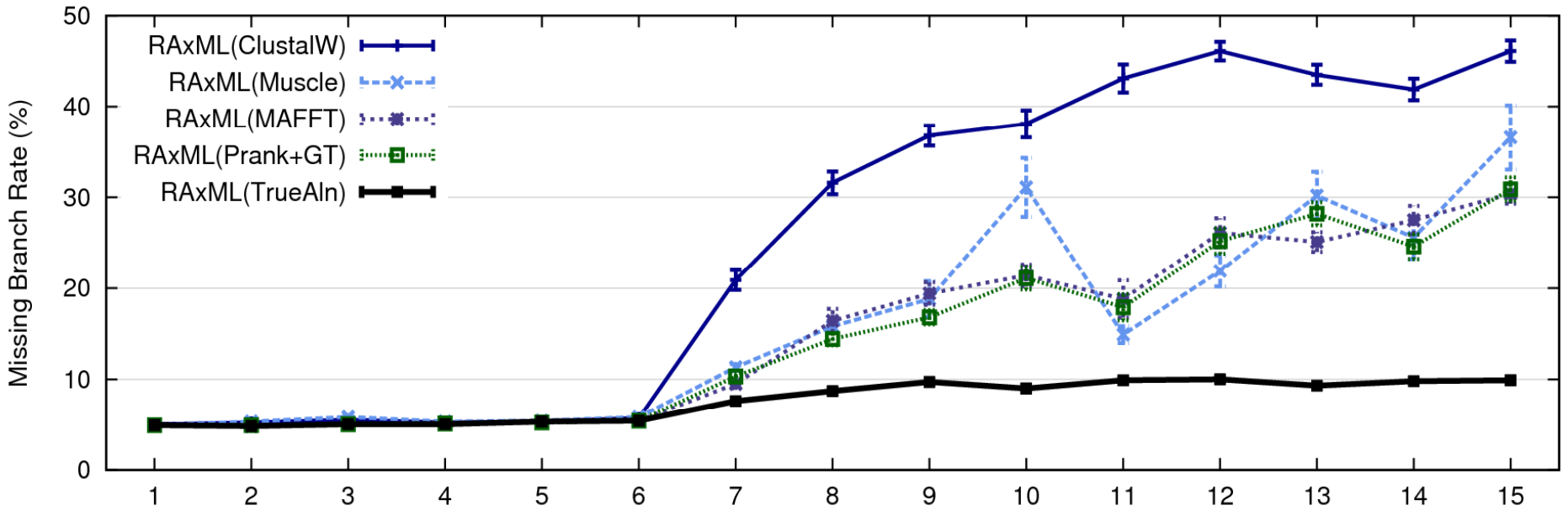


If new alignment/tree pair has worse ML score, realign using a different decomposition

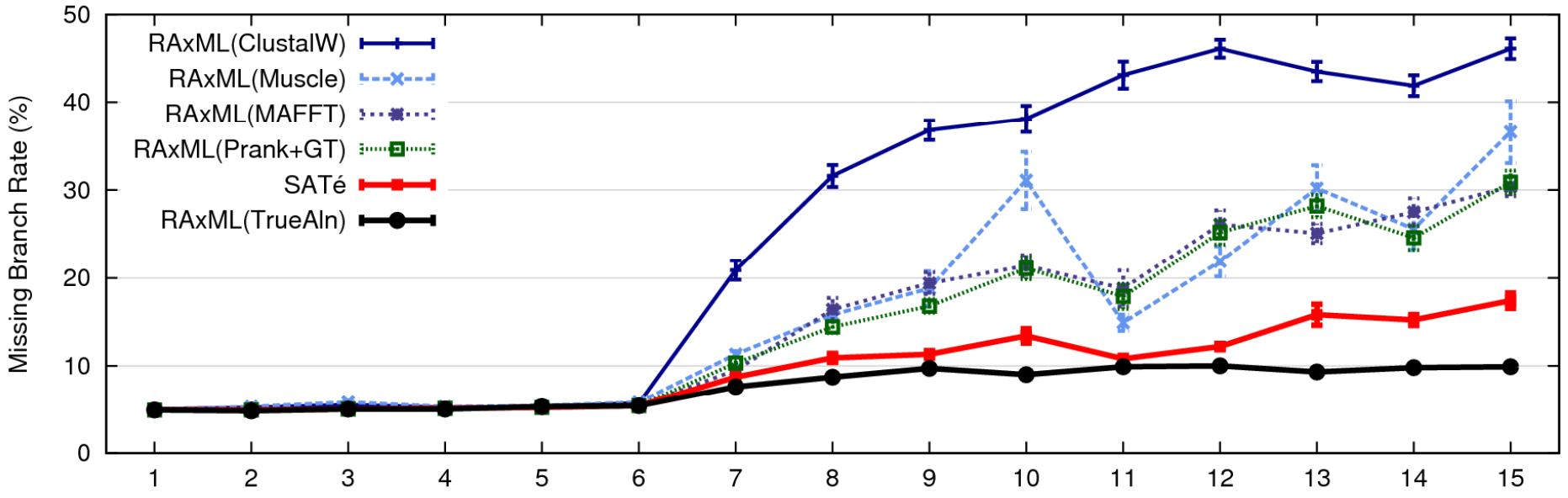
Repeat until termination condition (typically, 24 hours)

One SATé iteration (really 32 subsets)





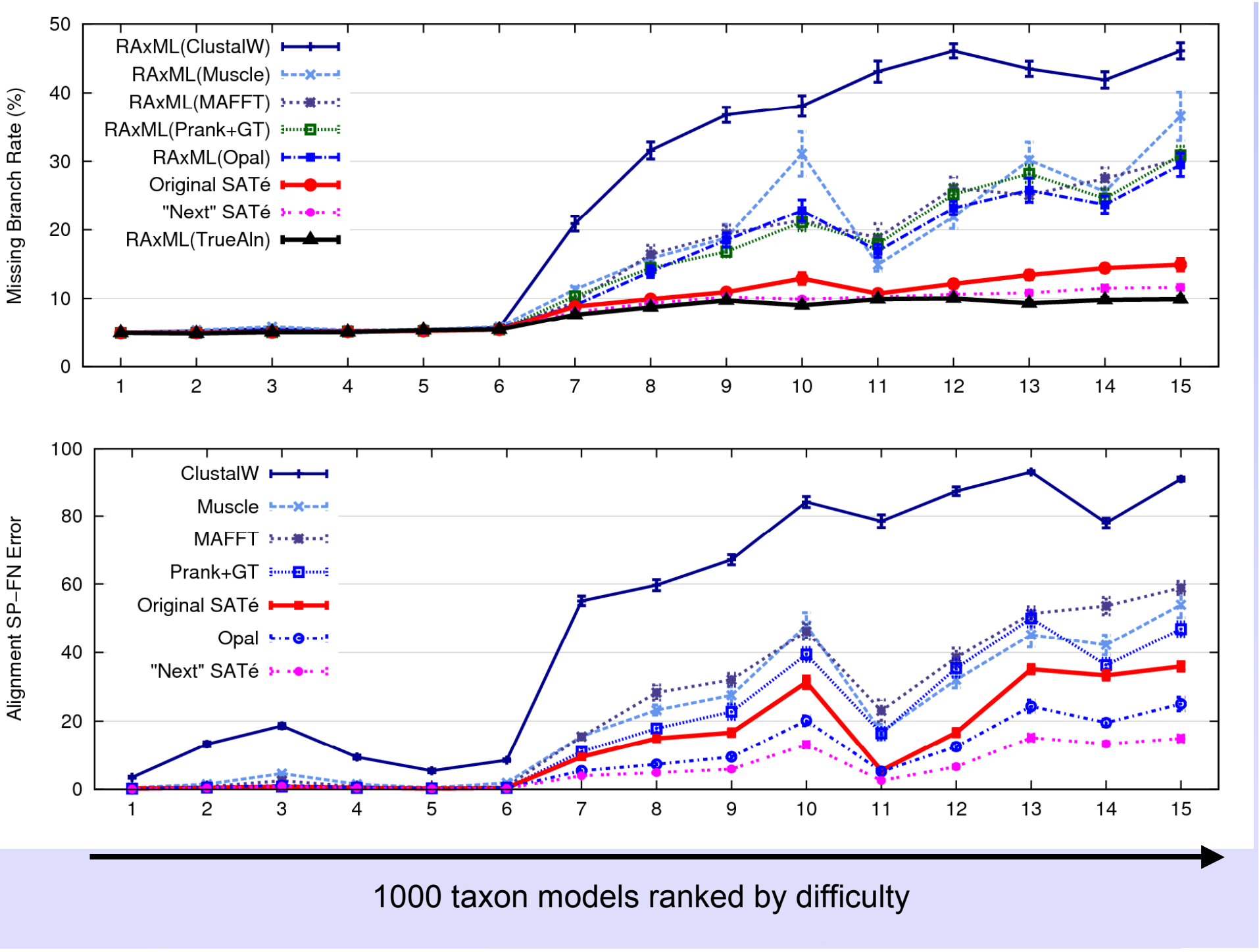
1000 taxon models, ordered by difficulty



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

Understanding SATé

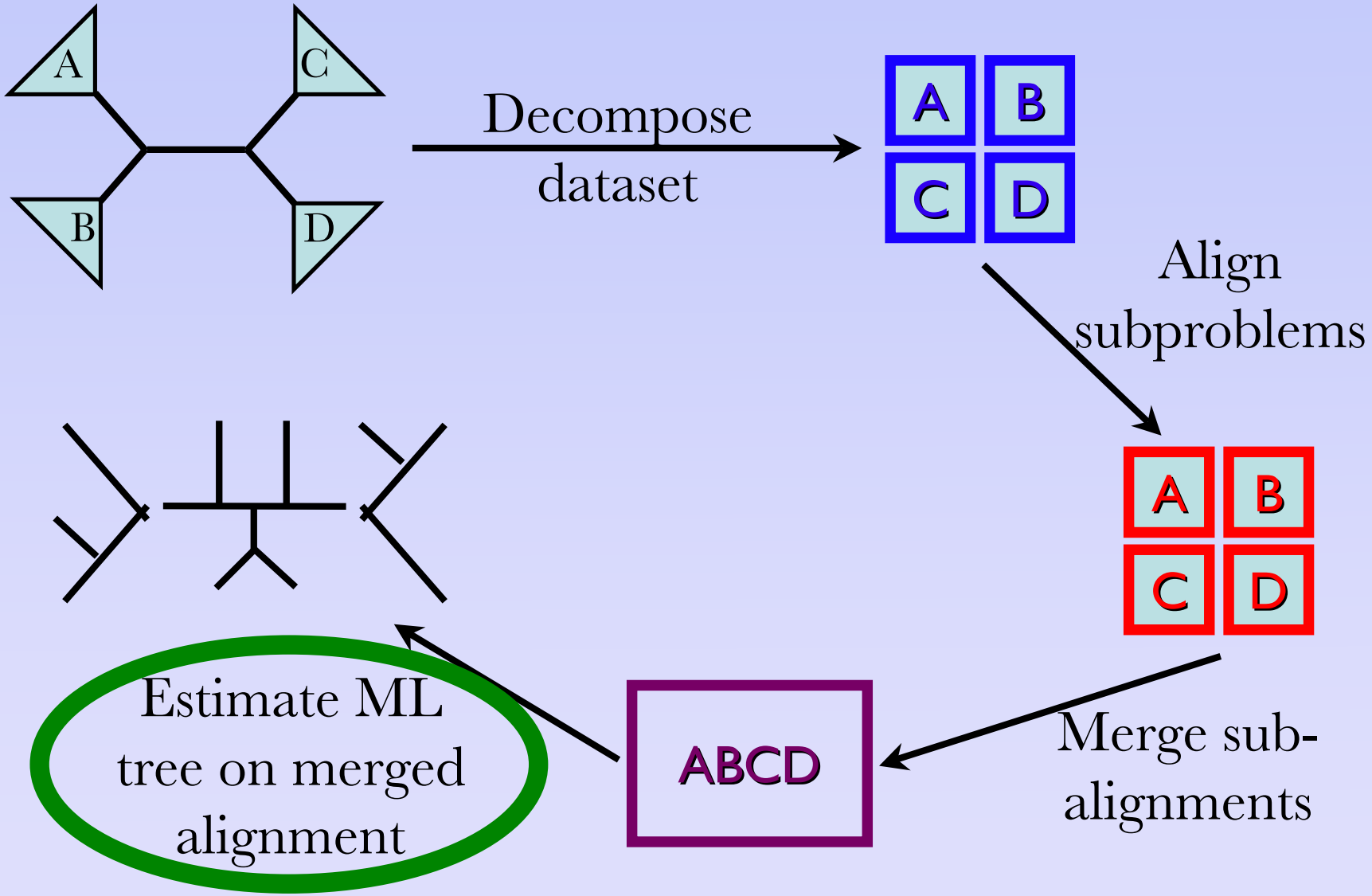
- Observations: (1) subsets of taxa that are small enough, closely related, and densely sampled are aligned more accurately than others.
- SATé-1 produces subsets that are closely related and densely sampled, but not small enough.
- SATé-2 (“next SATé”) changes the design to produce smaller subproblems.
- The next iteration starts with a more accurate tree. This leads to a better alignment, and a better tree.

Software

In use by research groups around the world

- Kansas SATé software developers: Mark Holder, Jiaye Yu, and Jeet Sukumaran
- Downloadable software for various platforms
- Easy-to-use GUI
- http://phylo.bio.ku.edu/software/sate/sate.html

Limitations of SATé-I and -II



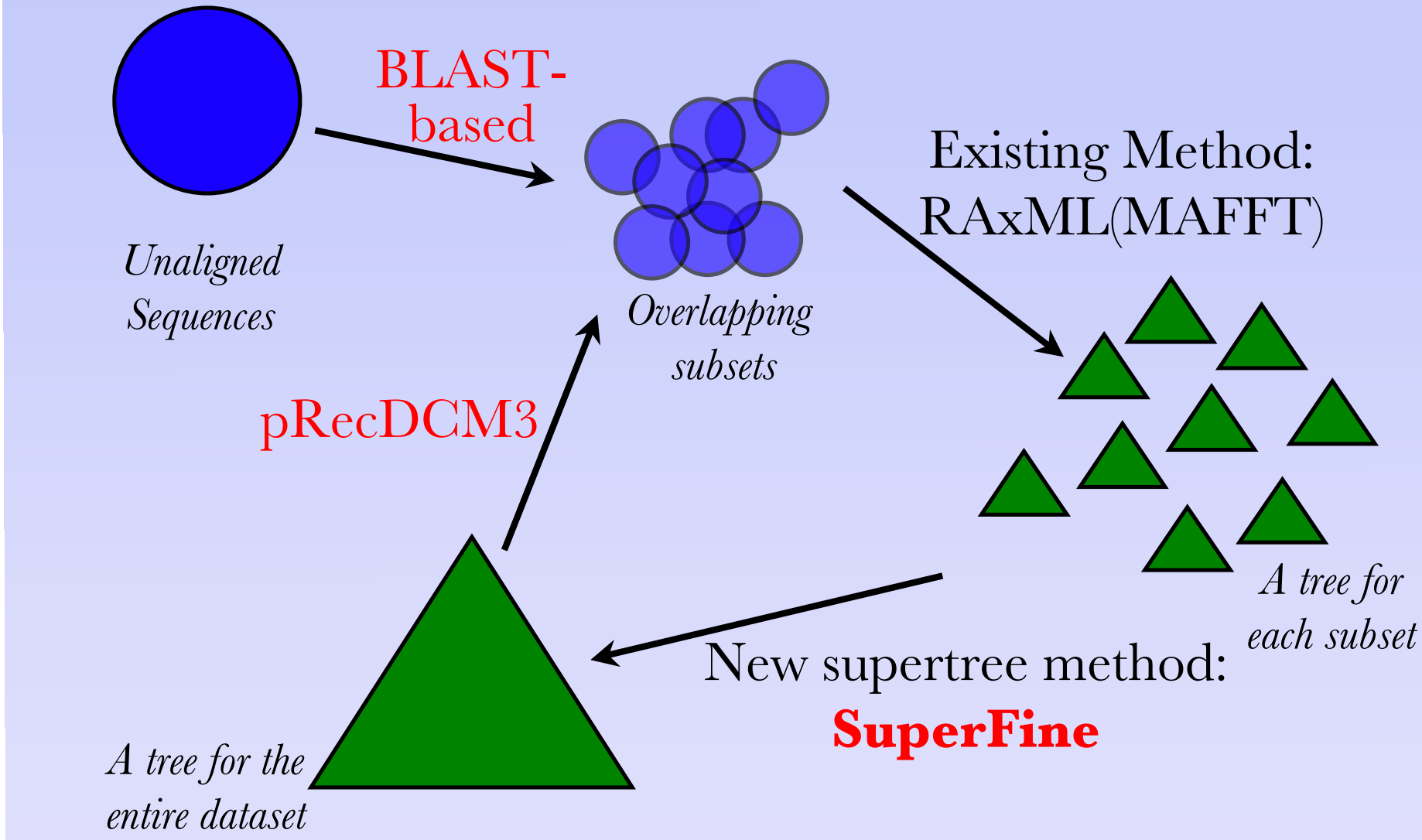
Part II: DACTAL

(Divide-And-Conquer Trees (without) ALignments)

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

(Nelesen, Liu, Wang, Linder, and Warnow, in preparation)

DACTAL


BLAST-based
Unaligned Sequences
Overlapping subsets
Existing Method: RAxML(MAFFT)
A tree for each subset
pRecDCM3
New supertree method: SuperFine
A tree for the entire dataset

Average of 3 Largest CRW Datasets

CRW: Comparative RNA database, datasets 16S.B.ALL, 16S.T, and 16S.3

6,323 to 27,643 sequences

These datasets have curated alignments based on secondary structure

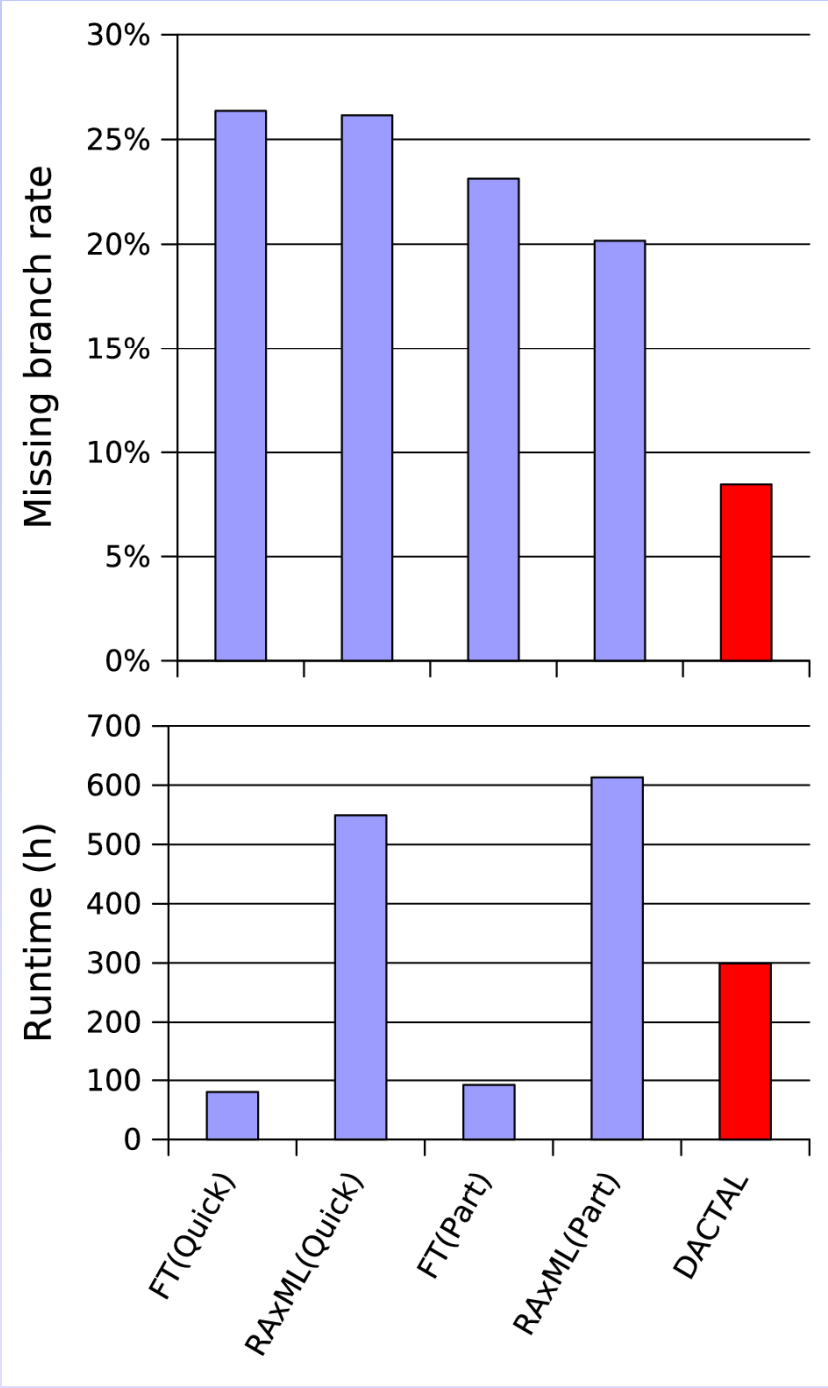
Reference trees are 75% RAxML bootstrap trees

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

DACTAL is robust to starting trees

PartTree and Quicktree are the only MSA methods that run on all 3 datasets

FastTree (FT) and RAxML are ML methods



DACTAL outperforms SATé

- DACTAL faster and matches or improves upon accuracy of SATé for datasets with 1000 or more taxa

- The biggest gains are on the very large datasets

Part III: SEPP

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

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

Metagenomic data analysis

NGS data produce fragmentary sequence data

Metagenomic analyses include unknown species

Taxon identification: given short sequences, identify the species for each fragment

Applications: Human Microbiome

Issues: accuracy and speed

Phylogenetic Placement

- **Input**: **Backbone** alignment and tree on full-length sequences, and a set of **query** sequences (short fragments)
- **Output**: Placement of query sequences on backbone tree

Phylogenetic placement can be used for taxon identification, but it has general applications for phylogenetic analyses of NGS data.

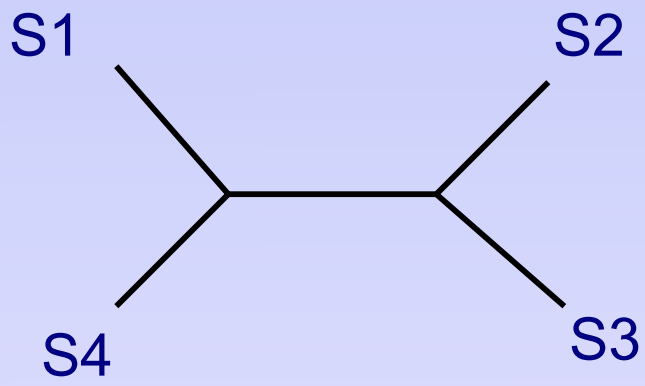
Phylogenetic Placement

- Align each query sequence to backbone alignment

- Place each query sequence into backbone tree, using extended alignment

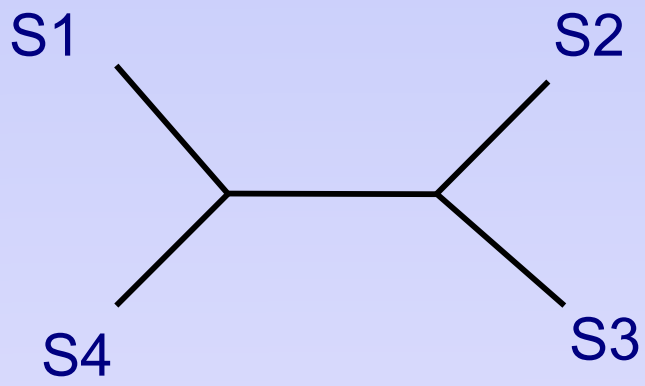
Align Sequence

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



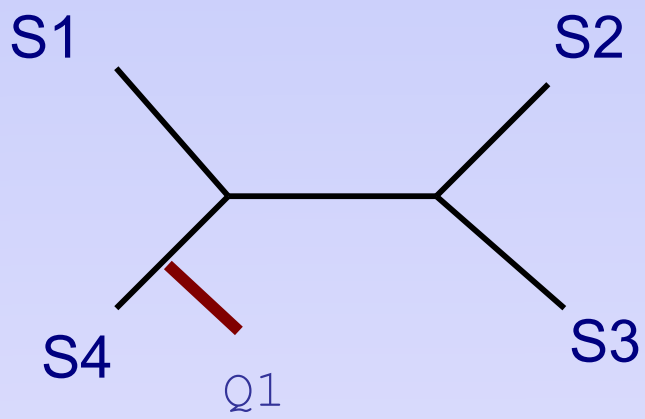
Align Sequence

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

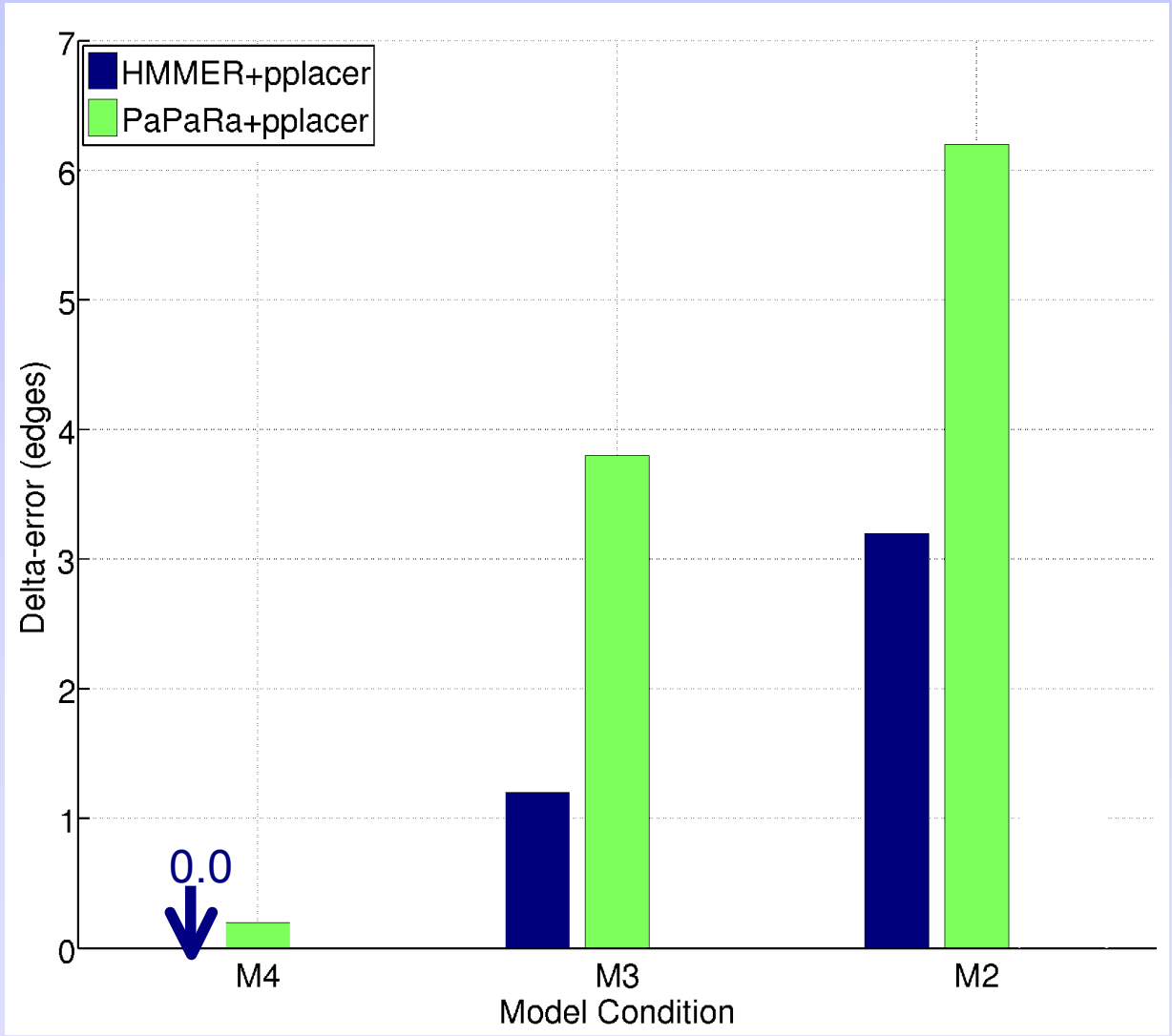


Phylogenetic Placement

- 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)

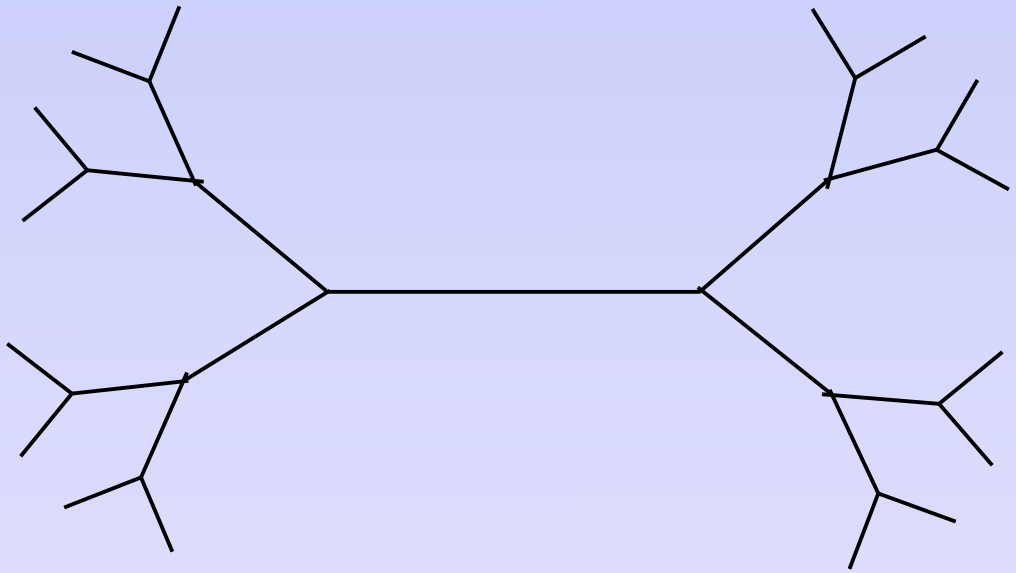
Note: pplacer and EPA use maximum likelihood

HMMER vs. PaPaRa Alignments

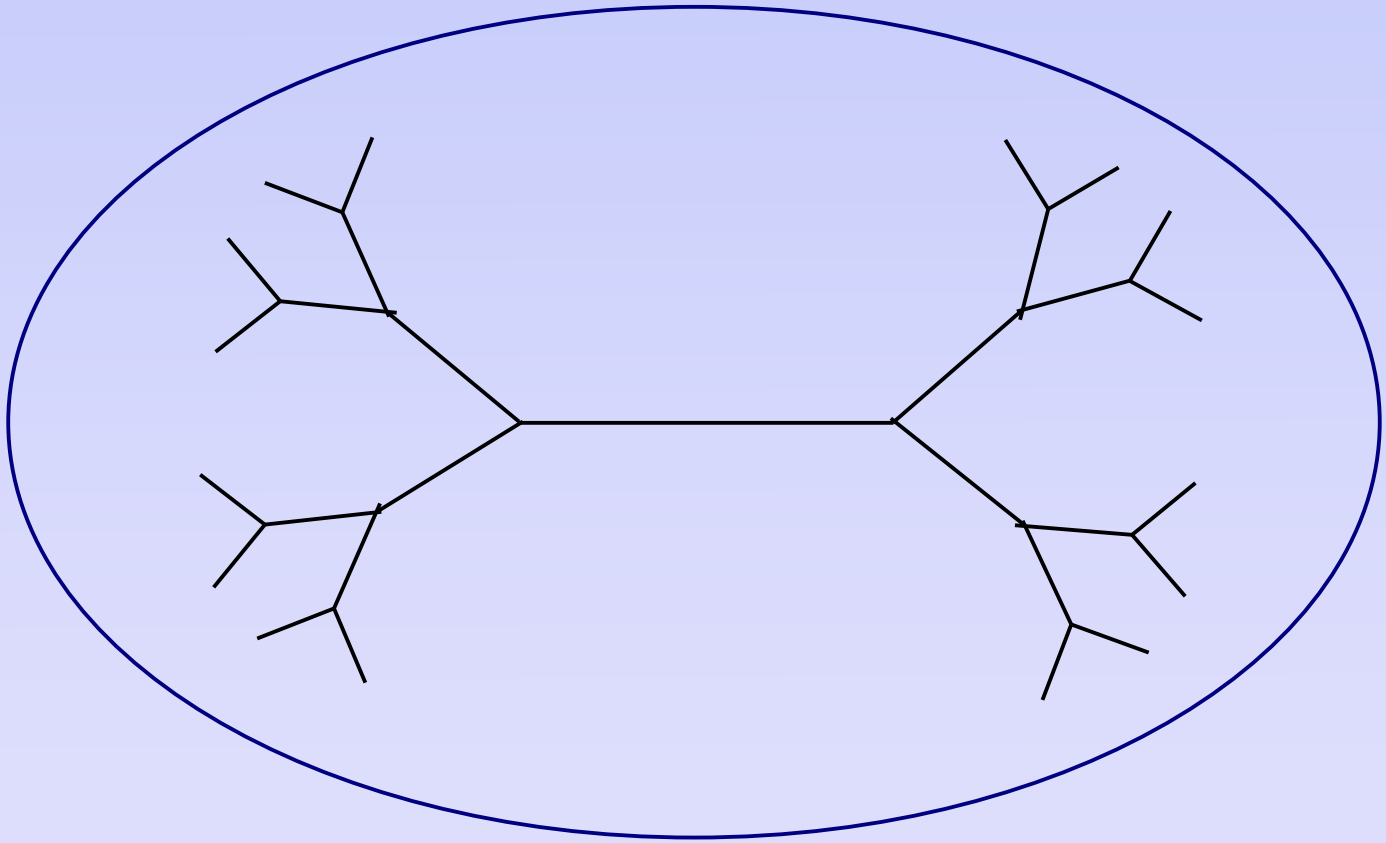


Increasing rate of evolution

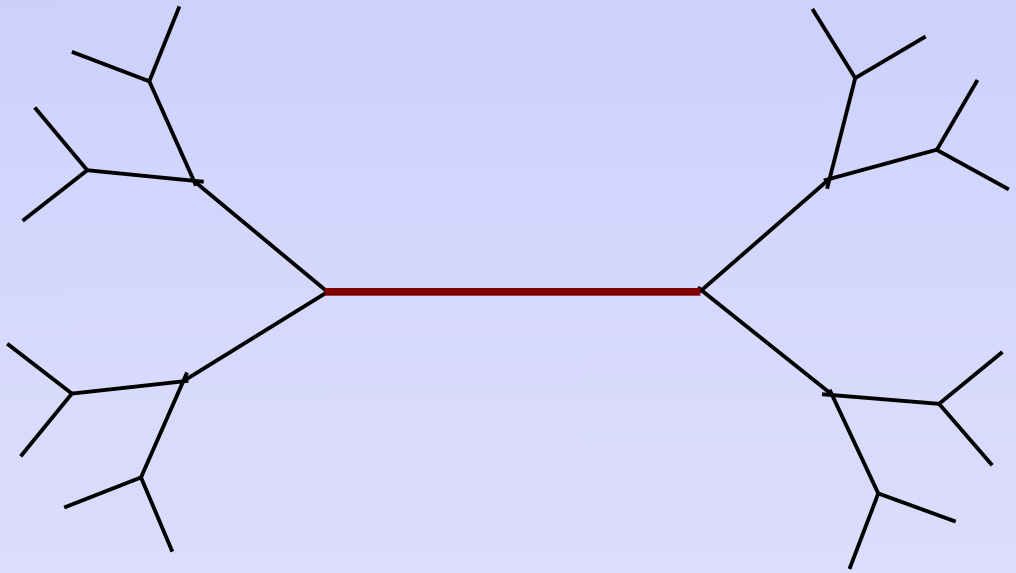
Insights from SATé



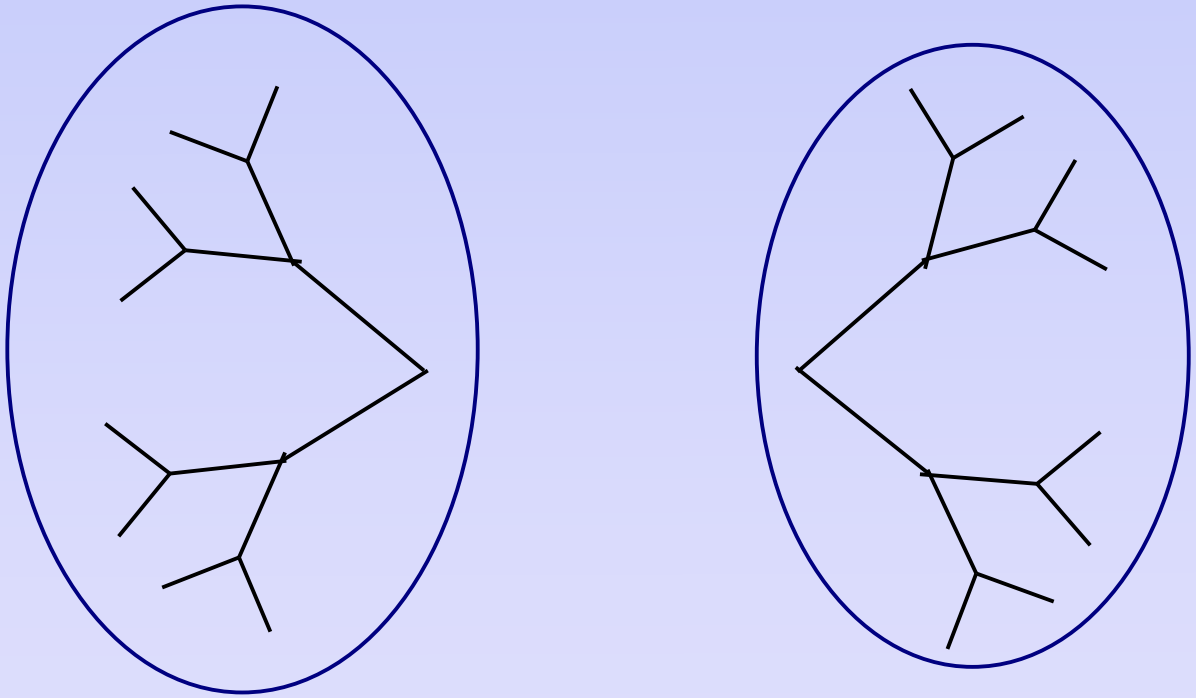
Insights from SATé



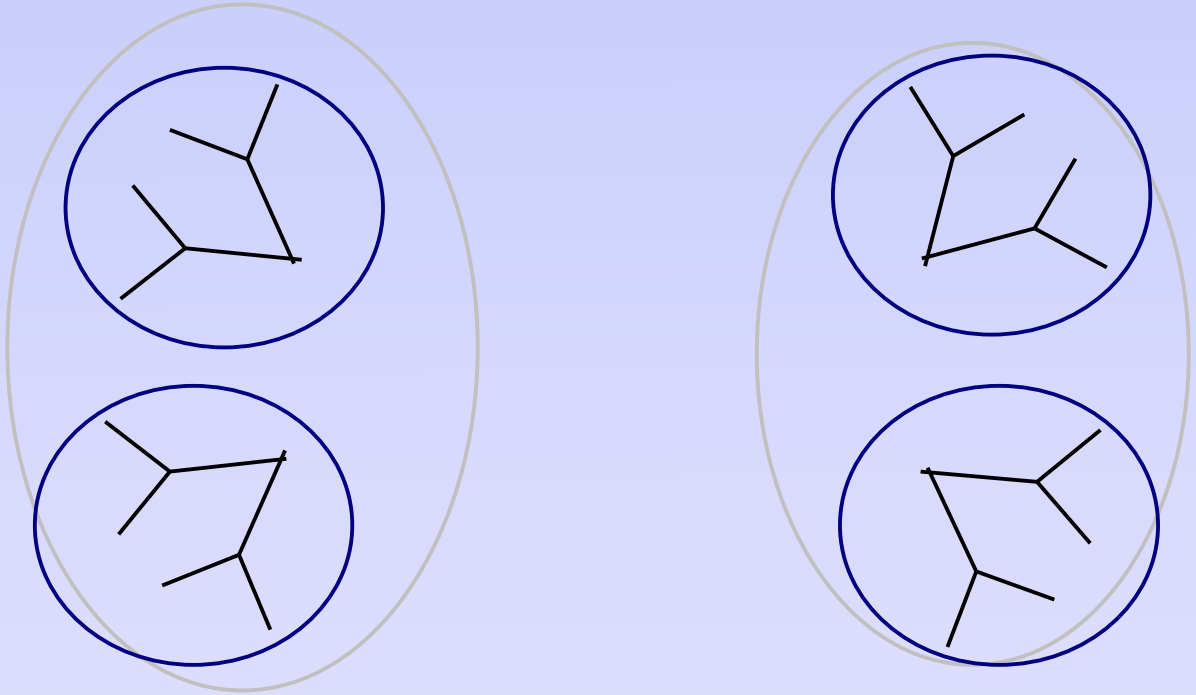
Insights from SATé



Insights from SATé



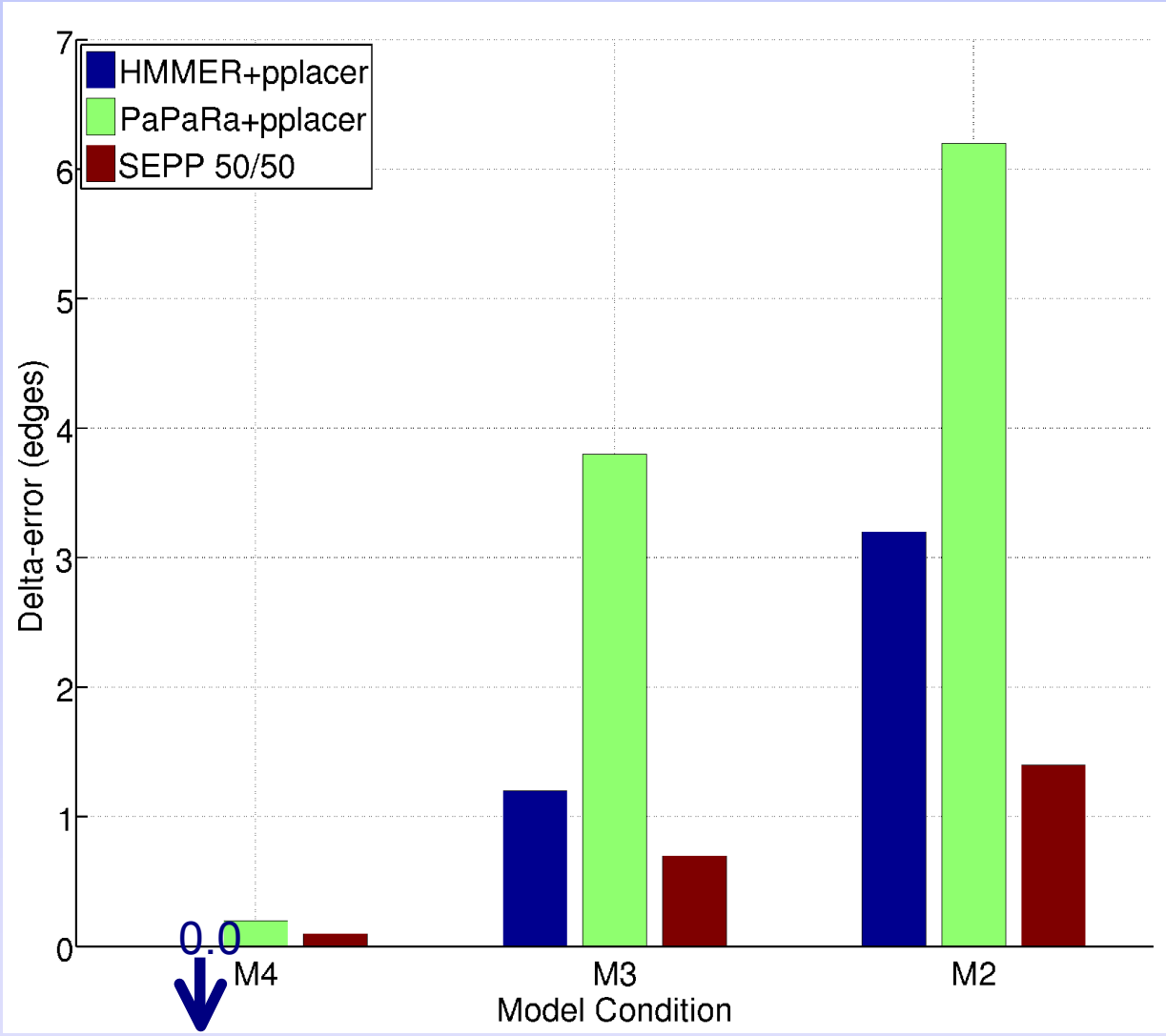
Insights from SATé



SEPP Parameter Exploration

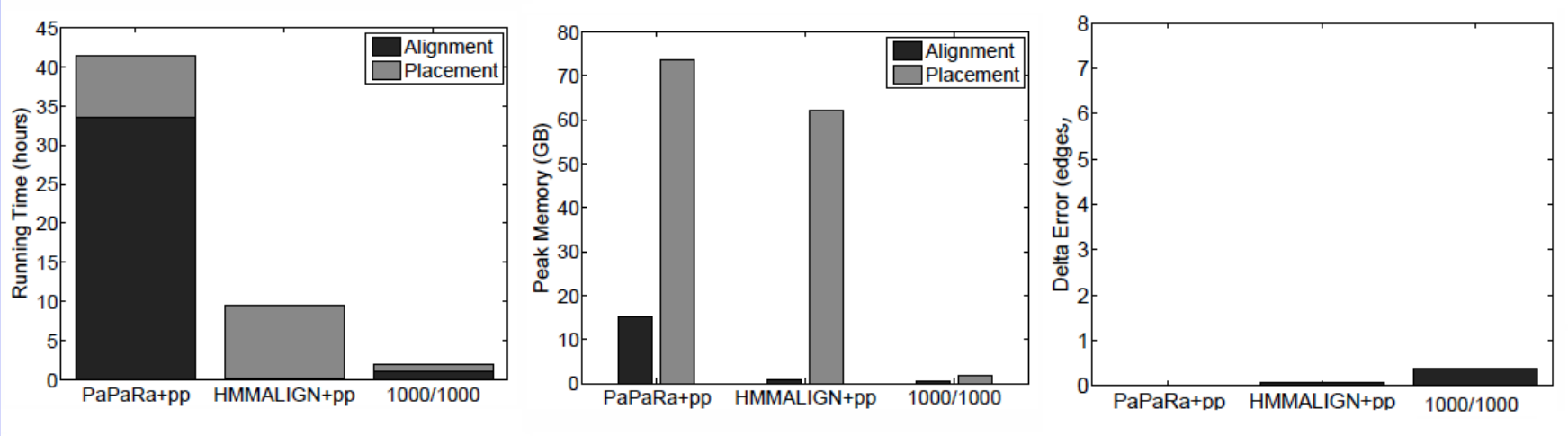
- Alignment subset size and placement subset size impact the accuracy, running time, and memory of SEPP
- 10% rule (subset sizes 10% of backbone) had best overall performance

SEPP (10%-rule) on simulated data



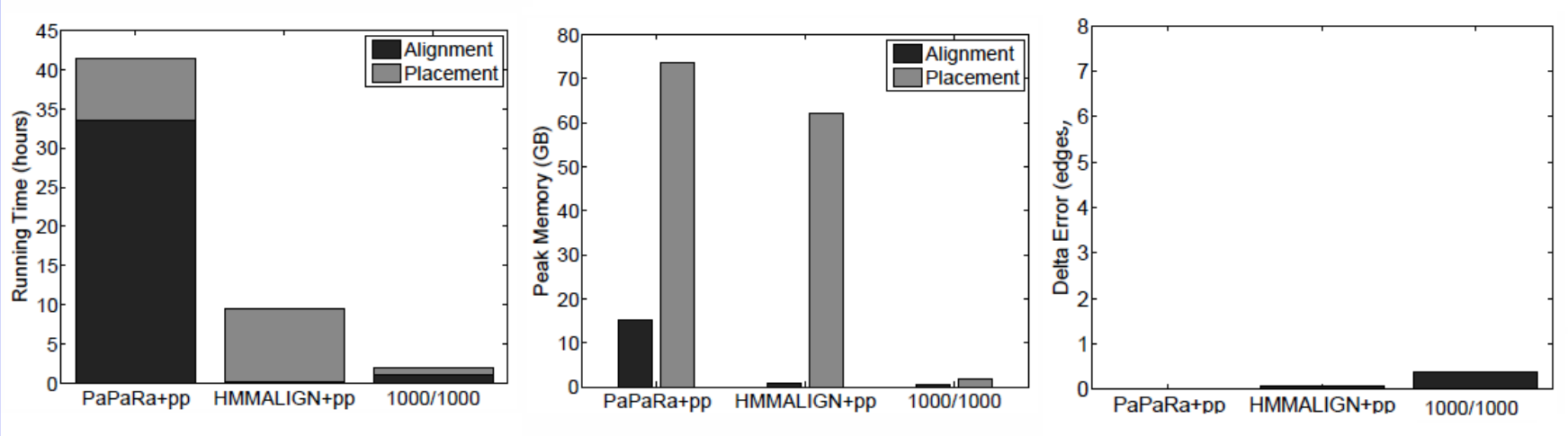
Increasing rate of evolution

SEPP (10%) on Biological Data



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

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

Three “Boosters”

- **SATé**: co-estimation of alignments and trees
- **DACTAL**: large trees without full alignments
- **SEPP**: phylogenetic analysis of fragmentary data

Algorithmic strategies: divide-and-conquer and iteration to improve the accuracy and scalability of a *base method*

Summary

- Standard alignment and phylogeny estimation methods do not provide adequate accuracy on large datasets, and NGS data present novel challenges
- When markers tend to yield poor alignments and trees, develop better methods - don’t throw out the data.

Current Research Projects

Method development:

- Large-scale multiple sequence alignment and phylogeny estimation
- Metagenomics
- Comparative genomics
- Estimating species trees from gene trees
- Supertree methods
- Phylogenetic estimation under statistical models

Dataset analyses (multi-institutional collaborations):

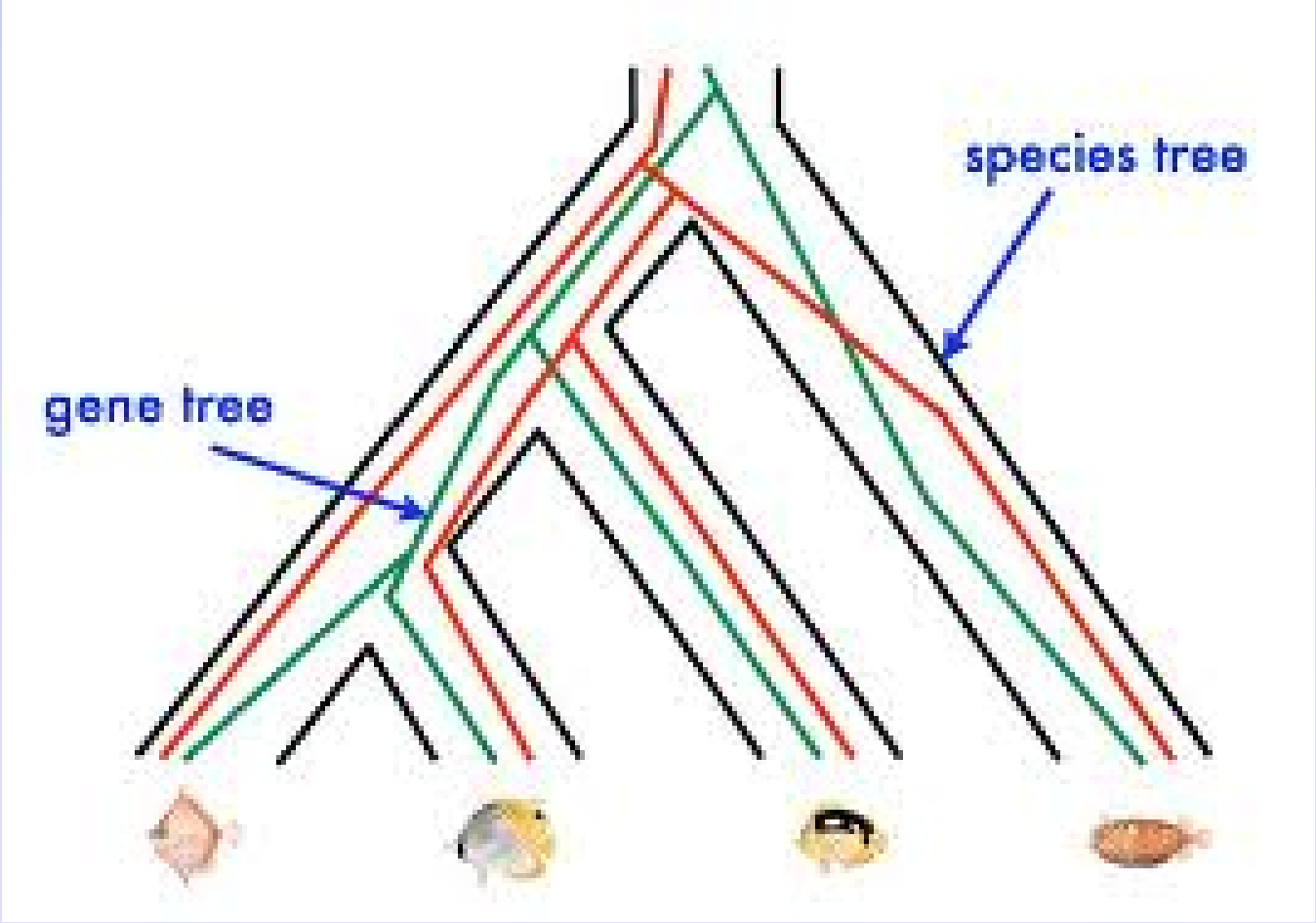
- Avian Phylogeny (and brain evolution)
- Human Microbiome
- Thousand Transcriptome (1KP) Project
- Conifer evolution

Acknowledgments

- Guggenheim Foundation Fellowship, Microsoft Research New England, National Science Foundation: Assembling the Tree of Life (ATOL), ITR, and IGERT grants

- Collaborators:
  - SATé: Kevin Liu, Serita Nelesen, Sindhu Raghavan, and Randy Linder
  - DACTAL: Serita Nelesen, Li-San Wang, and Randy Linder
  - SEPP: Siavash Mirarab and Nam Nguyen

Red gene tree ≠ species tree (green gene tree okay)



Multi-marker species tree estimation

- Species phylogenies are estimated using multiple gene trees. Most methods assume that all gene trees are identical to the species tree.

- This is known to be unrealistic in some situations, due to processes such as
  - Deep Coalescence
  - Gene duplication and loss
  - Horizontal gene transfer

- MDC problem: Given set of gene trees, find a species tree that minimizes the total number of “deep coalescences”.

Yu, Warnow and Nakhleh, 2011

- Previous software for MDC assumed all gene trees are correct, completely resolved, and rooted.
- Our methods allow for error in estimated gene trees.
- We provide exact algorithms and heuristics to find an optimal species tree with respect to a given set of partially resolved, unrooted gene trees, minimizing the total number of deep coalescences.
- Software at http://bioinfo.cs.rice.edu/phylonet/

To appear, RECOMB 2011 and J. Computational Biology, special issue for RECOMB 2011.

Talk about this topic today at 2 PM in OEB.

Markov Model of Site Evolution

Simplest (Jukes-Cantor):

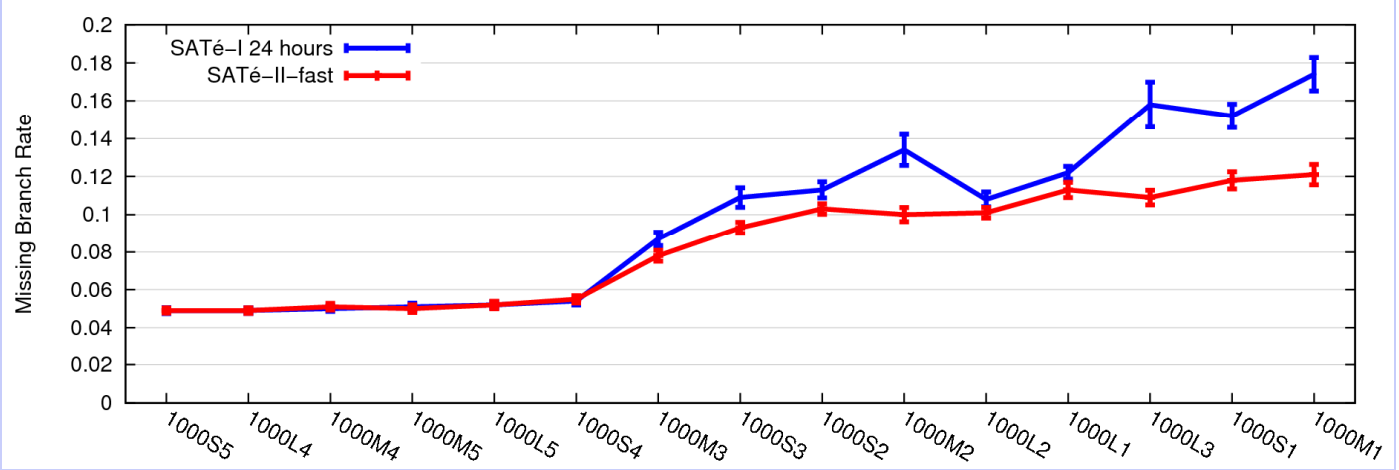
- The model tree T is binary and has substitution probabilities p(e) on each edge e.
- The state at the root is randomly drawn from {A,C,T,G} (nucleotides)
- If a site (position) changes on an edge, it changes with equal probability to each of the remaining states.
- The evolutionary process is Markovian.

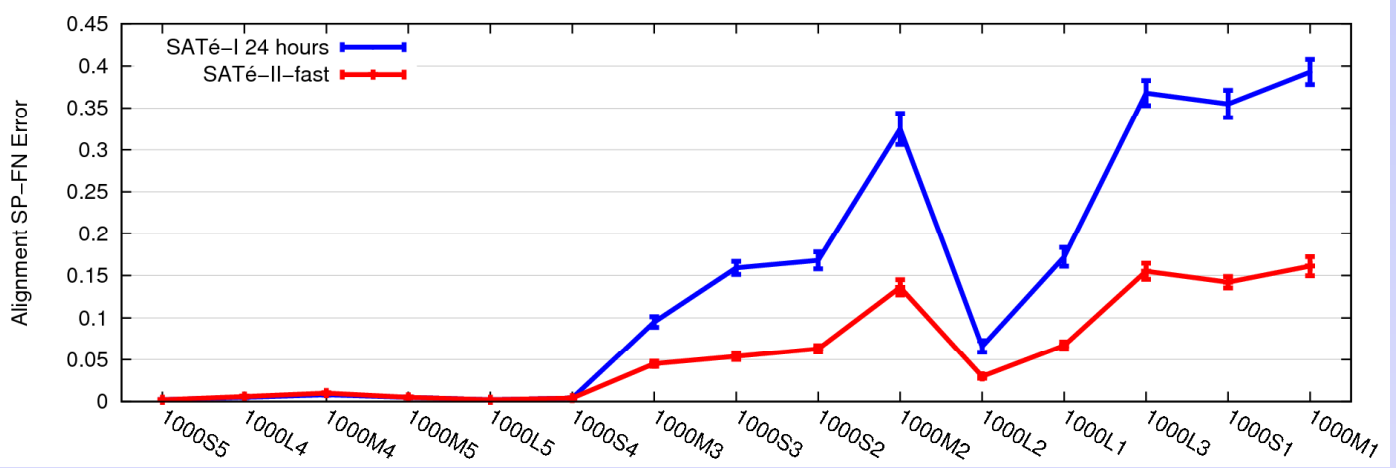
More complex models (such as the General Markov model) are also considered, often with little change to the theory.

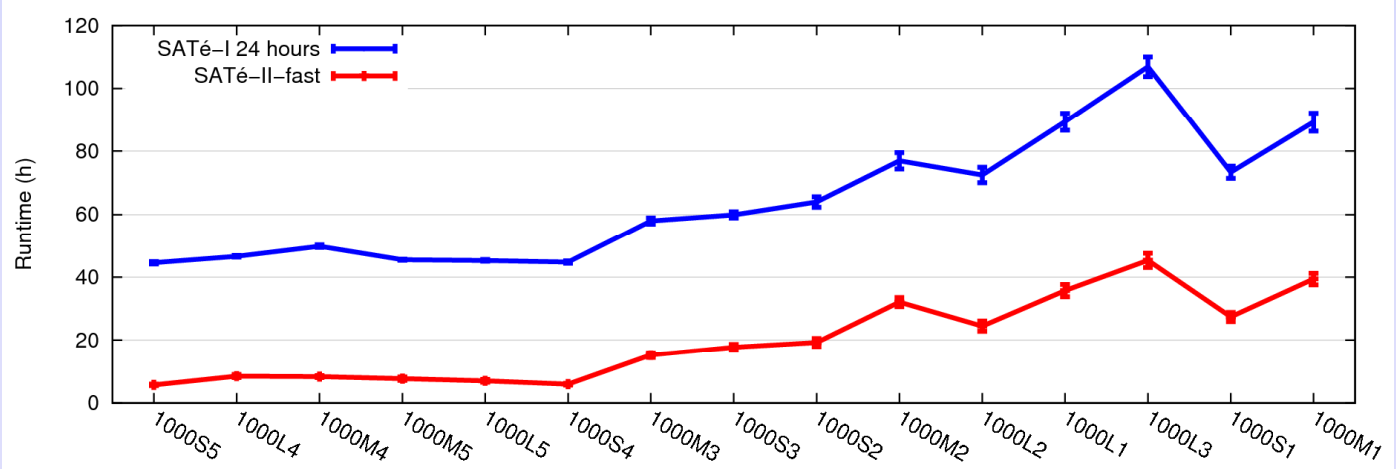
SATé-I vs. SATé-II

SATé-II

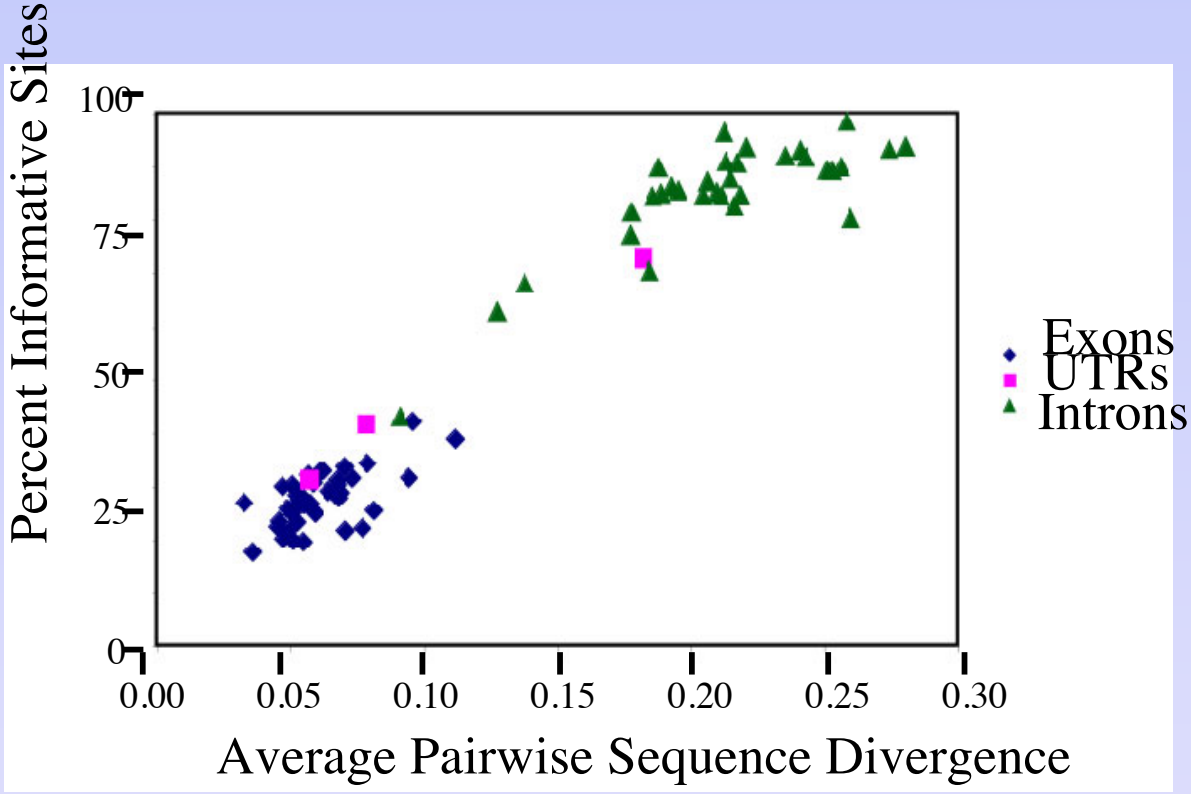
- **Faster** and more accurate than **SATé-I**
- Longer analyses or use of ML to select tree/alignment pair slightly better results







Divergence & Information Content



Analysis and figure provided by Mike Braun
Smithsonian Institution