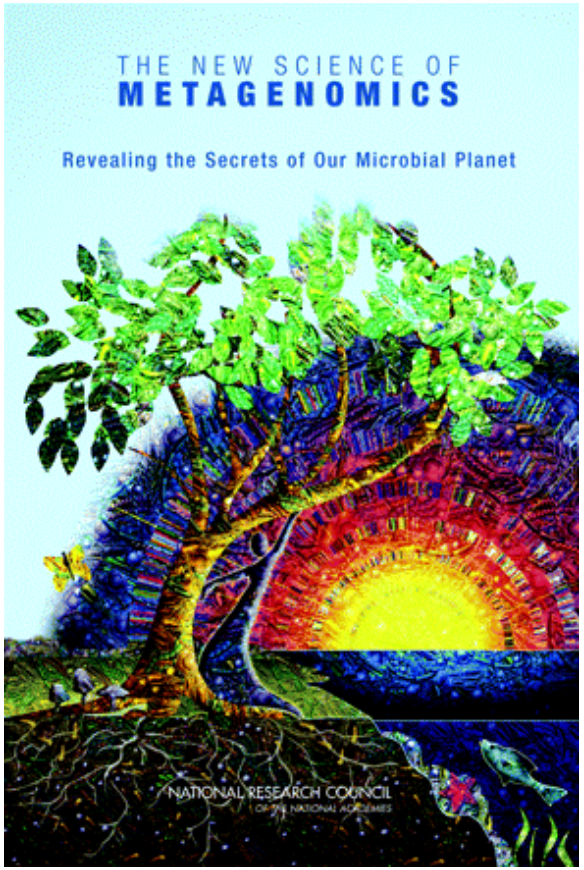
SEPP and TIPP: New Phylogenetic Placement and Taxon Identification Methods for Metagenomic Data

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

Computational Phylogenetics and Metagenomics



Courtesy of the Tree of Life project

Major Challenges

- Many phylogenetic datasets contain hundreds to thousands of species, some with thousands of genes. *Current alignment and tree estimation methods have poor accuracy or cannot run on large datasets, especially if the data are fragmentary.*
- Metagenomic datasets contain millions of short reads or contigs. *Current taxon identification methods have insufficient sensitivity, and high throughput is essential.*

Goals

- High accuracy
- Able to analyze large datasets
- Robust to model violations
- Robust to algorithmic parameters (e.g., starting trees)
- Improved biological analyses

Research Projects

Theory: Phylogenetic estimation under statistical models

Method development:

- "Absolute fast converging" methods
- Very large-scale multiple sequence alignment and phylogeny estimation
- Estimating species trees and networks from gene trees
- Supertree methods
- Comparative genomics (genome rearrangement phylogenetics)
- Metagenomic taxon identification
- Alignment and Phylogenetic Placement of NGS data

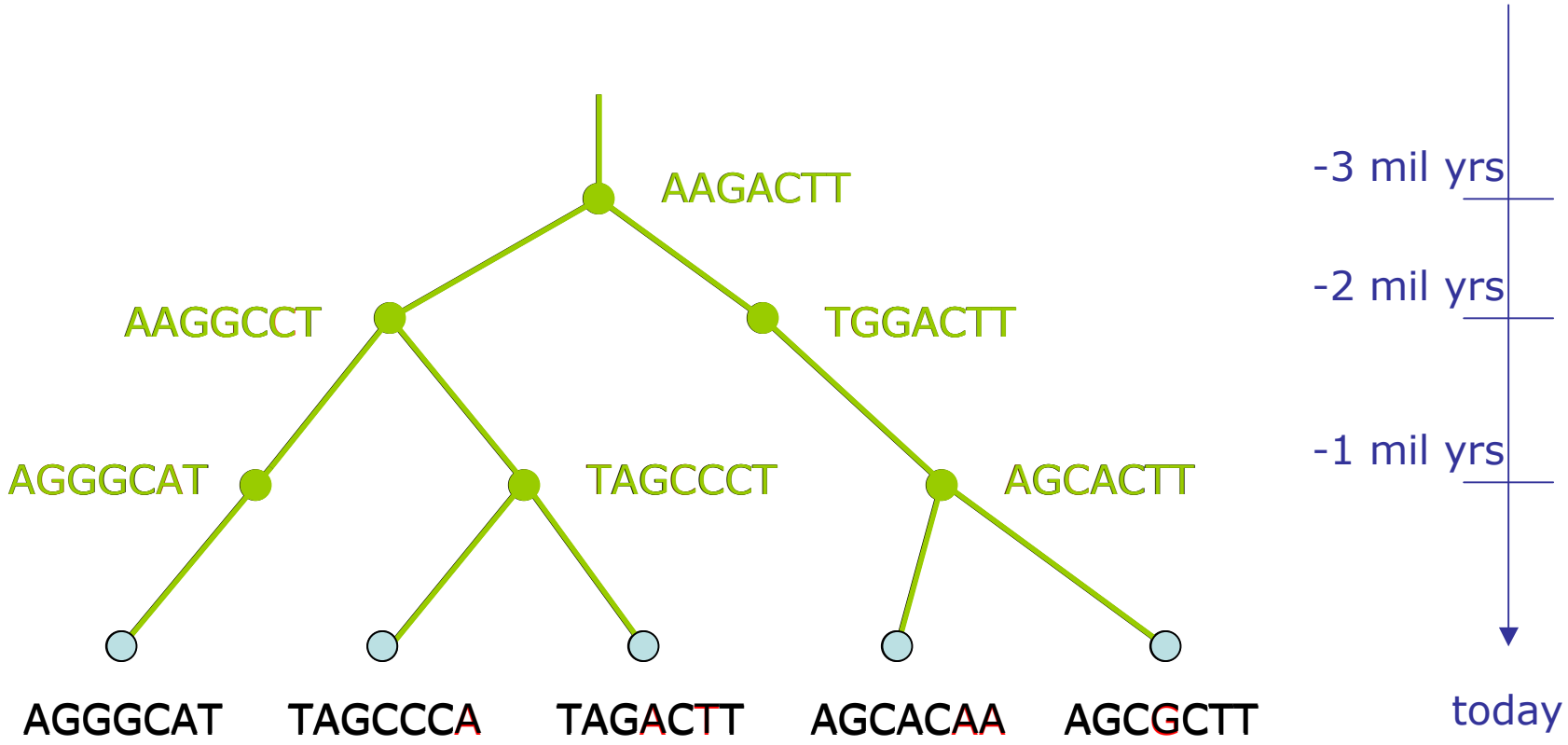
Dataset analyses

- Avian Phylogeny: 50 species and 8000+ genes
- Thousand Transcriptome (1KP) Project: 1000 species and 1000 genes
- Chloroplast genomics

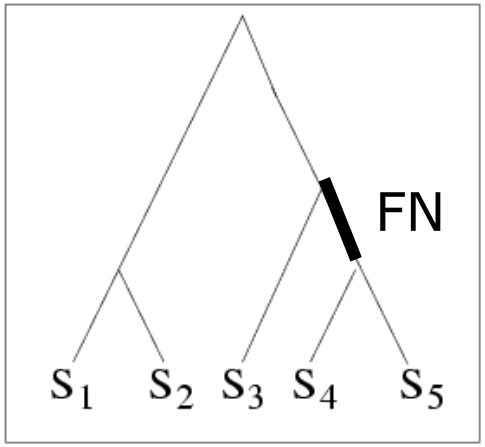
Samples of My Research

- Absolute Fast Converging Methods (SODA 2001, ISMB 2001, TCS 1999, RSA 1999, ICALP 1997)
- SATé (Co-estimation of alignments and trees), Science 2009 and Systematic Biology (in press)
- DACTAL (almost alignment-free estimation of trees), ISMB and Bioinformatics 2012)
- SEPP/TIPP (Phylogenetic placement and taxon identification of short reads for metagenomic analysis), 2012 Pacific Symposium on Biocomputing (SEPP) and in preparation (TIPP)

DNA Sequence Evolution



Quantifying Error



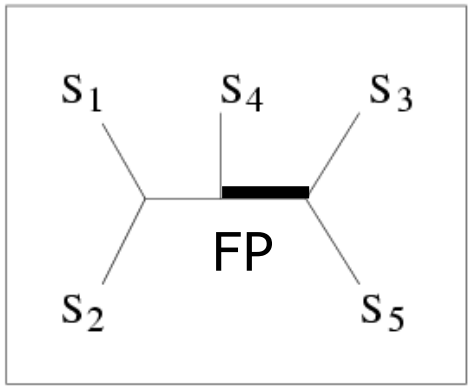
TRUE TREE

S_1	ACAATTAGAAC
S_2	ACCCTTAGAAC
S_3	ACCATTCCAAC
S_4	ACCAGACCAAC
S_5	ACCAGACCGGA

DNA SEQUENCES

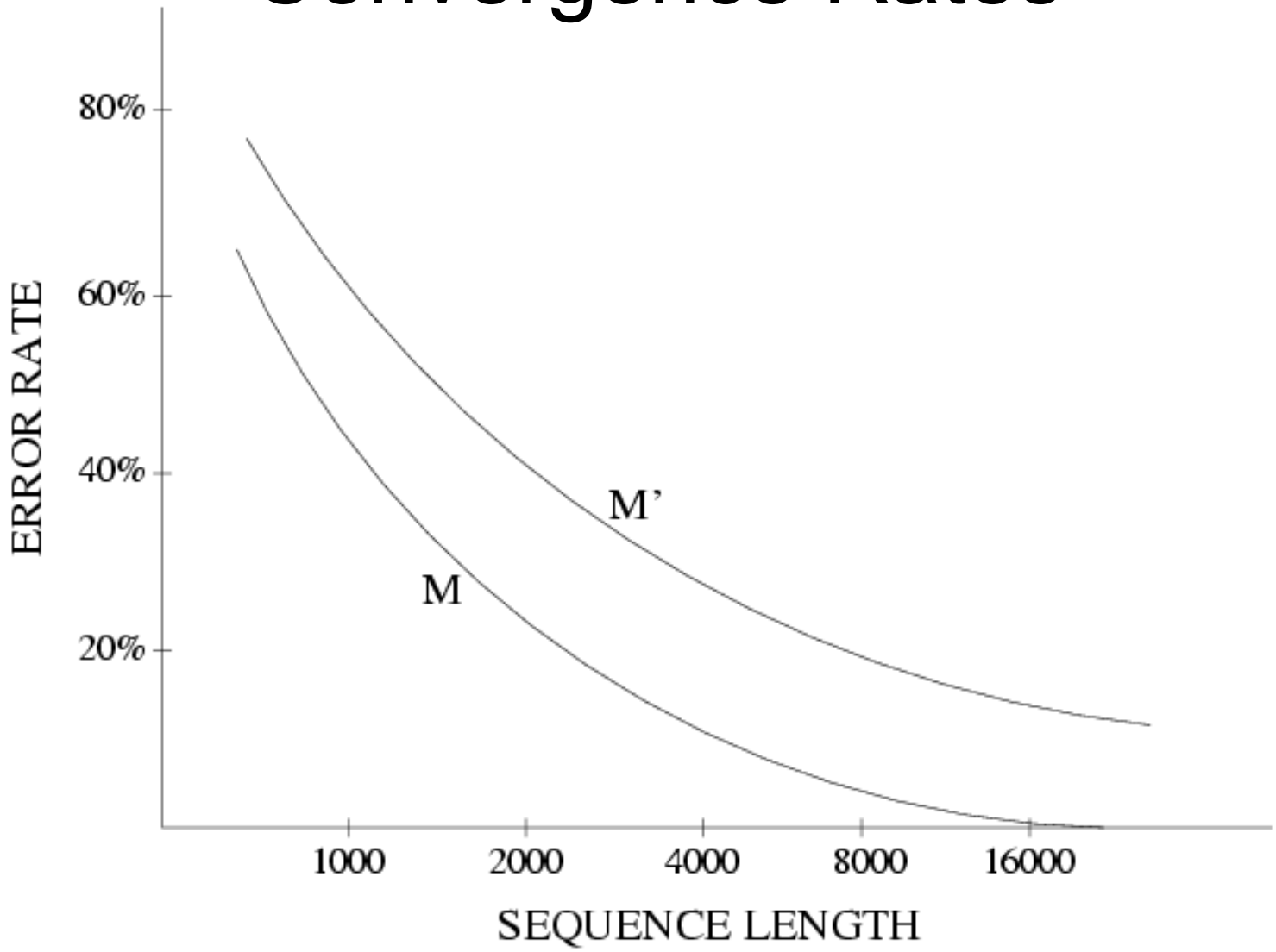
FN: false negative
(missing edge)
FP: false positive
(incorrect edge)

50% error rate

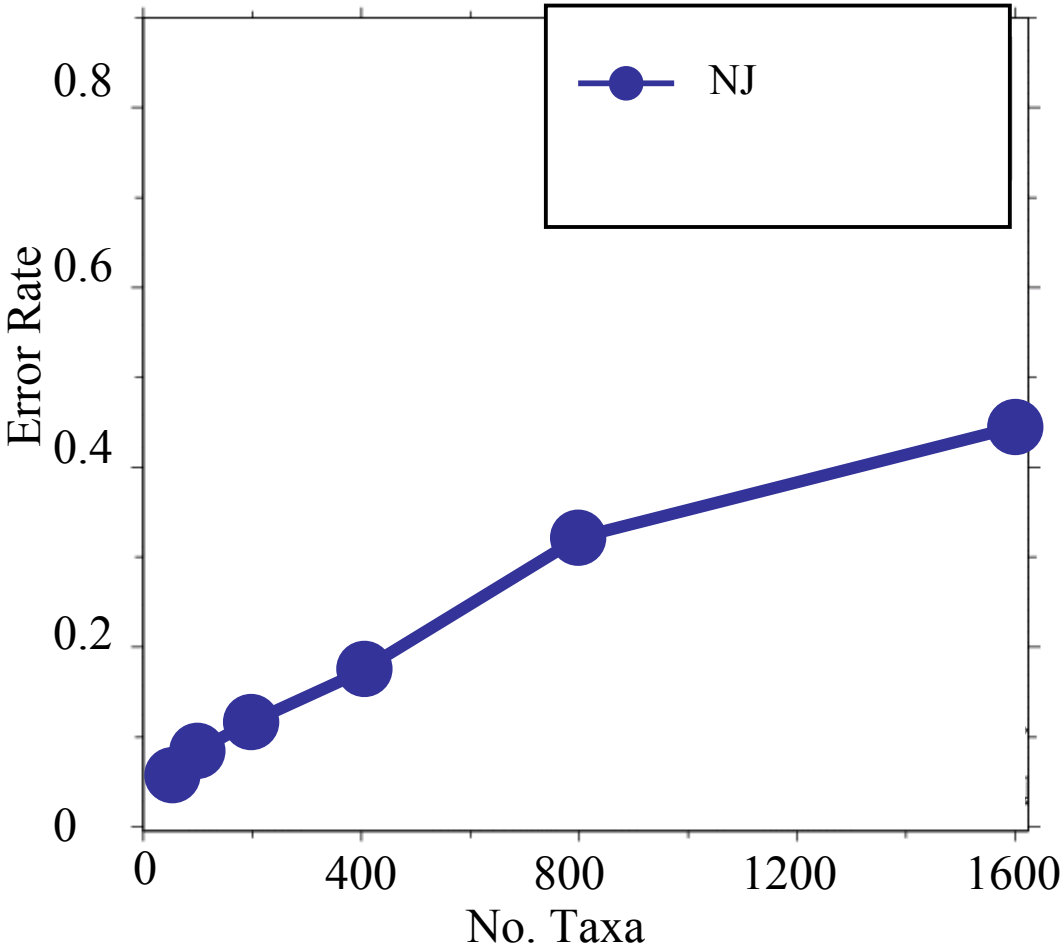


INFERRED TREE

Statistical Consistency and Convergence Rates

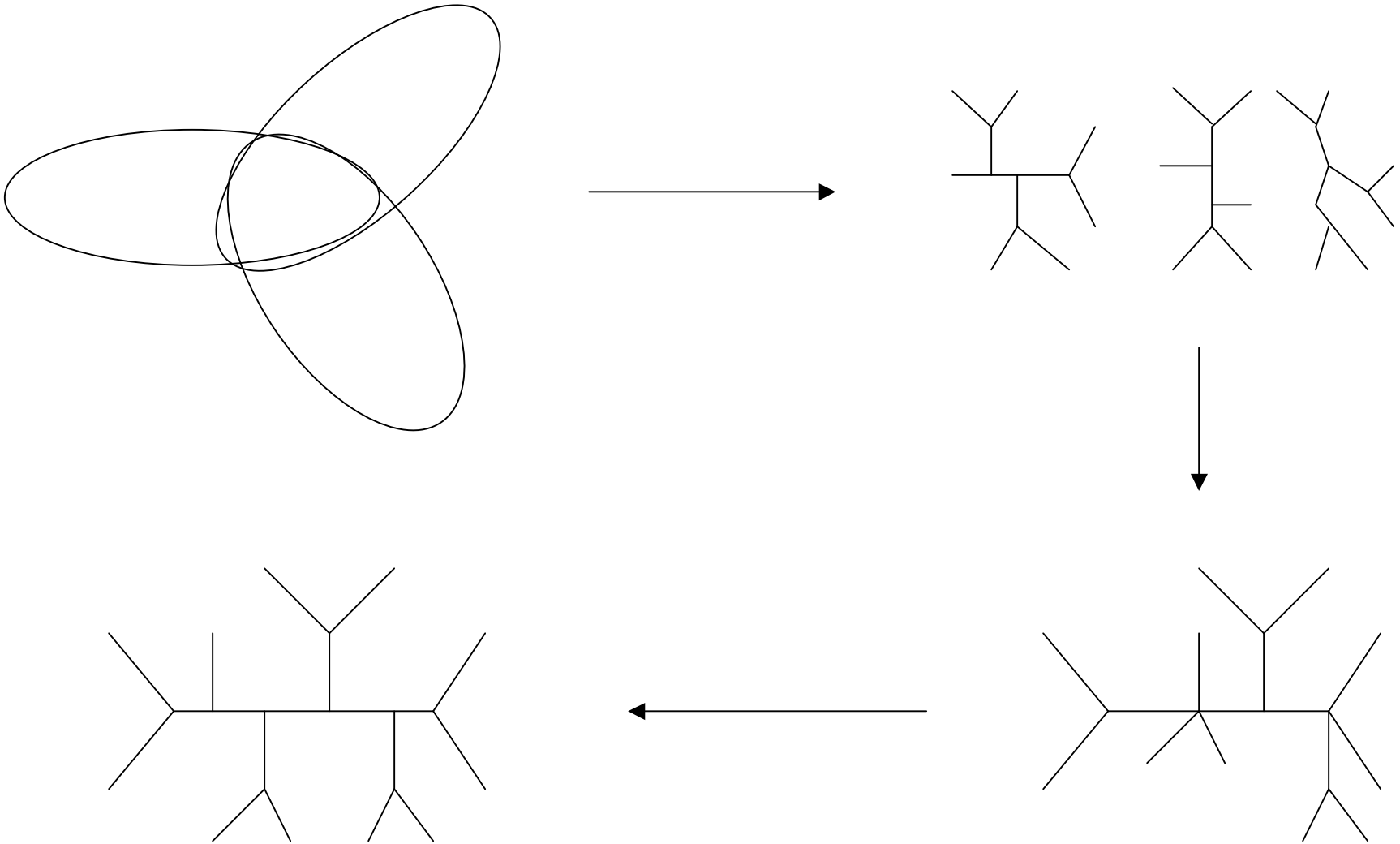


Neighbor Joining has Poor Performance on Large Diameter Trees *[Nakhleh et al. ISMB 2001]*



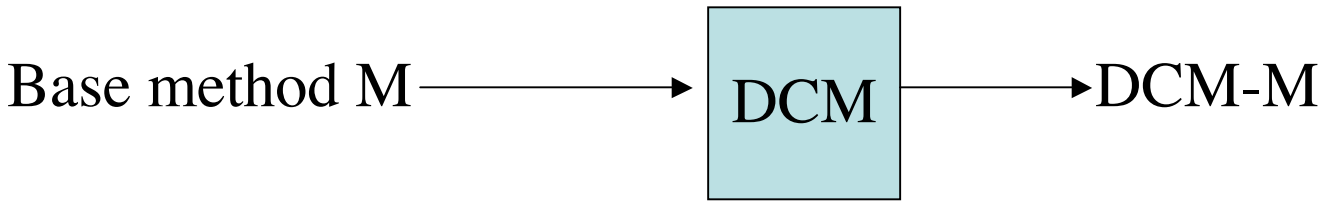
Exponential sequence length requirement for Neighbor Joining (Lacey and Chang, 2006)

Disk-Covering Methods (DCMs)

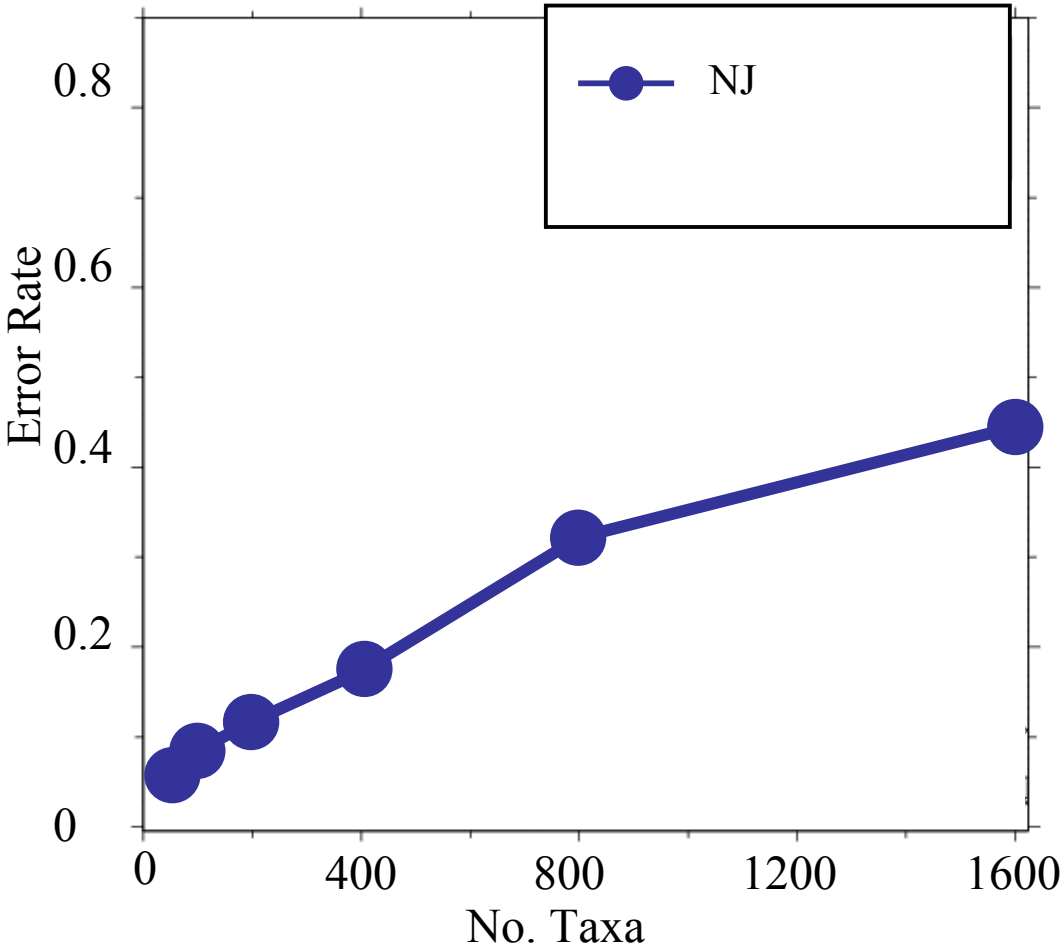


Disk-Covering Methods (DCMs)

- DCMs “boost” the performance of phylogeny reconstruction methods.



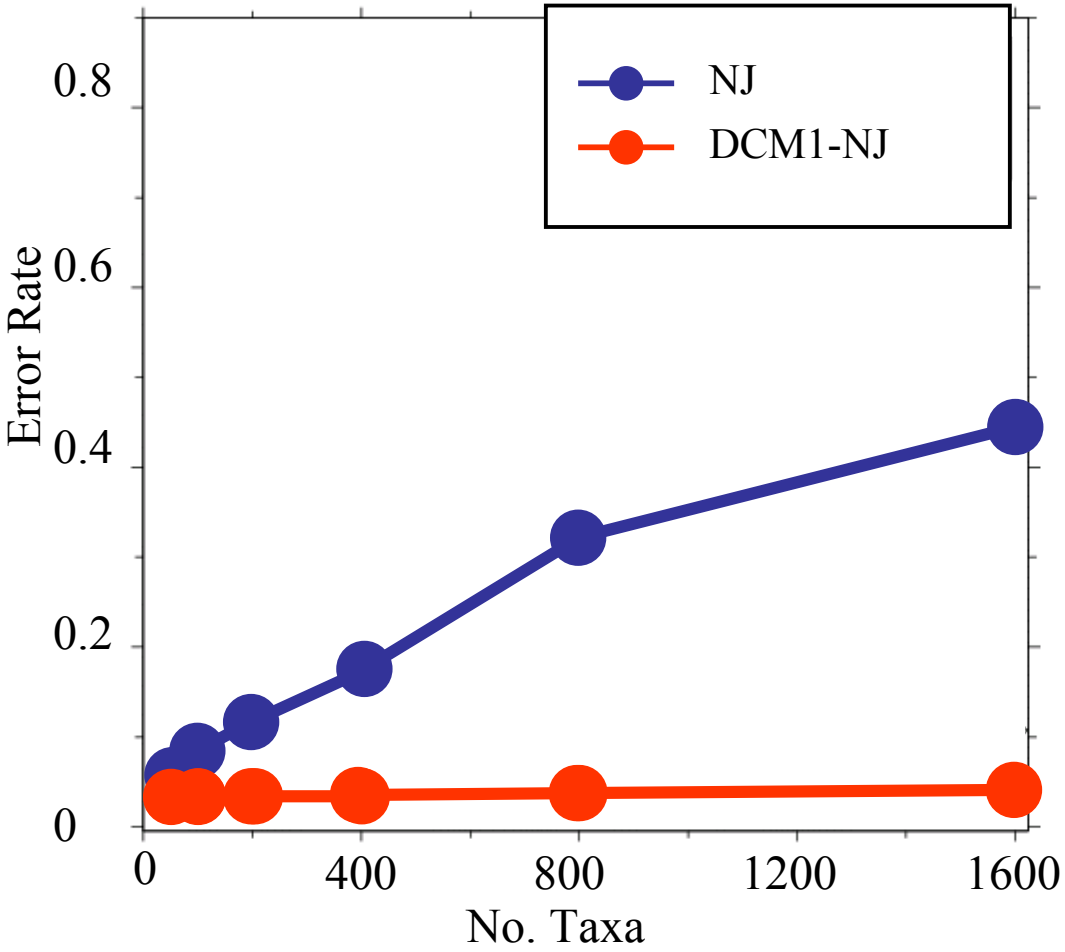
Neighbor Joining has Poor Performance on Large Diameter Trees *[Nakhleh et al. ISMB 2001]*



Exponential sequence length requirement for Neighbor Joining (Lacey and Chang, 2006)

DCM1-boosting Distance-based Methods

[Nakhleh et al. ISMB 2001]



Theorem:
DCM1-NJ converges to the true tree from polynomial length sequences (Warnow et al., SODA 2001)

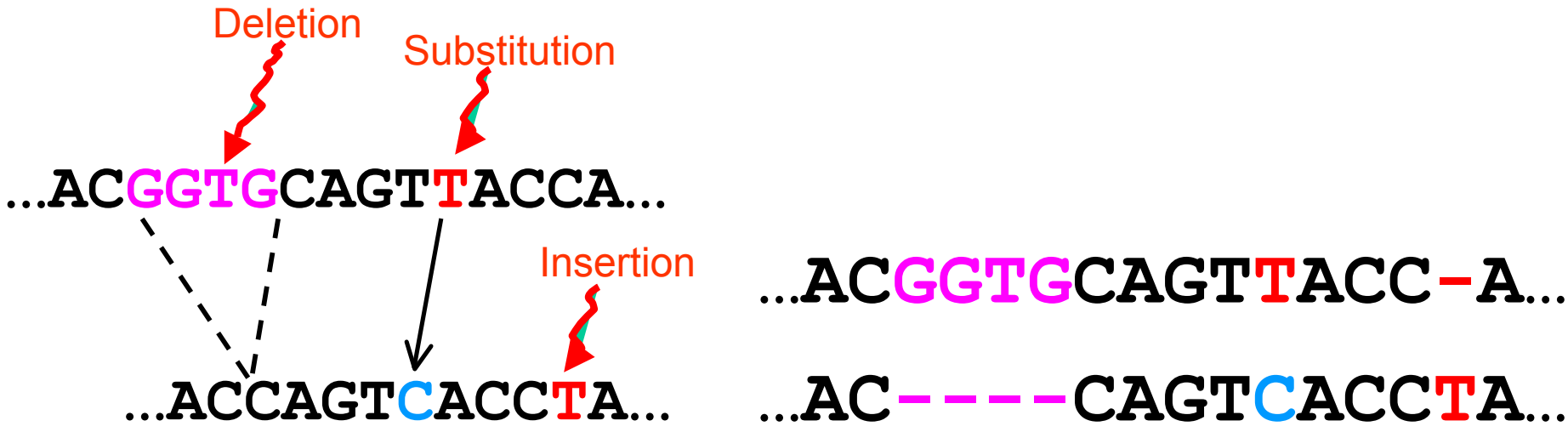
Part II: SATé

Simultaneous Alignment and Tree Estimation
(for nucleotide or amino-acid analysis)

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

Liu et al., Systematic Biology (in press)

Public software distribution (open source) through the University of Kansas (Mark Holder)



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

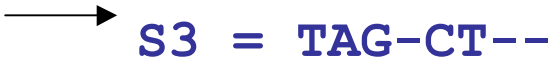
```
S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC
S3 = TAGCTGACCGC
S4 = TCACGACCGACA
```



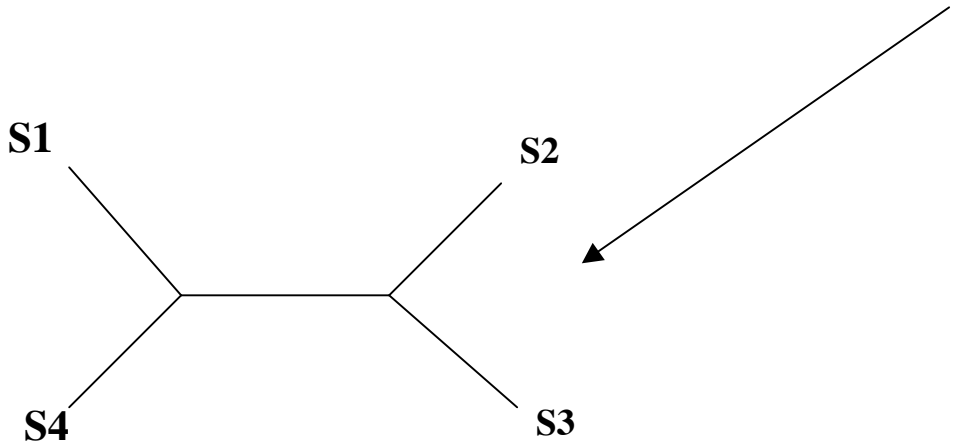
```
S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAC--GACCGC--
S3 = TAG-CT-------GACCGC--
S4 = -------TCAC--GACCGACA
```

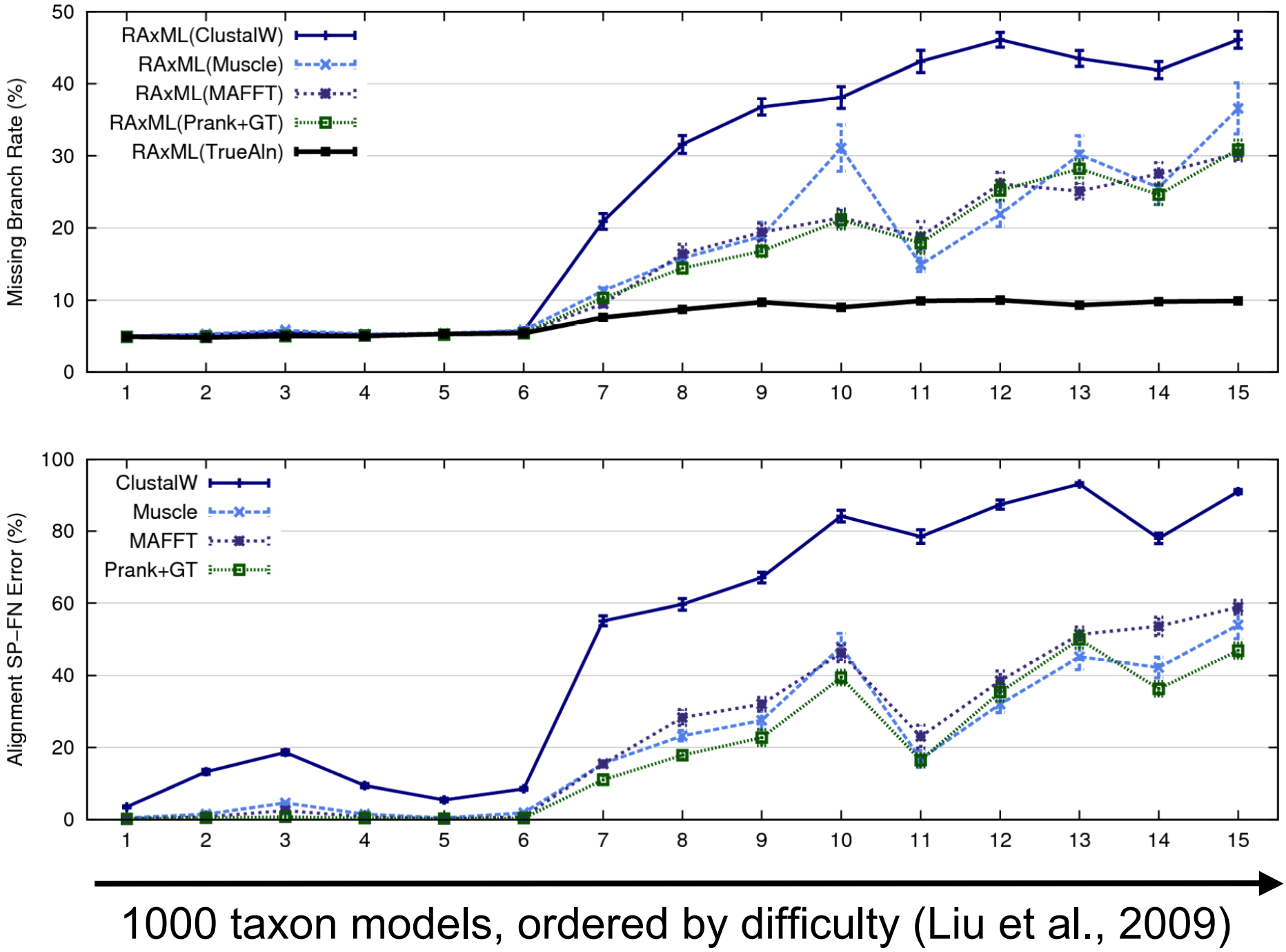
Phase 2: Construct Tree

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



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




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)

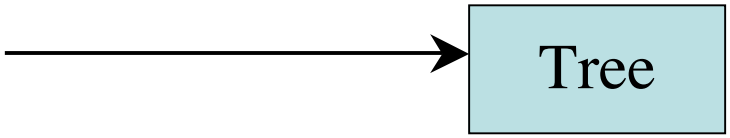
Problems

- Large datasets with high rates of evolution are hard to align accurately, and phylogeny estimation methods produce poor trees when alignments are poor.
- Many phylogeny estimation methods have poor accuracy on large datasets (even if given correct alignments)
- *Potentially useful genes are often discarded* if they are difficult to align.

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

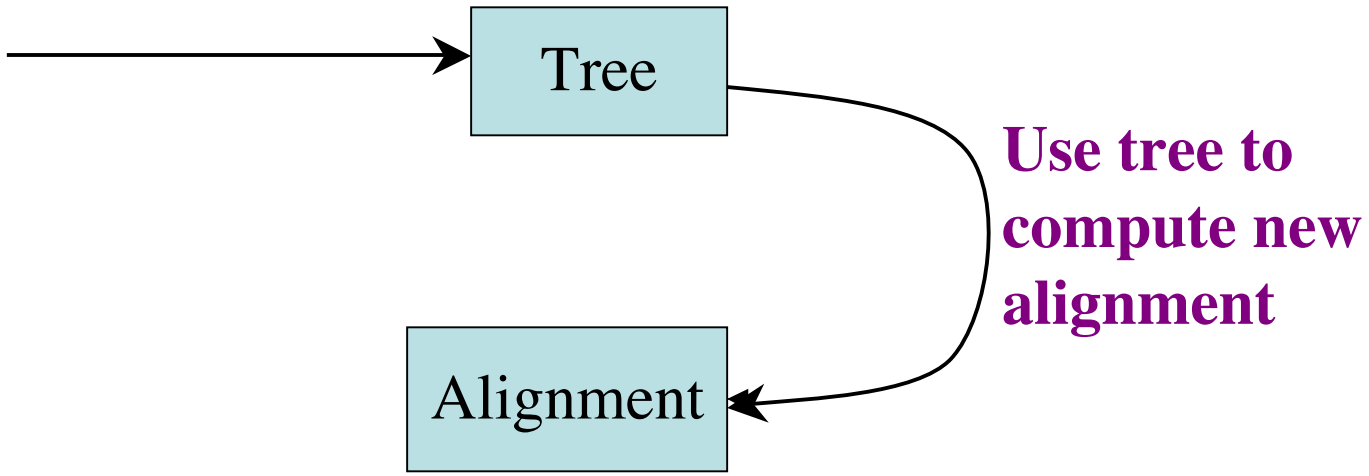
SATé Algorithm

Obtain initial alignment
and estimated ML tree



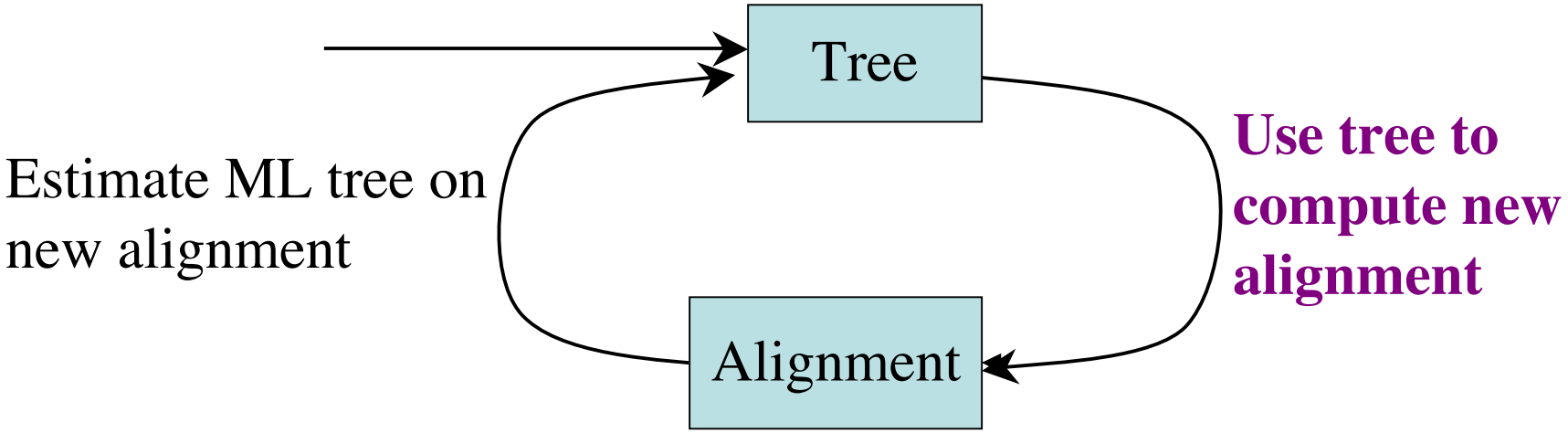
SATé Algorithm

Obtain initial alignment
and estimated ML tree

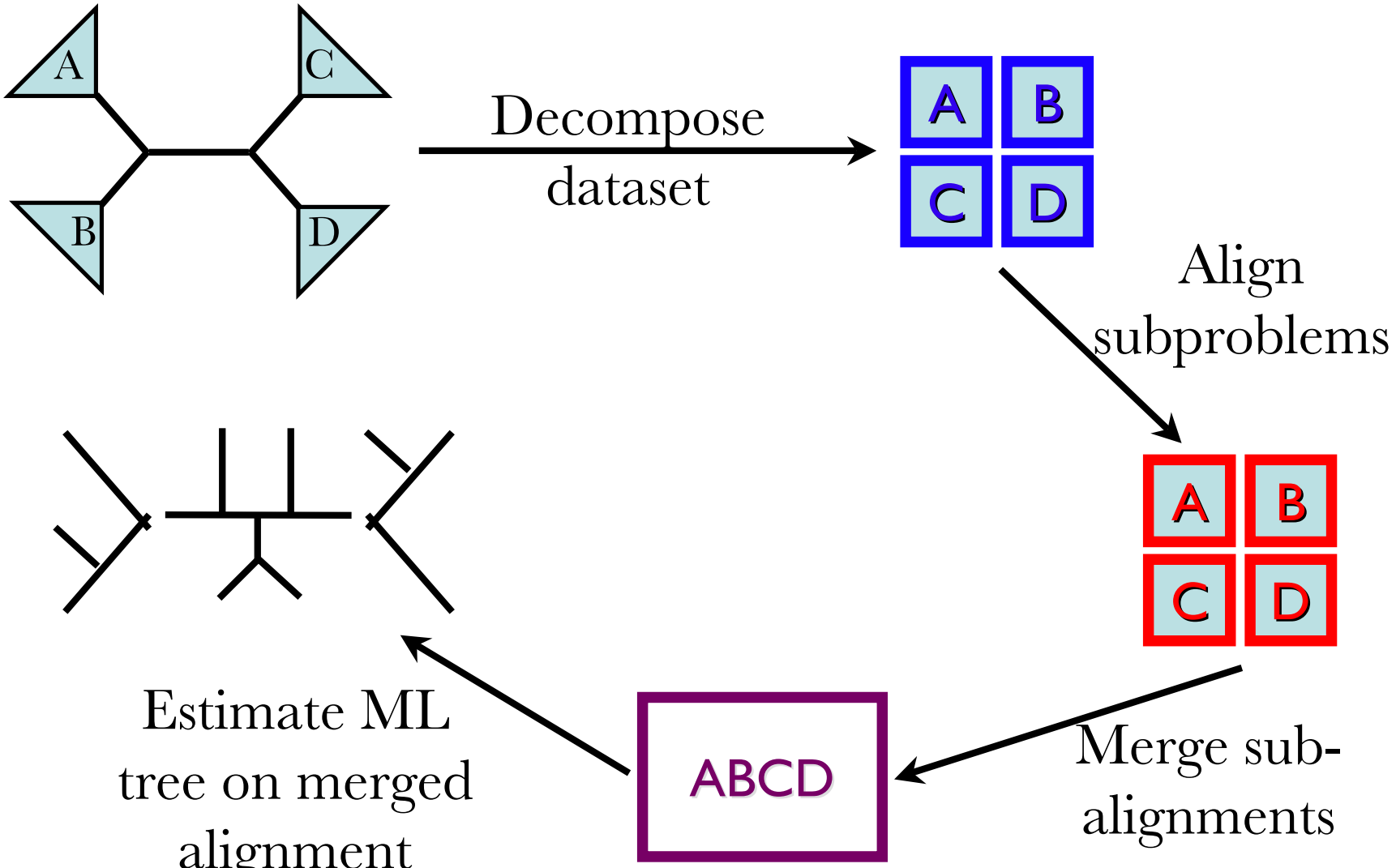


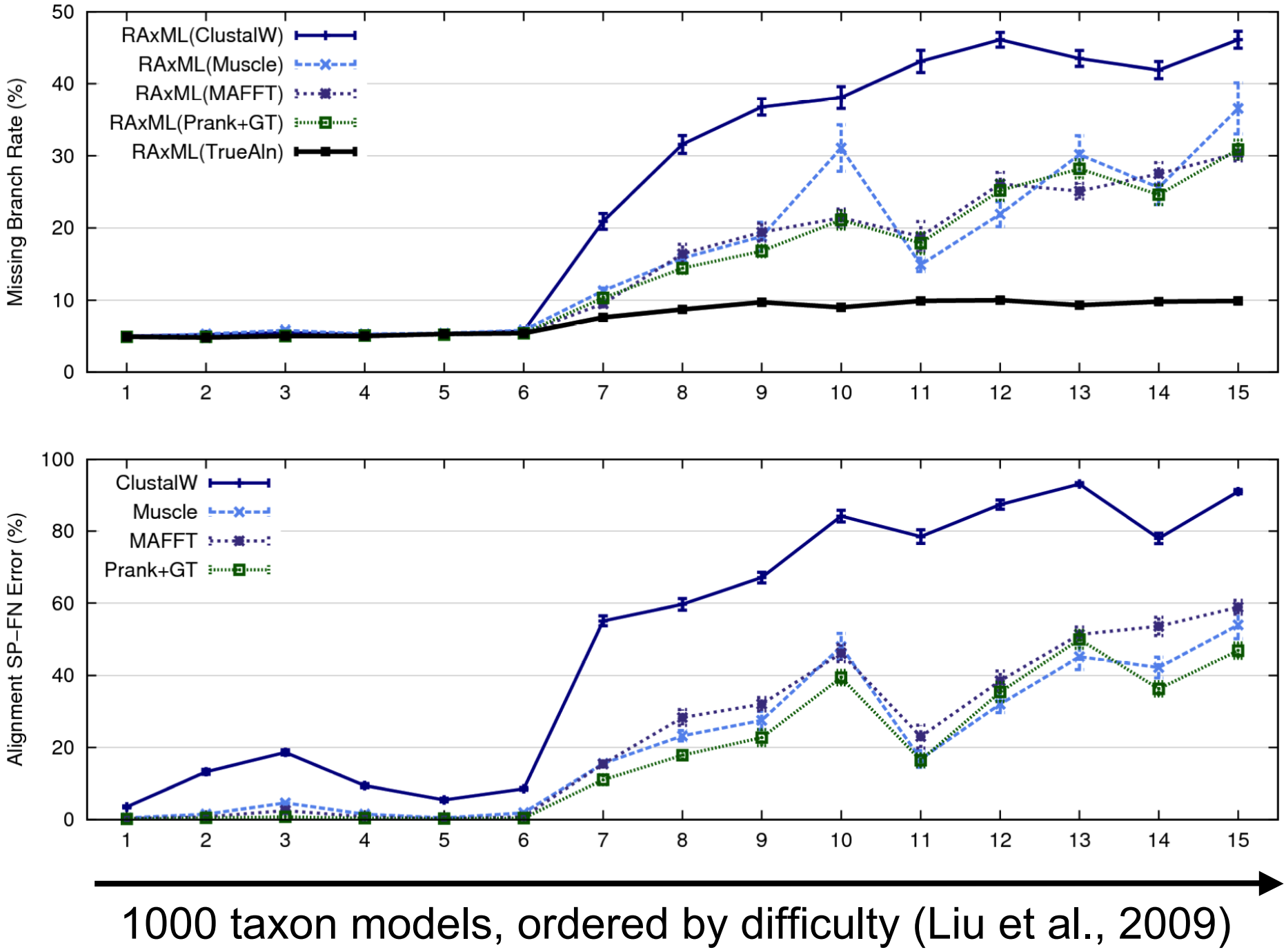
SATé Algorithm

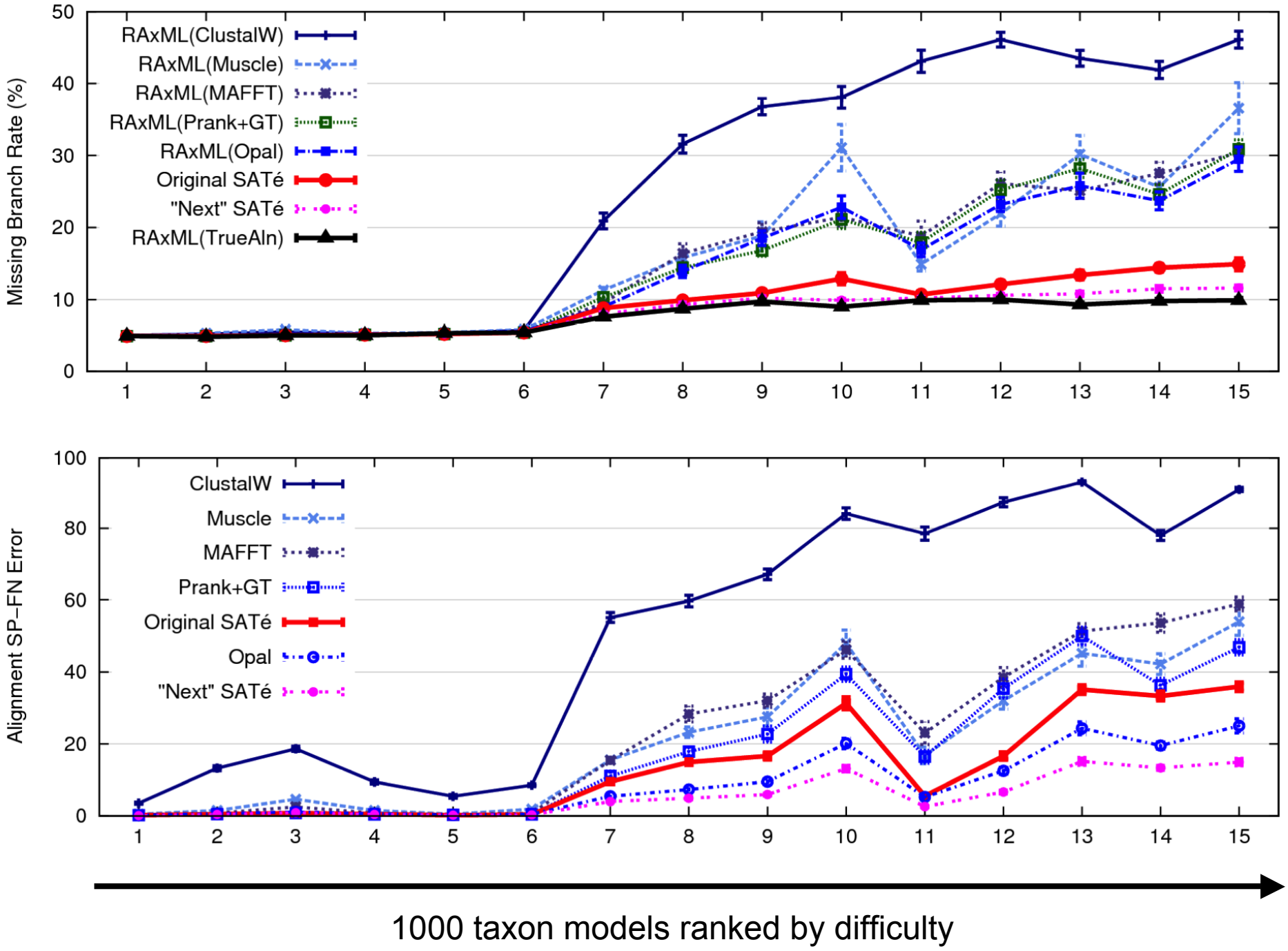
Obtain initial alignment
and estimated ML tree



Re-aligning on a 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)

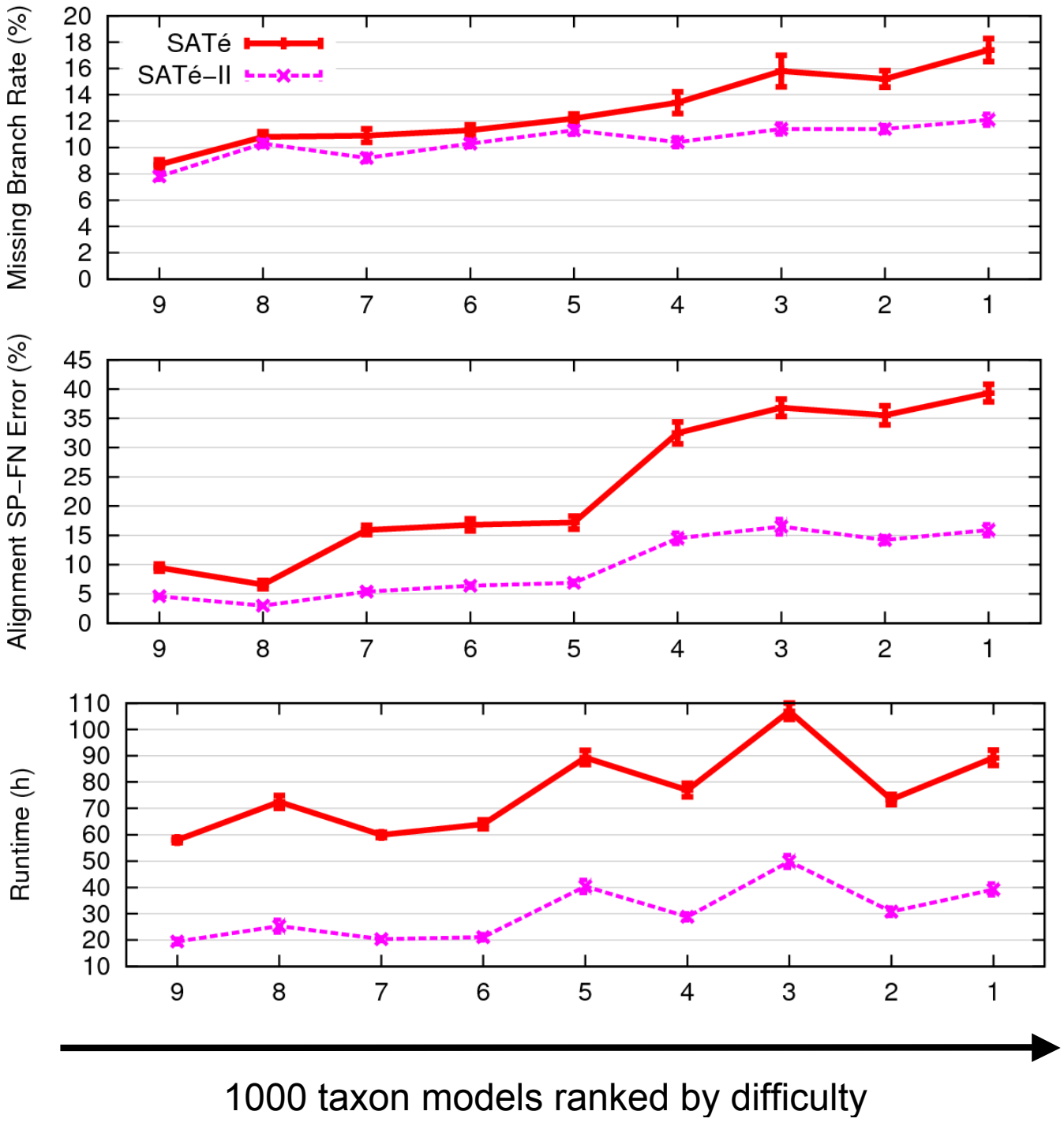

Missing Branch Rate (%)
RAxML(ClustalW)
RAxML(Muscle)
RAxML(MAFFT)
RAxML(Prank+GT)
RAxML(Opal)
Original SATé
"Next" SATé
RAxML(TrueAln)
Alignment SP–FN Error
ClustalW
Muscle
MAFFT
Prank+GT
Original SATé
Opal
"Next" SATé
1000 taxon models ranked by difficulty

Results

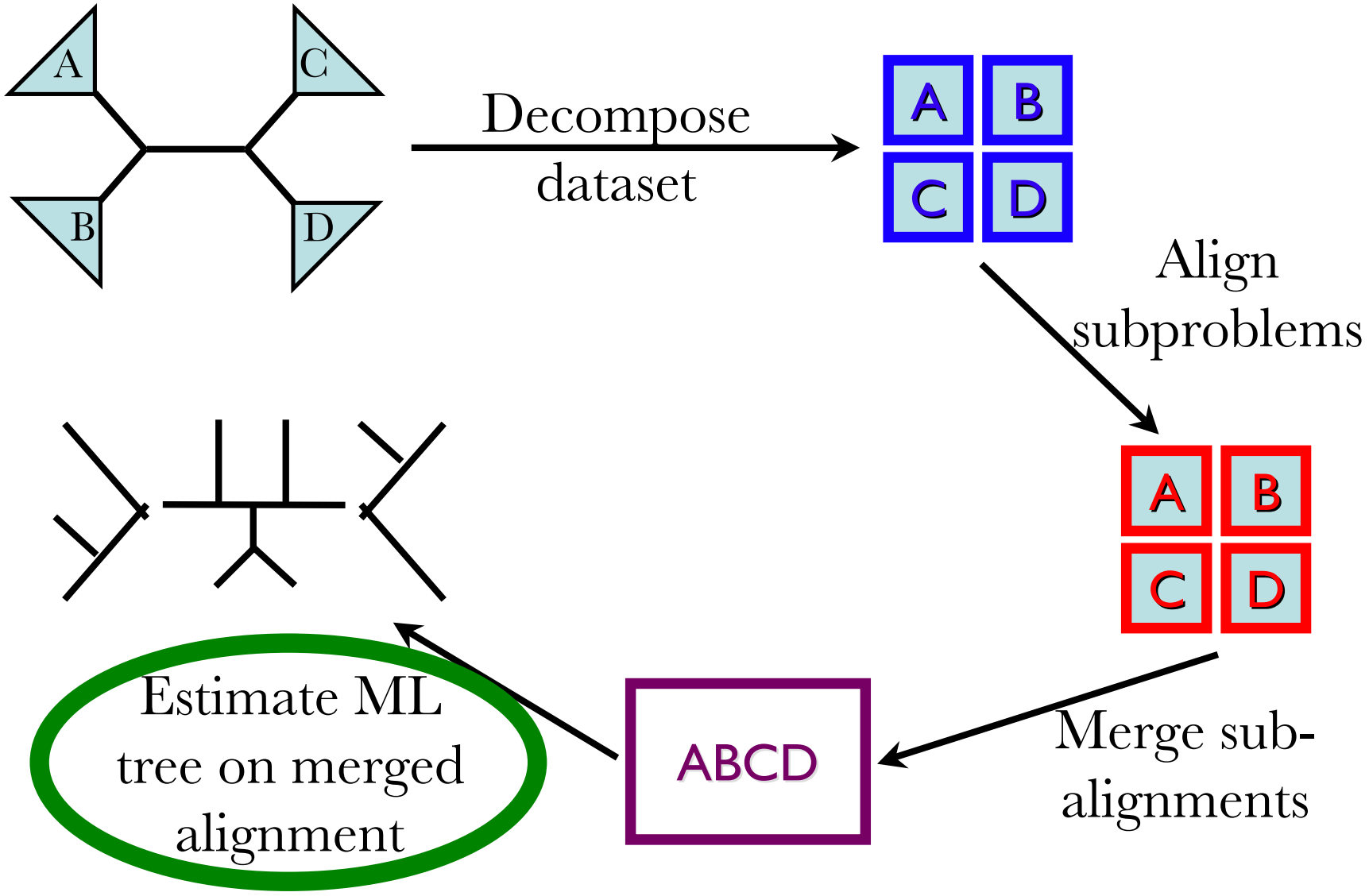
SATé-I

vs.

SATe-II



Limitations



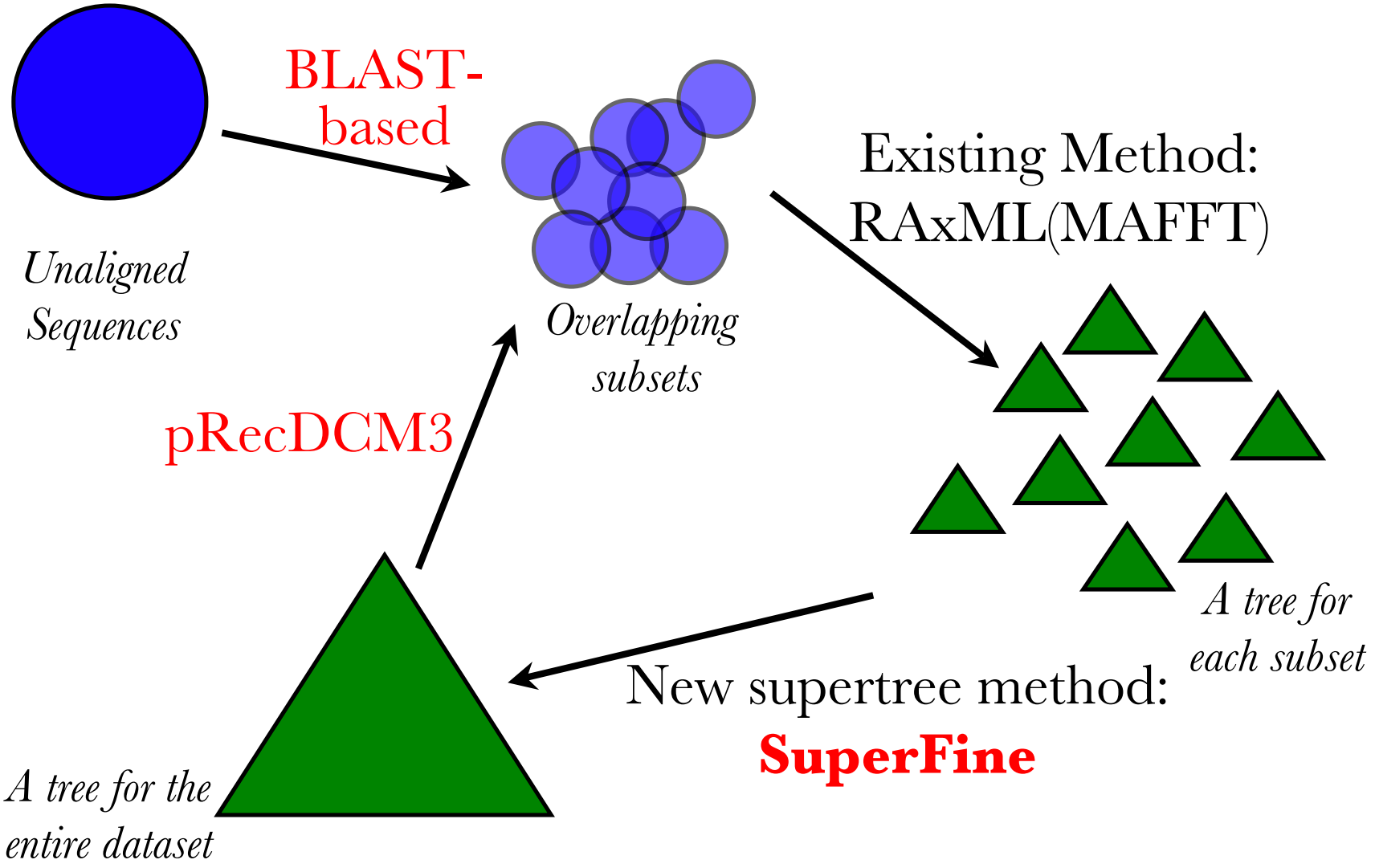
Part III: DACTAL

(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, In Press, ISMB 2012 and Bioinformatics 2012

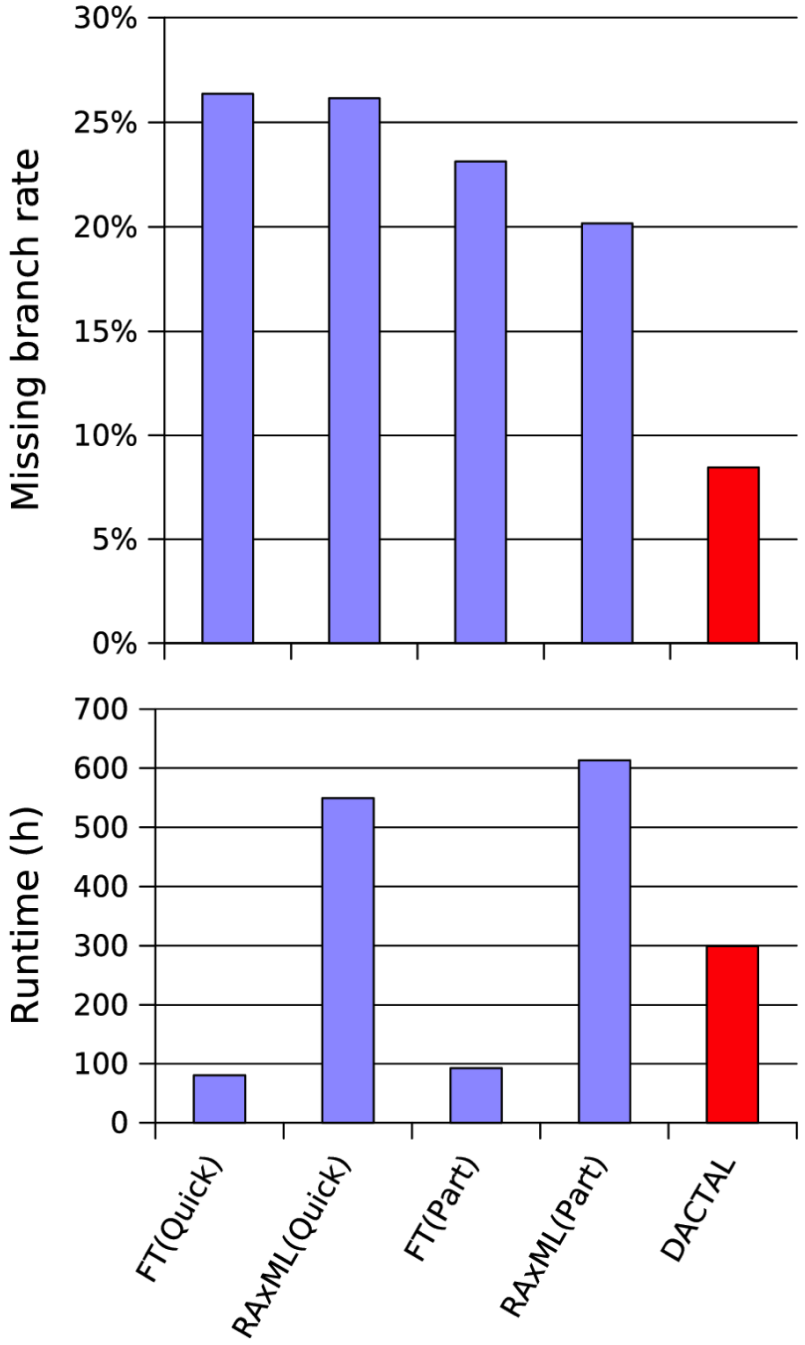
DACTAL


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

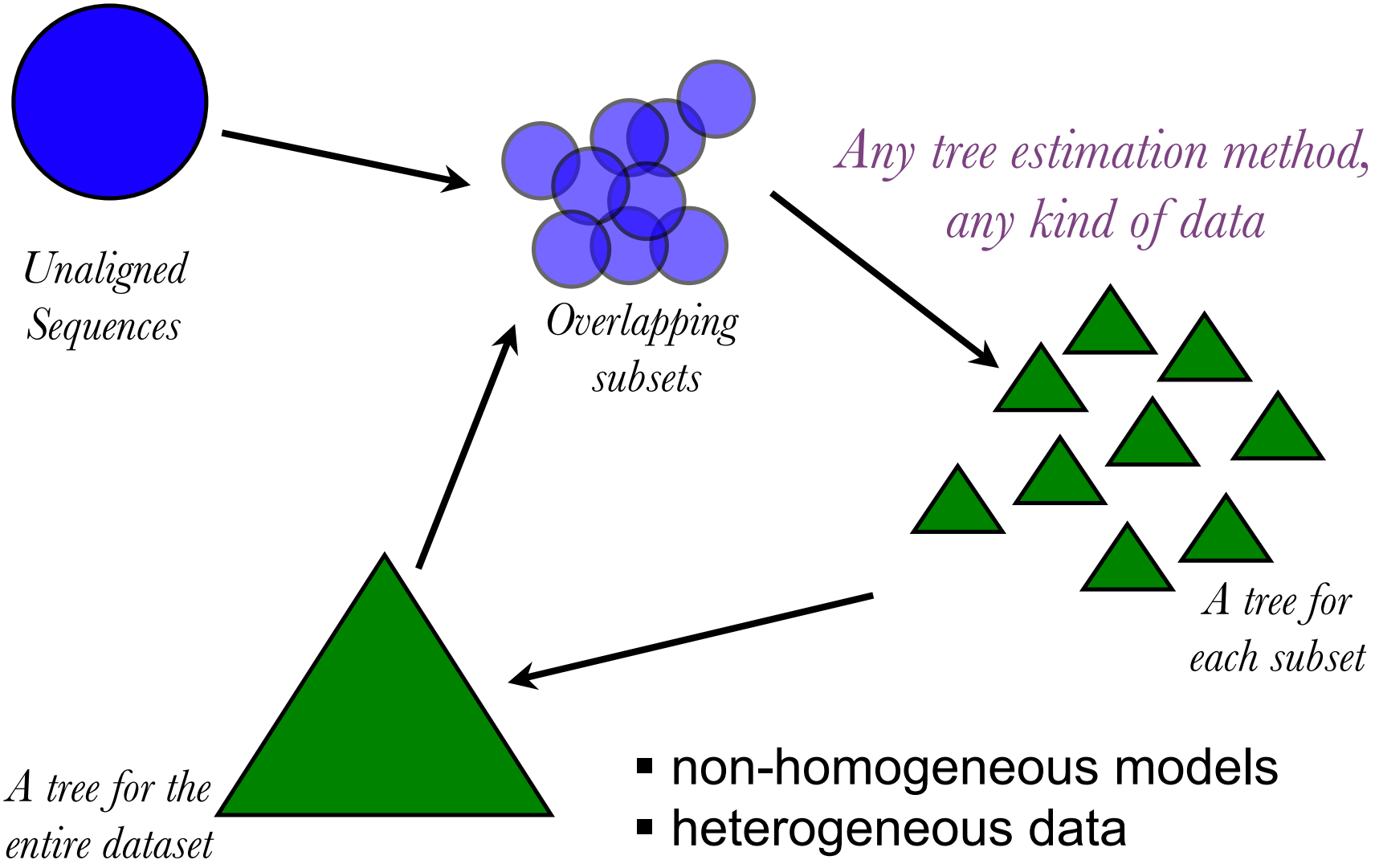
DACTAL: Better results than 2-phase methods

- Three 16S datasets from Gutell's database (CRW) 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 is Flexible



- non-homogeneous models
- heterogeneous data

Part IV: SEPP/TIPP, analysis of NGS and metagenomic data

- Fragmentary data (e.g., short reads):
  - How to align? How to insert into trees?
- Unknown taxa
  - How to identify the species, genus, family, etc?

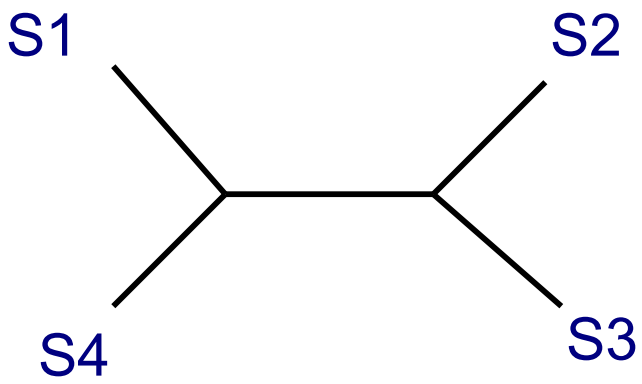
Phylogenetic Placement

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

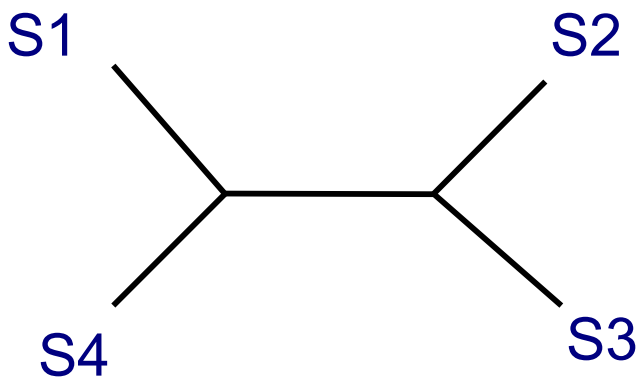
Align Sequence

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



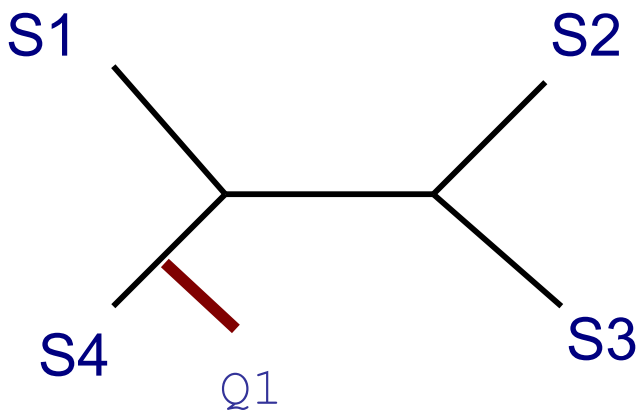
Align Sequence

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

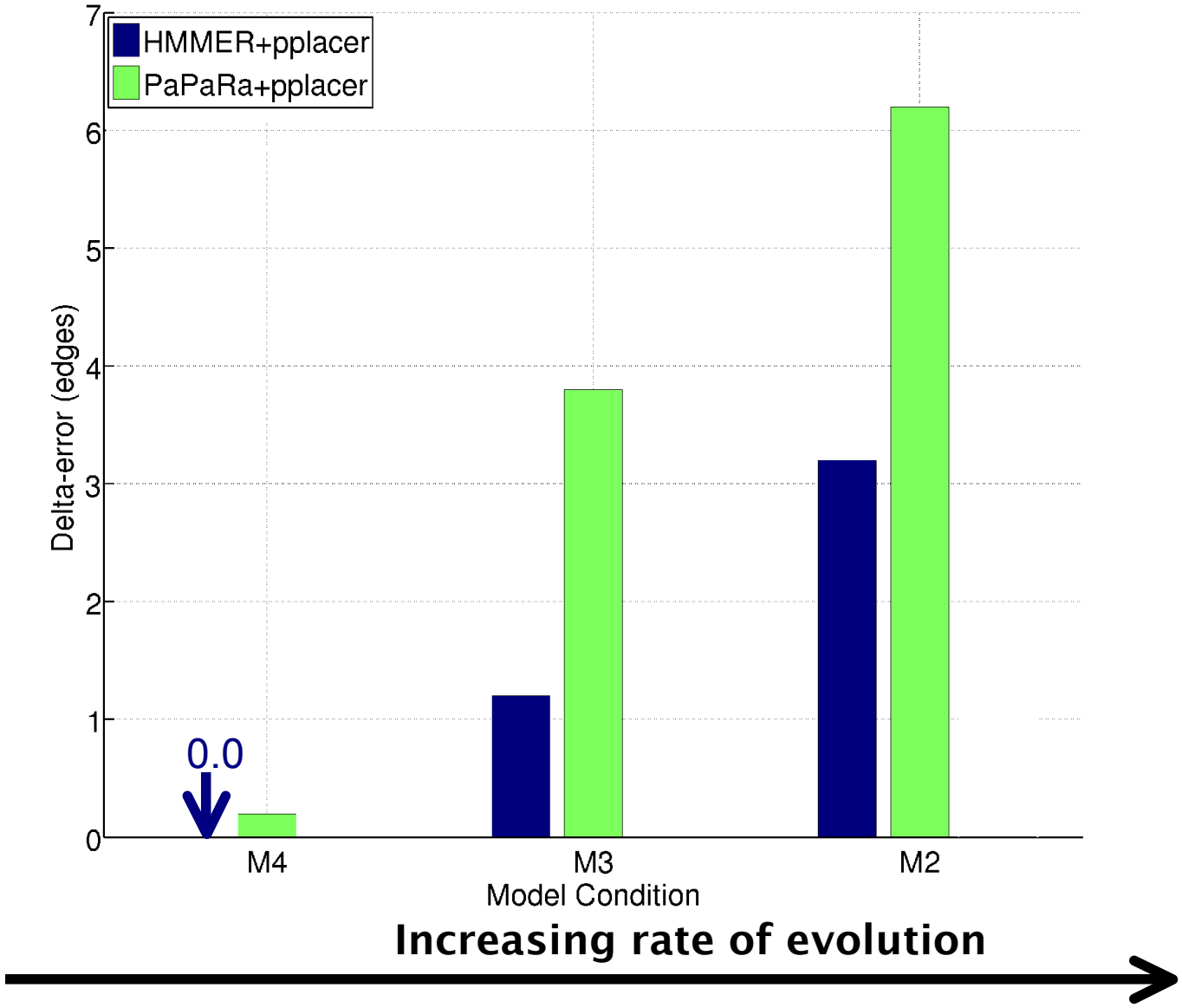


Phylogenetic Placement

- 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

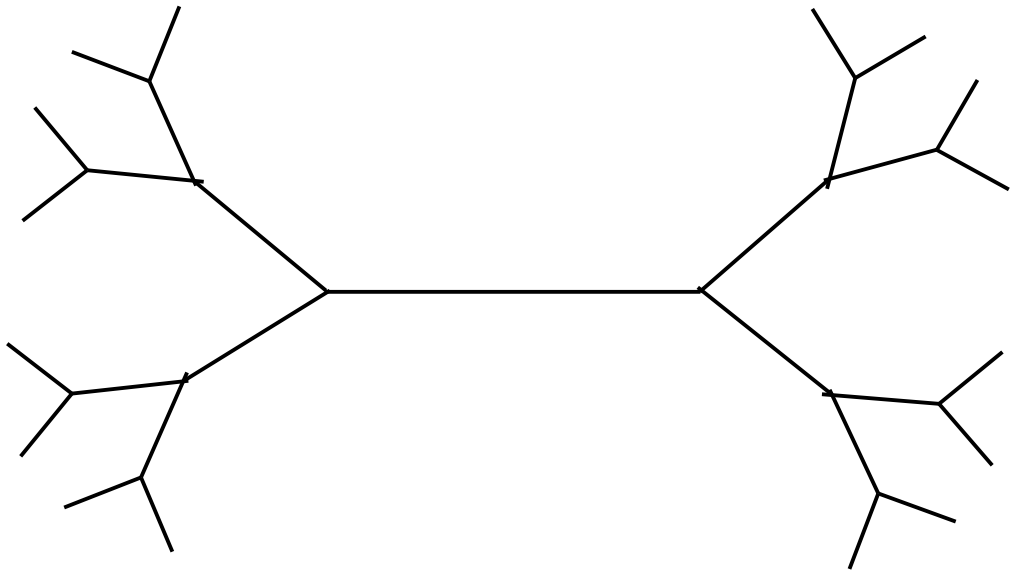
HMMER vs. PaPaRa



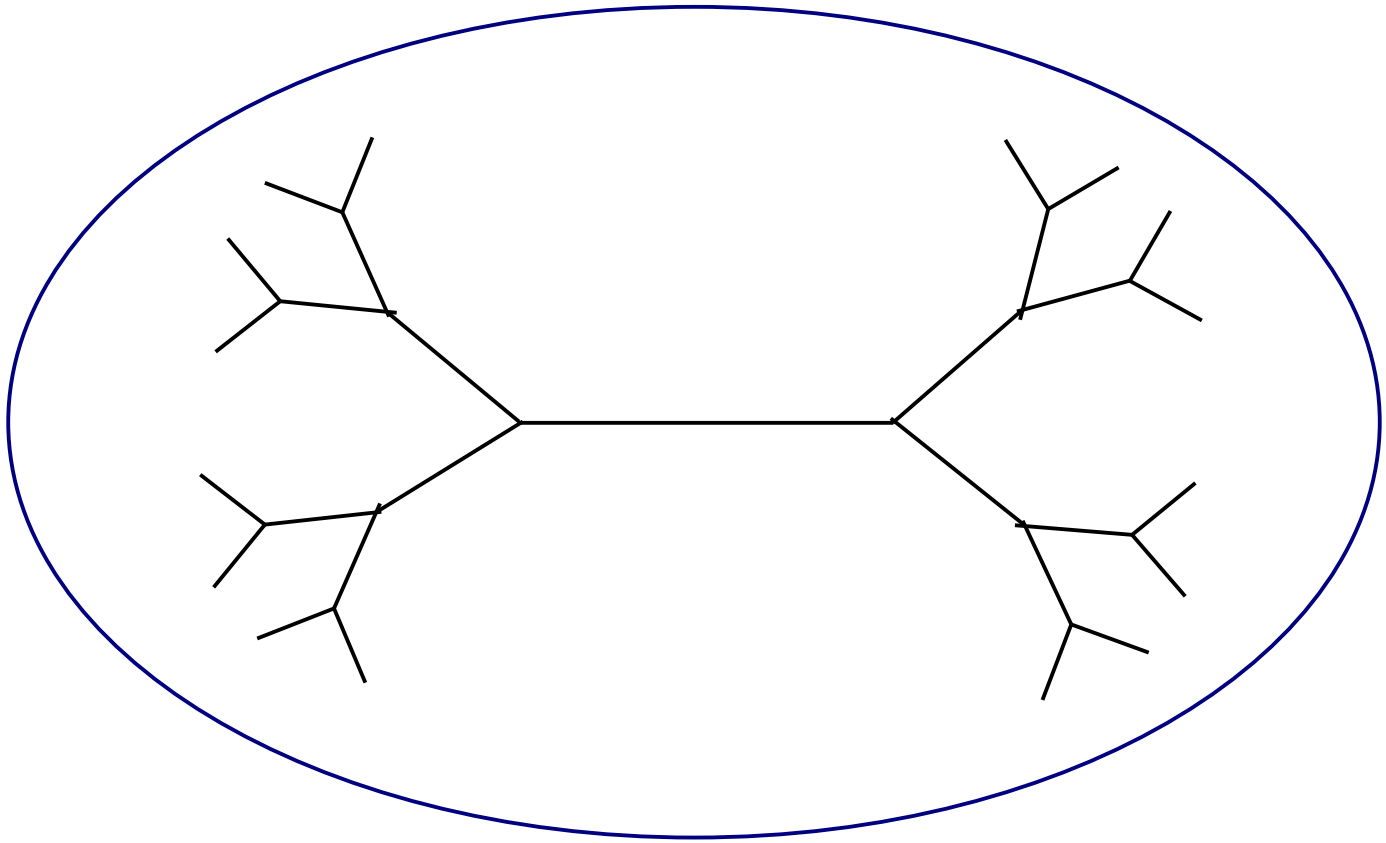
SEPP

- Key insight: HMMs are not very good at modelling MSAs on large, divergent datasets.
- Approach: insert fragments into taxonomy using estimated alignment of full-length sequences, and multiple HMMs (on different subsets of taxa).

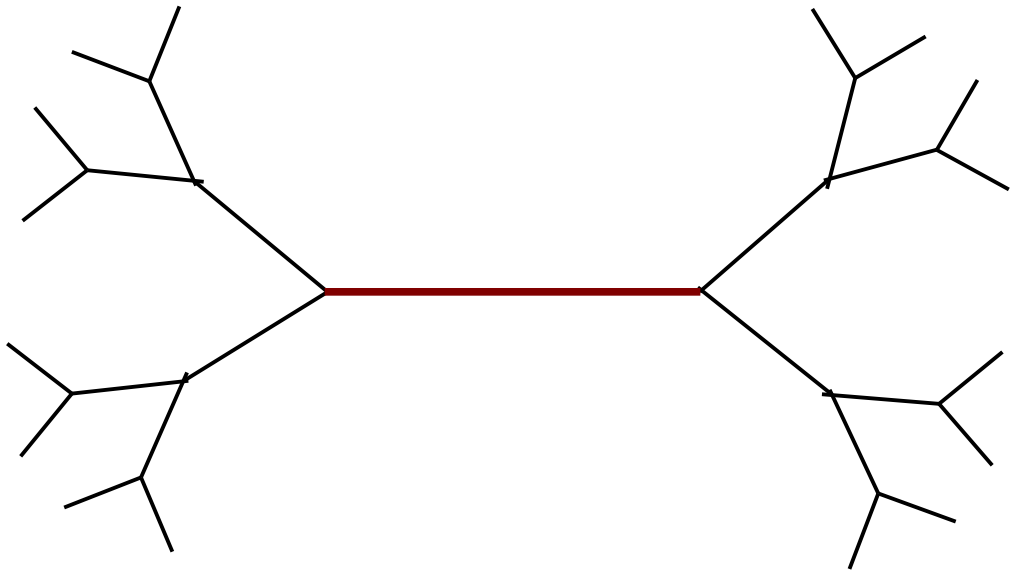
SEPP: SATé-enabled Phylogenetic Placement



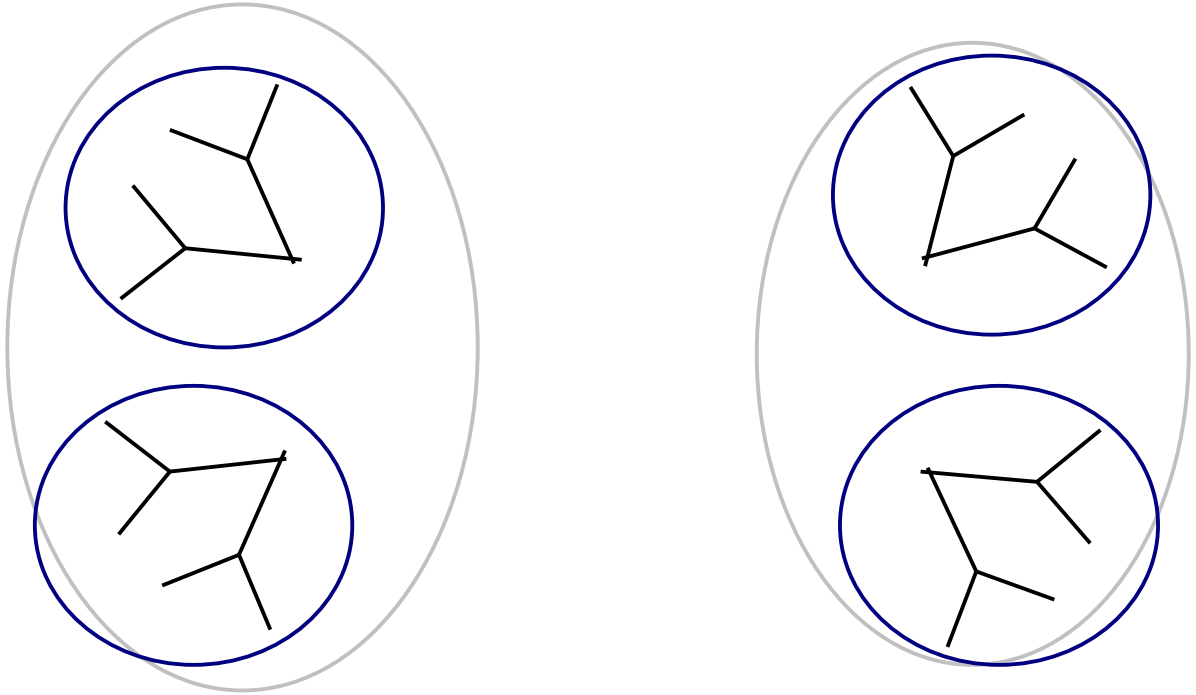
SEPP: SATé-enabled Phylogenetic Placement



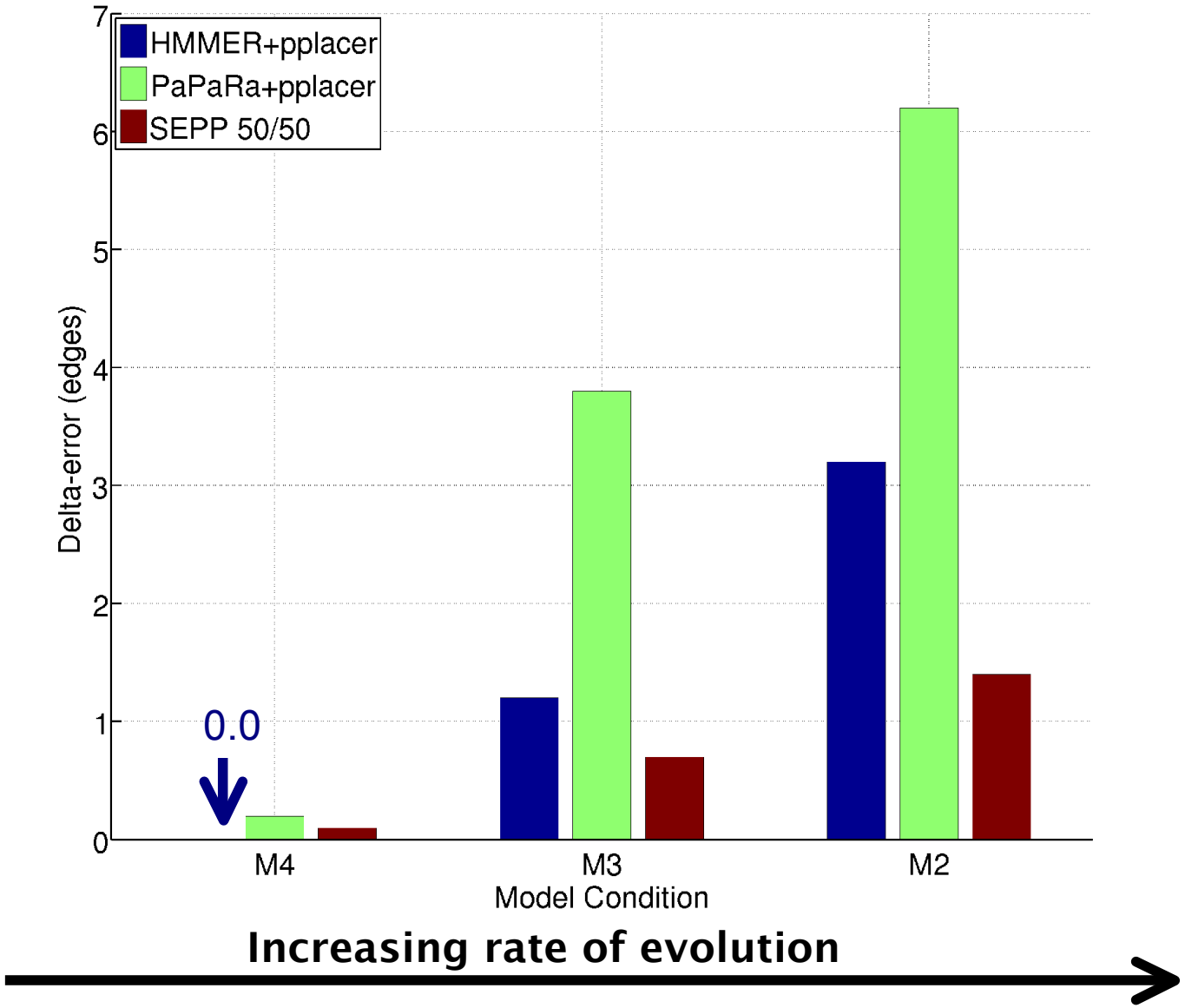
SEPP: SATé-enabled Phylogenetic Placement



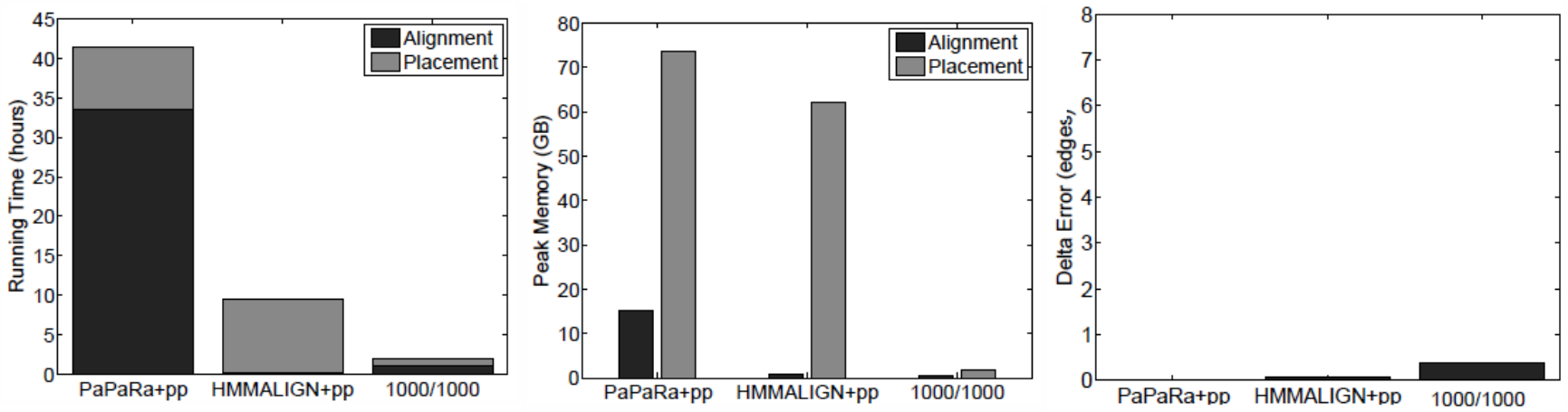
SEPP: SATé-enabled Phylogenetic Placement



SEPP (10%-rule) on Simulated Data

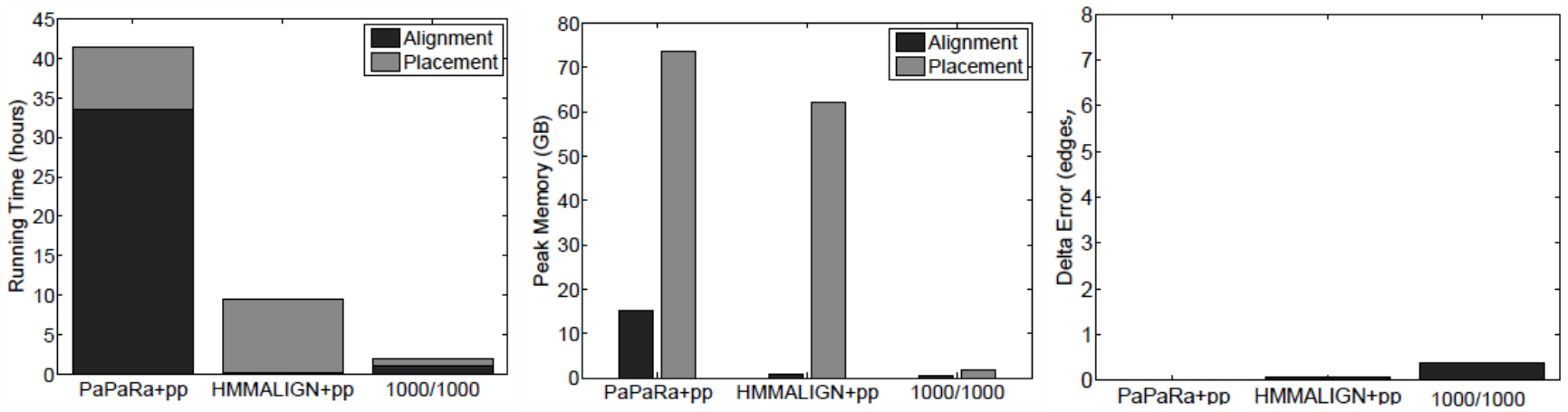


SEPP (10%) on Biological Data



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

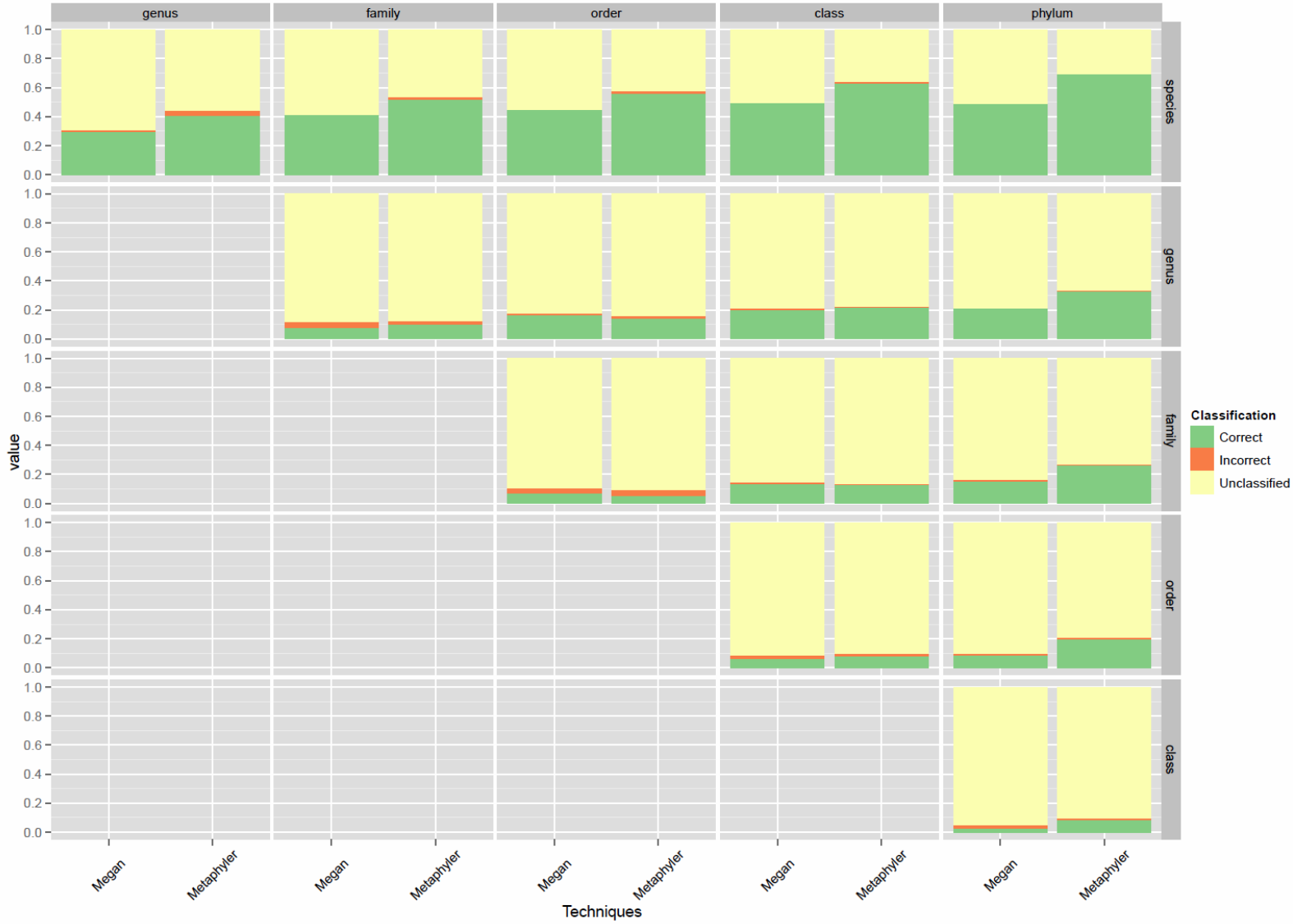
Taxon Identification

Metagenomic datasets include short reads from unknown species

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

Best current methods: Metaphyler, Phylopythia, and PhymmBL

60bp Error-Free Reads on rpsB Marker Gene: Megan and Metaphyler

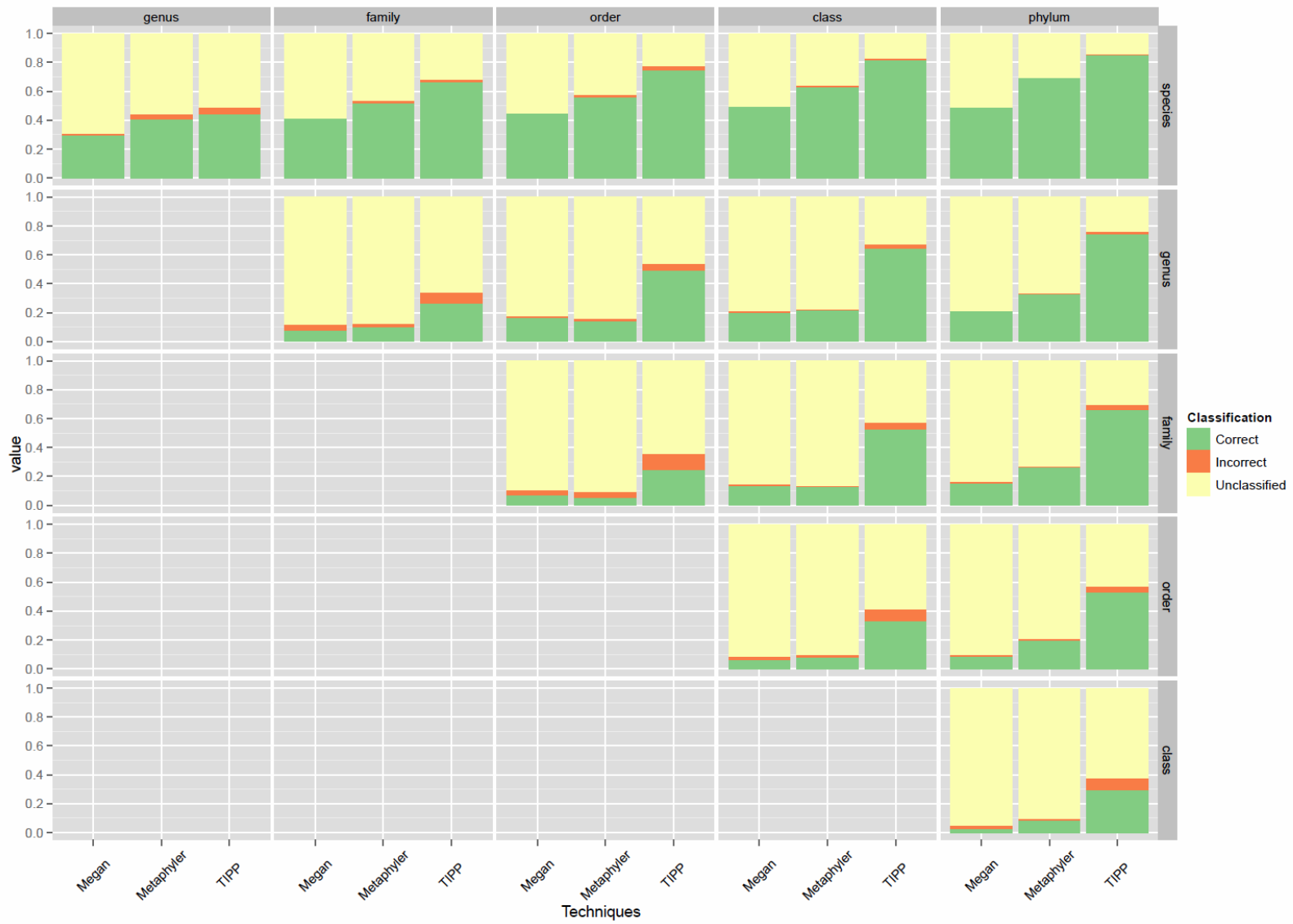


TIPP

- Taxon Identification using Phylogenetic Placement (Nguyen, Mirarab, and Warnow, in preparation)

- Approach: SEPP, modified to *take statistical uncertainty into account*

60bp Error-Free Reads on rpsB Marker Gene



Goals

- High accuracy
- Able to analyze large datasets
- Robust to model violations
- Robust to algorithmic parameters (e.g., starting trees)
- Improved biological analyses

Phylogenetic “boosters” (meta-methods)

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

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 for all phylogeny estimation methods (2012)
- SEPP-boosting for phylogenetic placement (2012)
- TIPP-boosting for taxon identification (in preparation)

Genomics and Big Data

- Relative performance of methods can change dramatically with dataset size.
- Standard statistical inference techniques often do not scale well.
- Divide-and-conquer and iteration can improve accuracy and speed of base methods.

Not Just Data Analysis, Not Just Algorithm Development

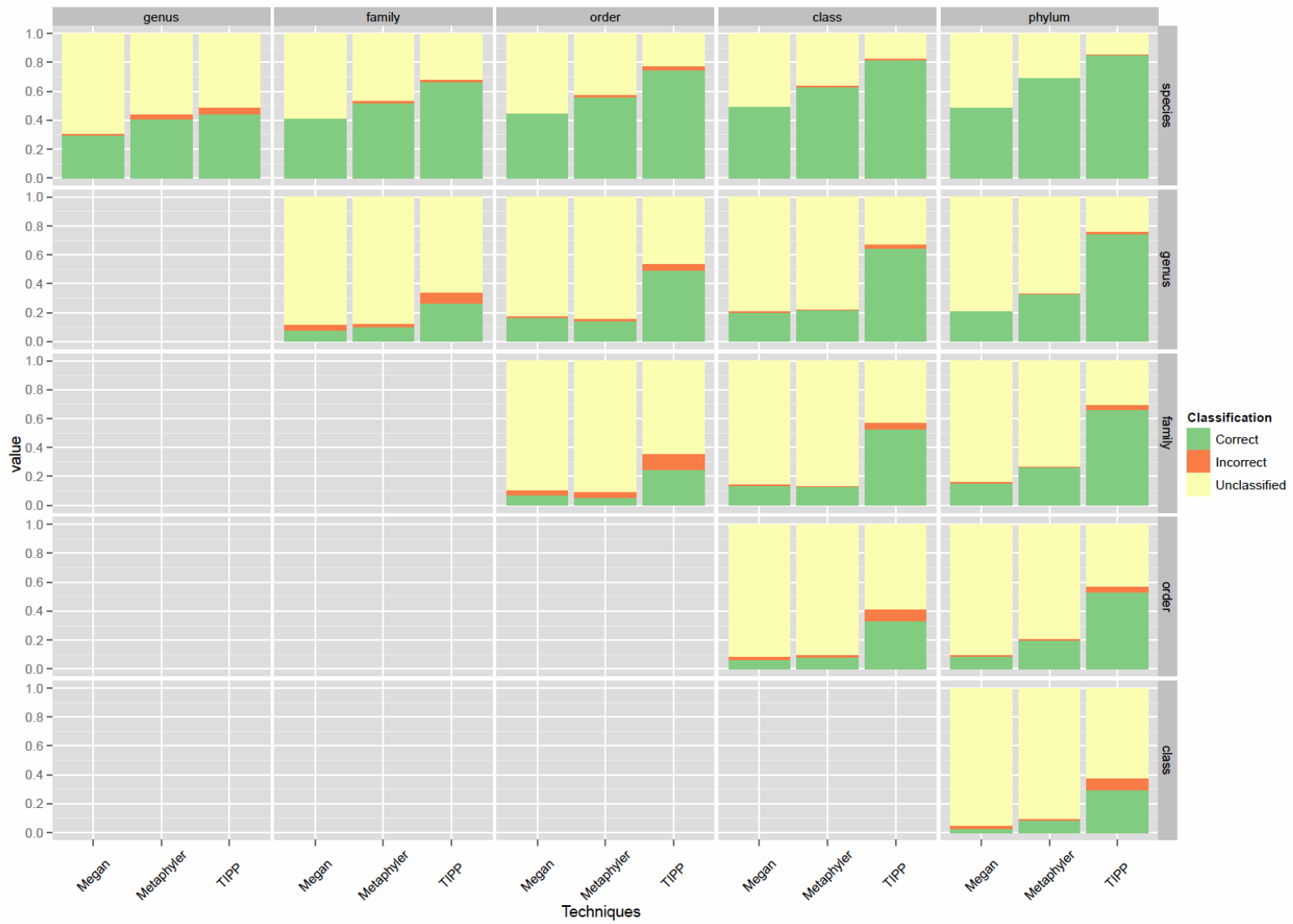
- Science is more complex than our mathematical models.
- Better analyses are needed in order to refine the models, and data are essential to accurate modelling.
- Hence, a ***cycle*** of mathematical modelling, statistical inference, methods for hard optimization problems, software development, extensive testing, …

Acknowledgments

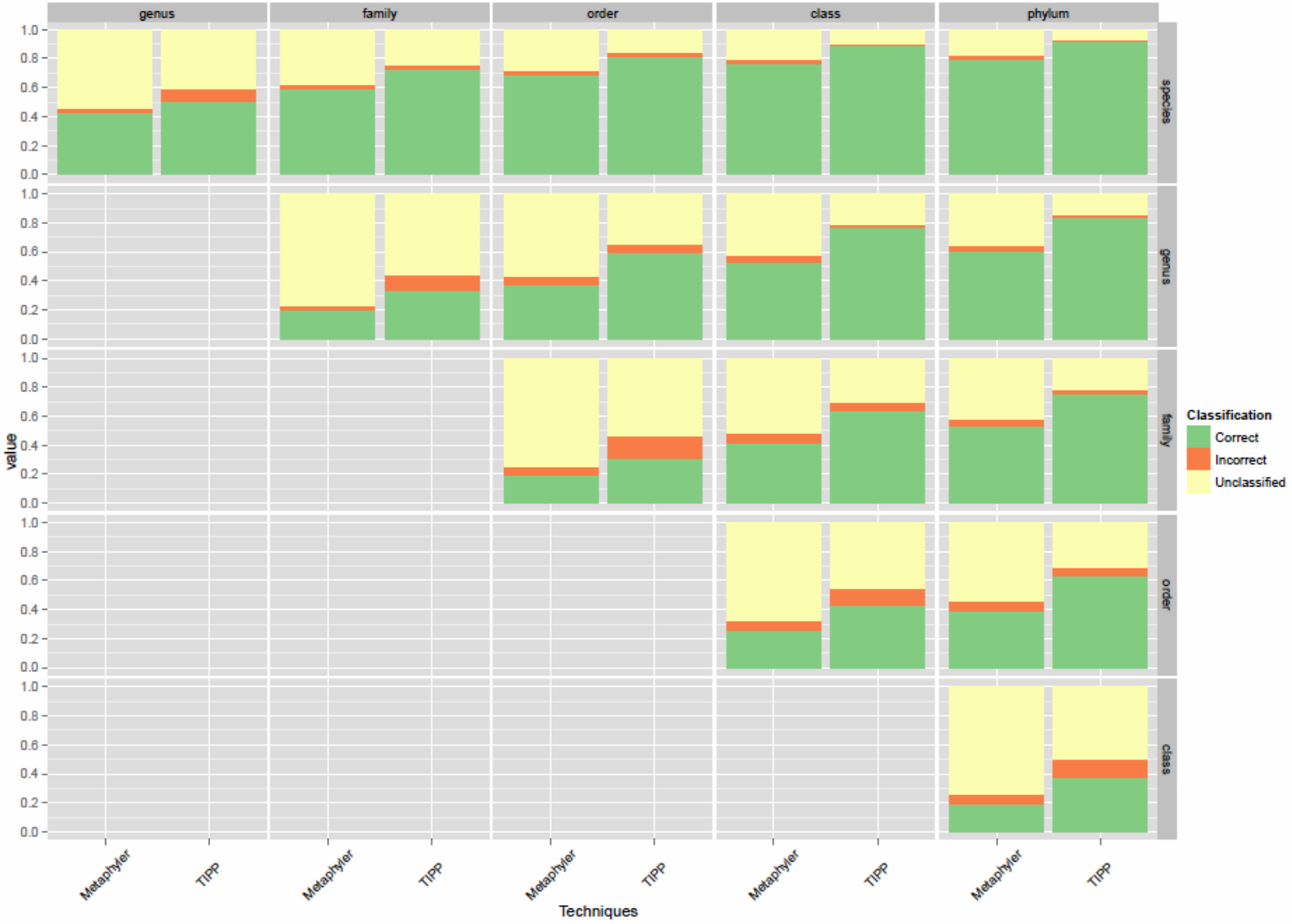
- Guggenheim Foundation Fellowship, Microsoft Research New England, National Science Foundation: Assembling the Tree of Life (ATOL), ITR, and IGERT grants, and David Bruton Jr. Professorship

- Collaborators:
  - DCM-NJ: Bernard Moret, Luay Nakhleh, and Katherine St. John
  - SATé: Randy Linder, Kevin Liu, Serita Nelesen, Sindhu Raghavan, and Li-San Wang
  - DACTAL: Serita Nelesen, Kevin Liu, Li-San Wang, and Randy Linder
  - SEPP/TIPP: Siavash Mirarab, Nam Nguyen (and Mihai Pop and Bo Liu)

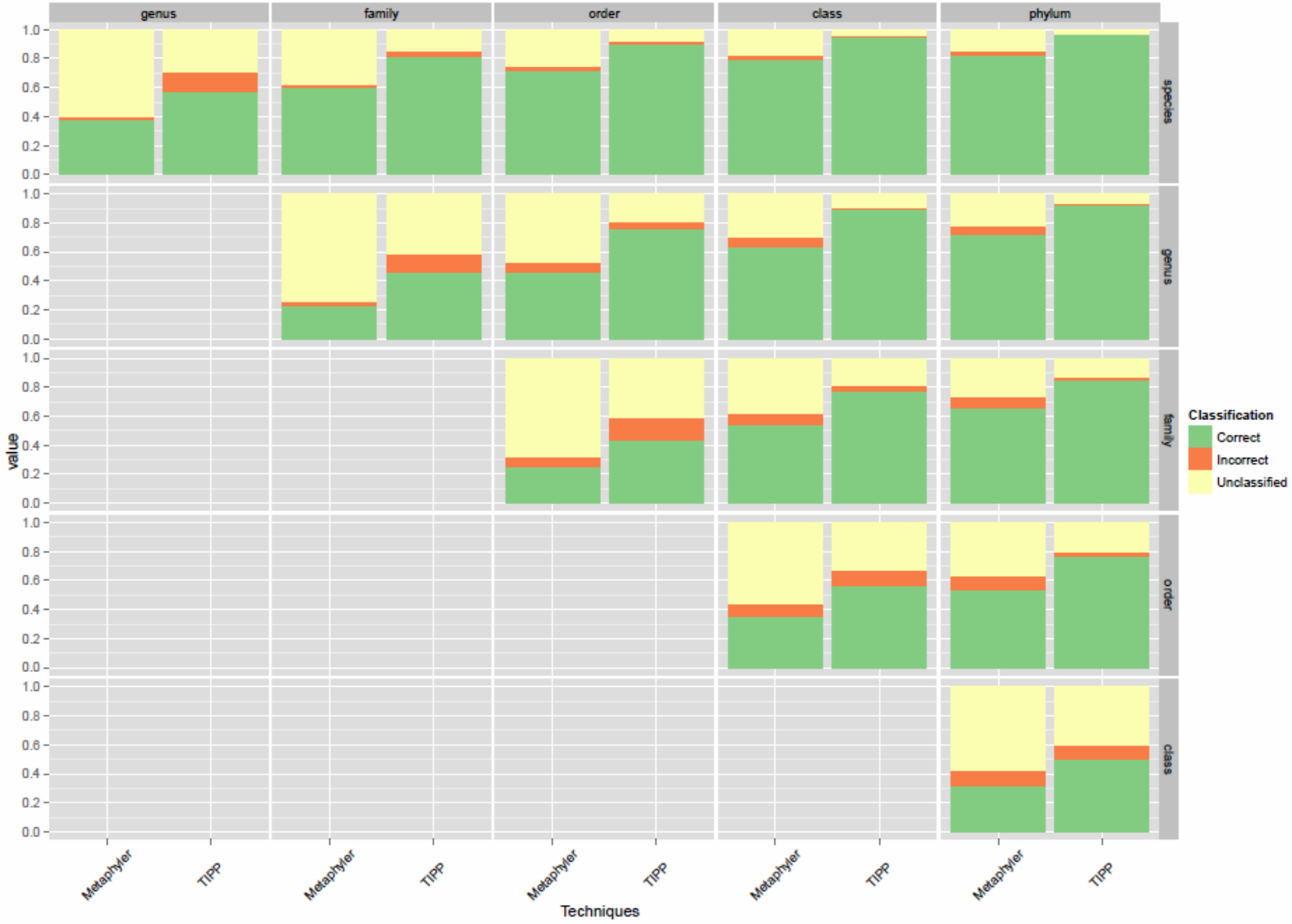
60bp error-free reads on rpsB marker gene

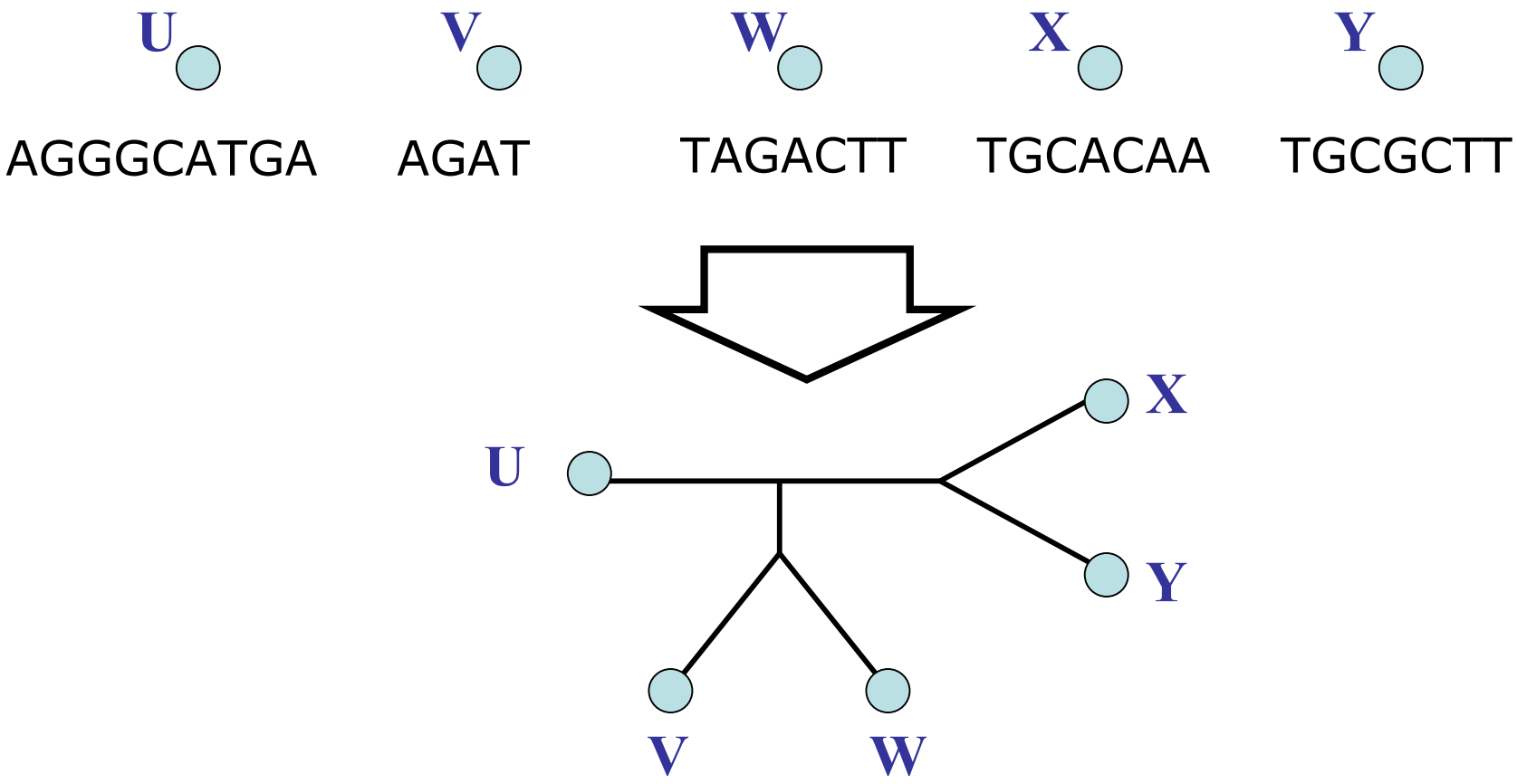


MetaPhyler versus TIPP on 100bp Illumina reads across 29 marker genes



MetaPhyler versus TIPP on 300bp 454 reads across 29 marker genes




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

Research Projects

Theory: Phylogenetic estimation under statistical models

Method development:

- "Absolute fast converging" methods
- Very large-scale multiple sequence alignment and phylogeny estimation
- Estimating species trees and networks from gene trees
- Supertree methods
- Comparative genomics (genome rearrangement phylogenetics)
- Metagenomic taxon identification
- Alignment and Phylogenetic Placement of NGS data

Dataset analyses

- Avian Phylogeny: 50 species and 8000+ genes
- Thousand Transcriptome (1KP) Project: 1000 species and 1000 genes
- Chloroplast genomics