Advances in Ultra-large Phylogeny Estimation
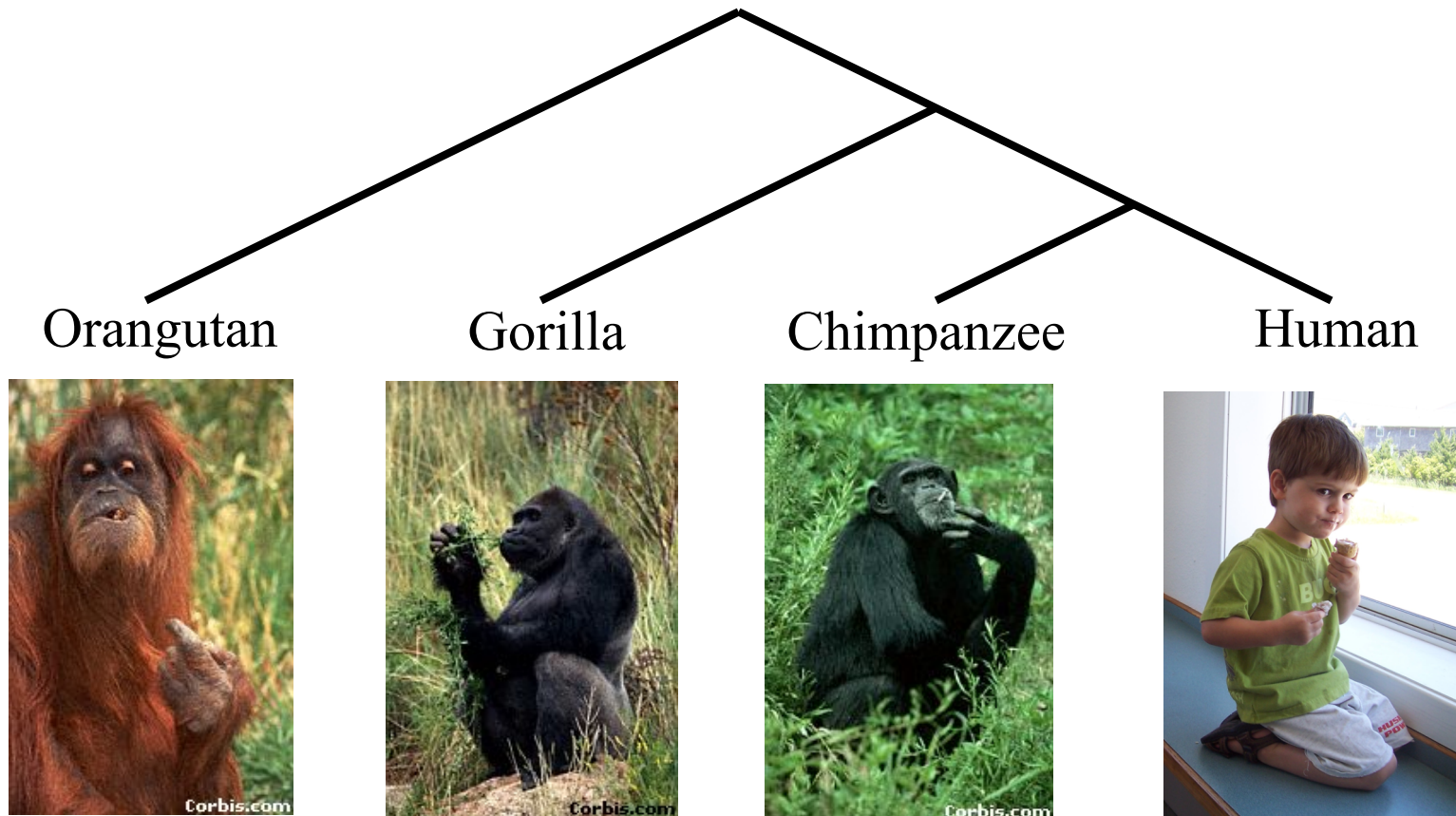
Tandy Warnow

Department of Computer Science

University of Texas

Phylogeny (evolutionary tree)
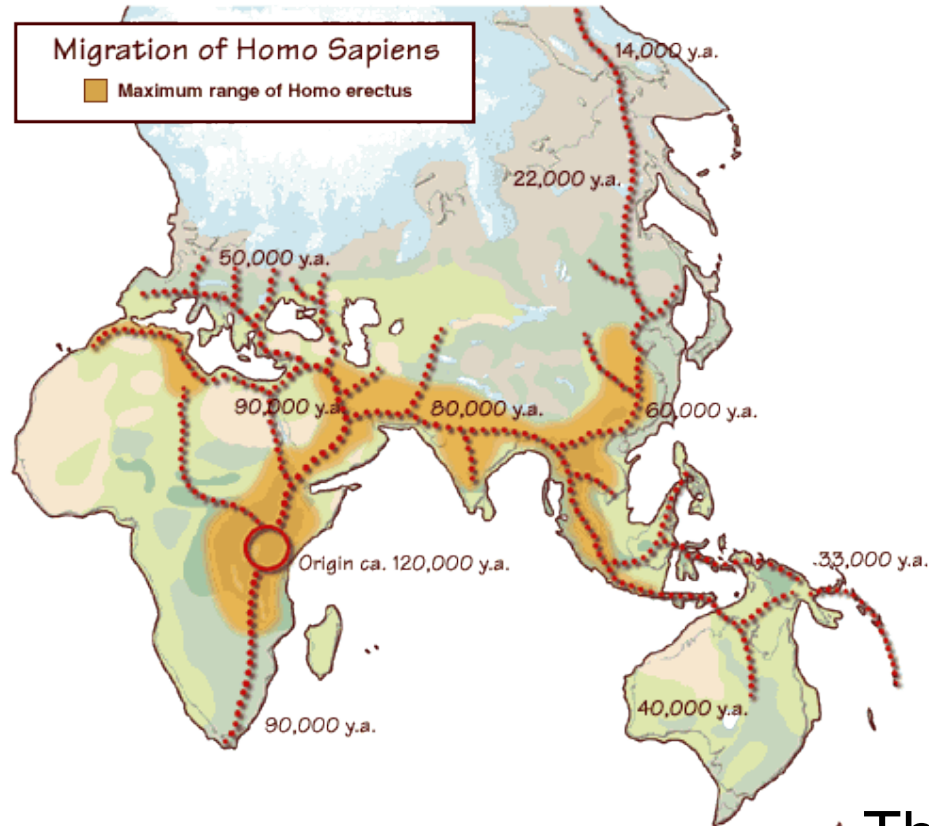
Orangutan   Gorilla   Chimpanzee   Human

From the Tree of the Life Website,
University of Arizona

How did life evolve on earth?

Courtesy of the Tree of Life project

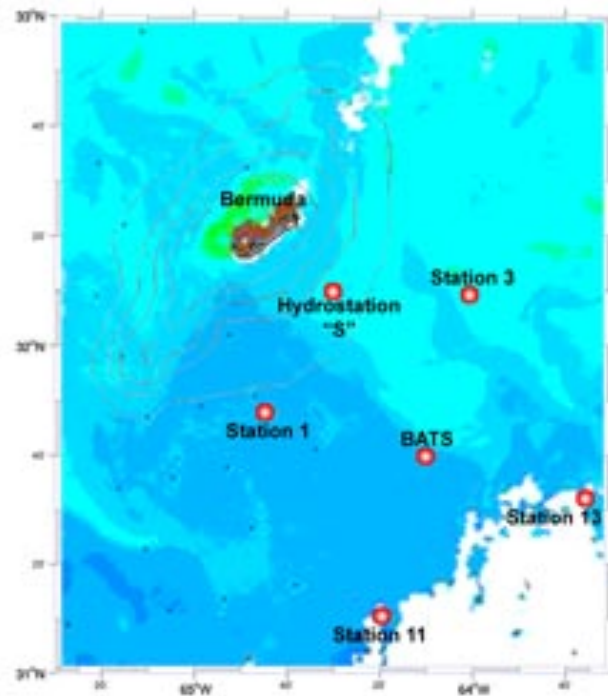Where did humans come from, and how did they move throughout the globe?

The 1000 Genome Project: using human genetic variation to better treat diseases

Metagenomics:

## C. Venter et al., Exploring the Sargasso Sea:

Scientists Discover One Million New Genes in Ocean Microbes

Major Challenges

- Current phylogenetic datasets contain hundreds to thousands of taxa, with multiple genes.

- Future datasets will be substantially larger (e.g., iPlant plans to construct a tree on 500,000 plant species)

- *Current methods have poor accuracy or cannot run on large datasets.*

Computational Phylogenetics
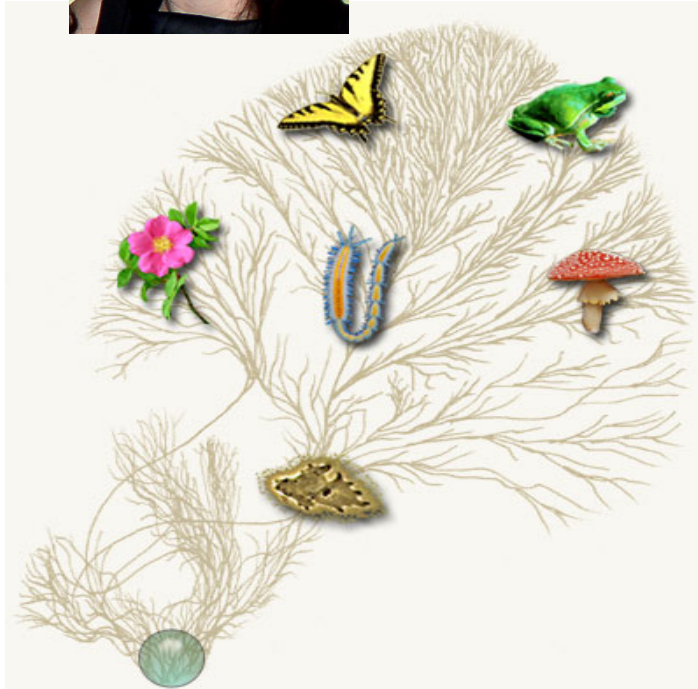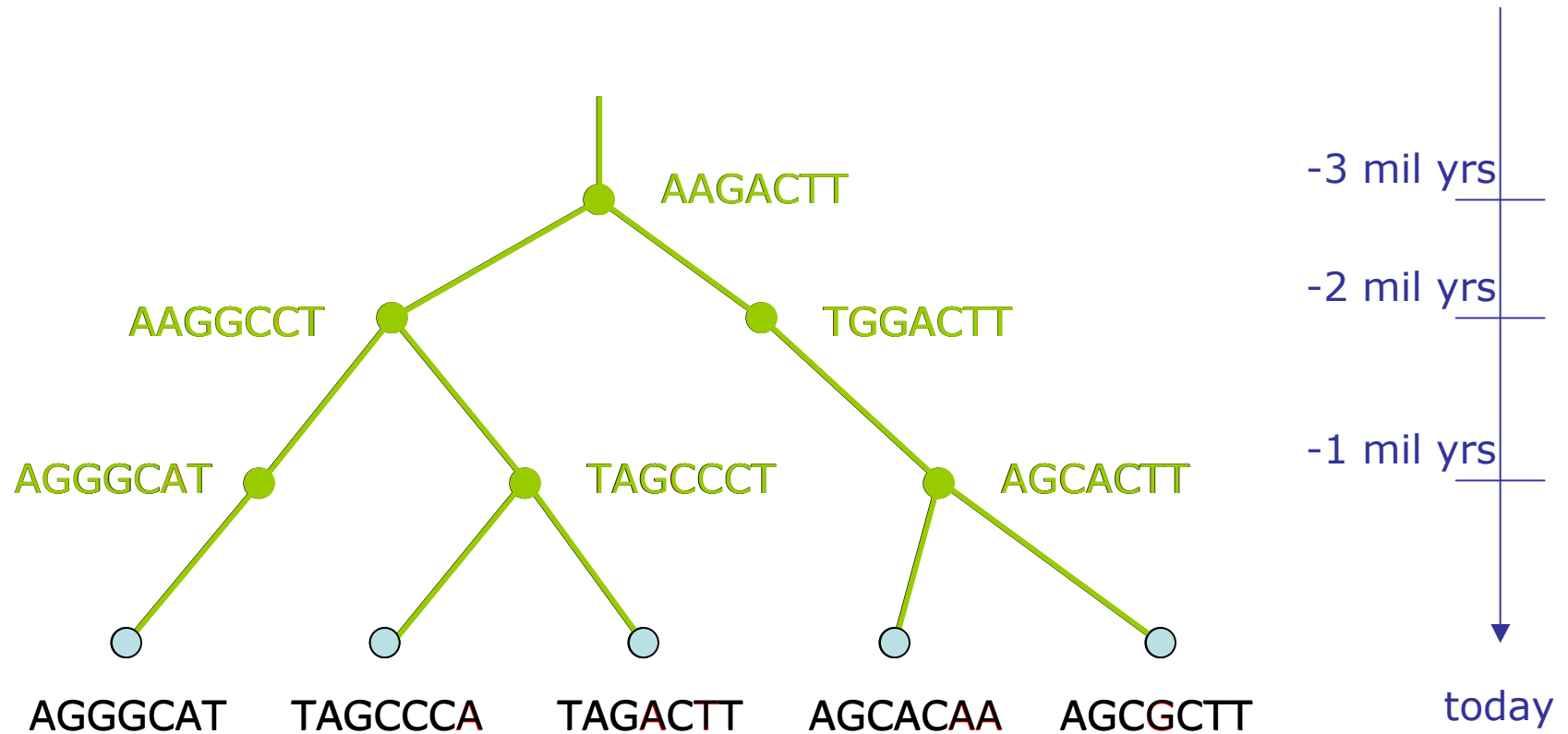

Courtesy of the Tree of Life project

Current methods can use months to

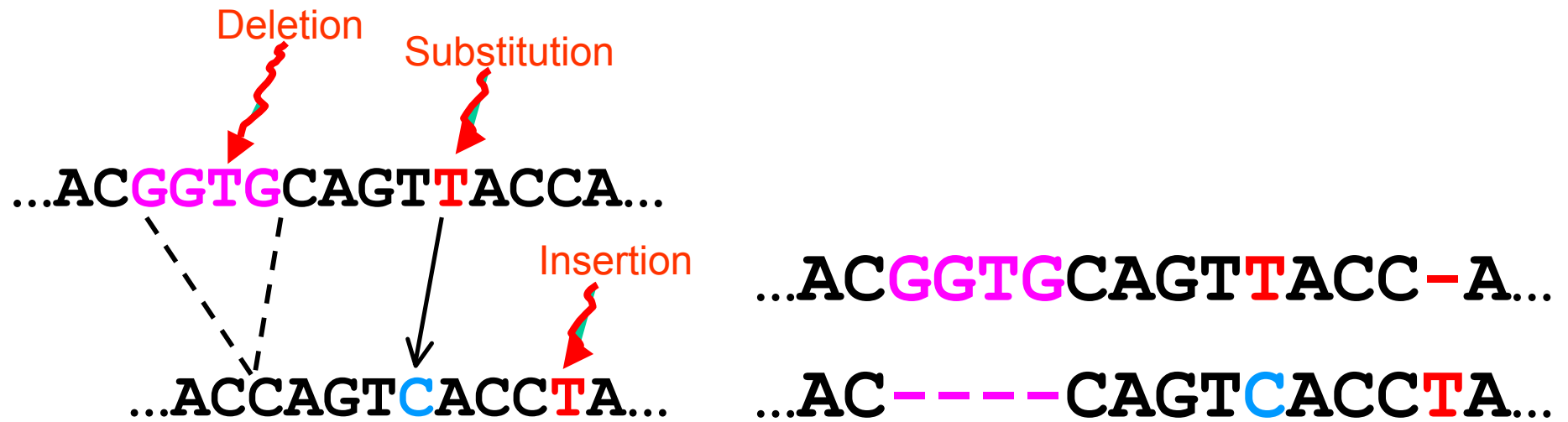estimate trees on 1000 DNA sequences

Our objective:

More accurate trees and alignments

on 500,000 sequences in under a week

We prove theorems using graph theory

and probability theory, and our

algorithms are studied on real and

simulated data.

DNA Sequence Evolution

AAGACTT

AAGGCCT

TGGACTT

AGGGCAT

TAGCCCT

AGCACTT

AGGGCAT TAGCCCA TAGACTT AGCACAA AGCGCTT

-3 mil yrs

-2 mil yrs

-1 mil yrs

today

Deletion **Substitution**

...AC**GGTG**CAGT**T**ACCA...

Insertion

...ACCAGT**C**ACC**T**A...

...AC**GGTG**CAGT**T**ACC**-**A...

...AC**----**CAGT**C**ACC**T**A...

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

U
AGGGCATGA

V
AGAT

W
TAGACTT

X
TGCACAA

Y
TGCGCTT

U

X

Y

V

W
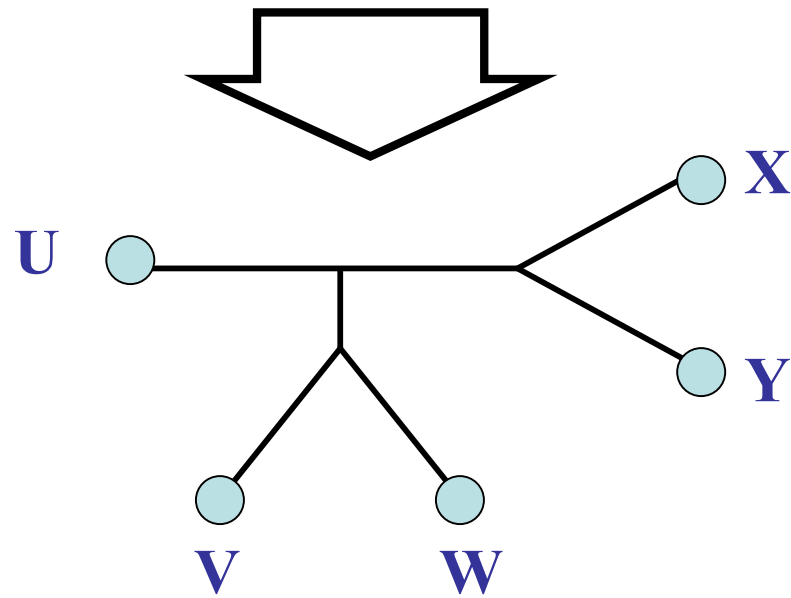
Input: unaligned sequences

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

Phase 1: Multiple Sequence Alignment

```
S1 = AGGCTATCACCTGACCTCCA          S1 = -AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC              S2 = TAG-CTATCAC--GACCGC--
S3 = TAGCTGACCGC           →       S3 = TAG-CT-------GACCGC--
S4 = TCACGACCGACA                  S4 = -------TCAC--GACCGACA
```
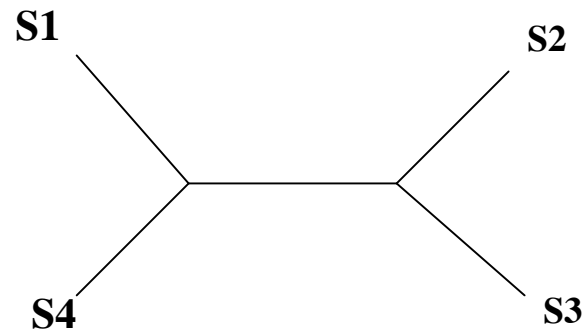
Phase 2: Construct tree

```
S1 = AGGCTATCACCTGACCTCCA              S1 = -AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC                  S2 = TAG-CTATCAC--GACCGC--
S3 = TAGCTGACCGC          ───────▶     S3 = TAG-CT-------GACCGC--
S4 = TCACGACCGACA                      S4 = -------TCAC--GACCGACA
```
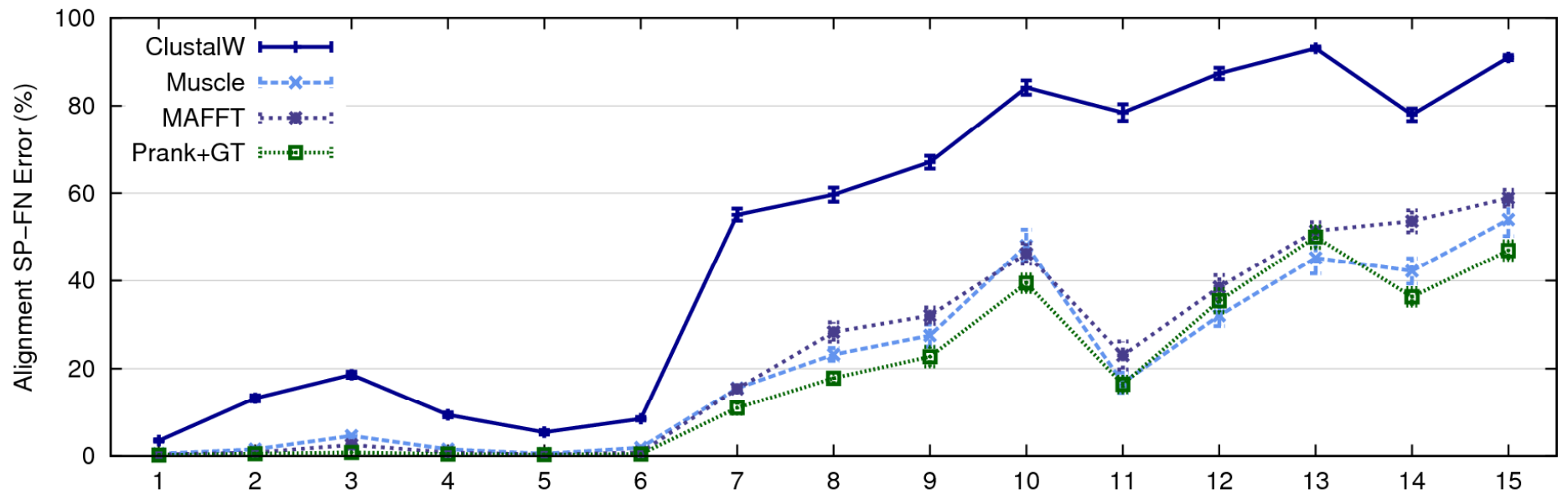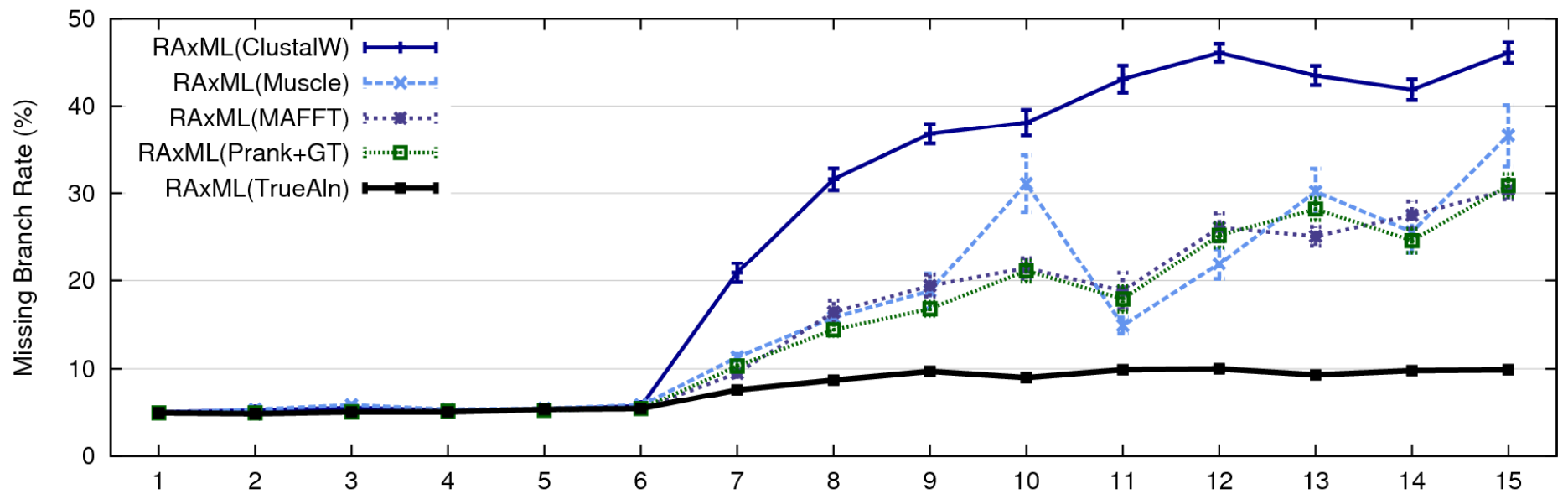
Simulation Studies
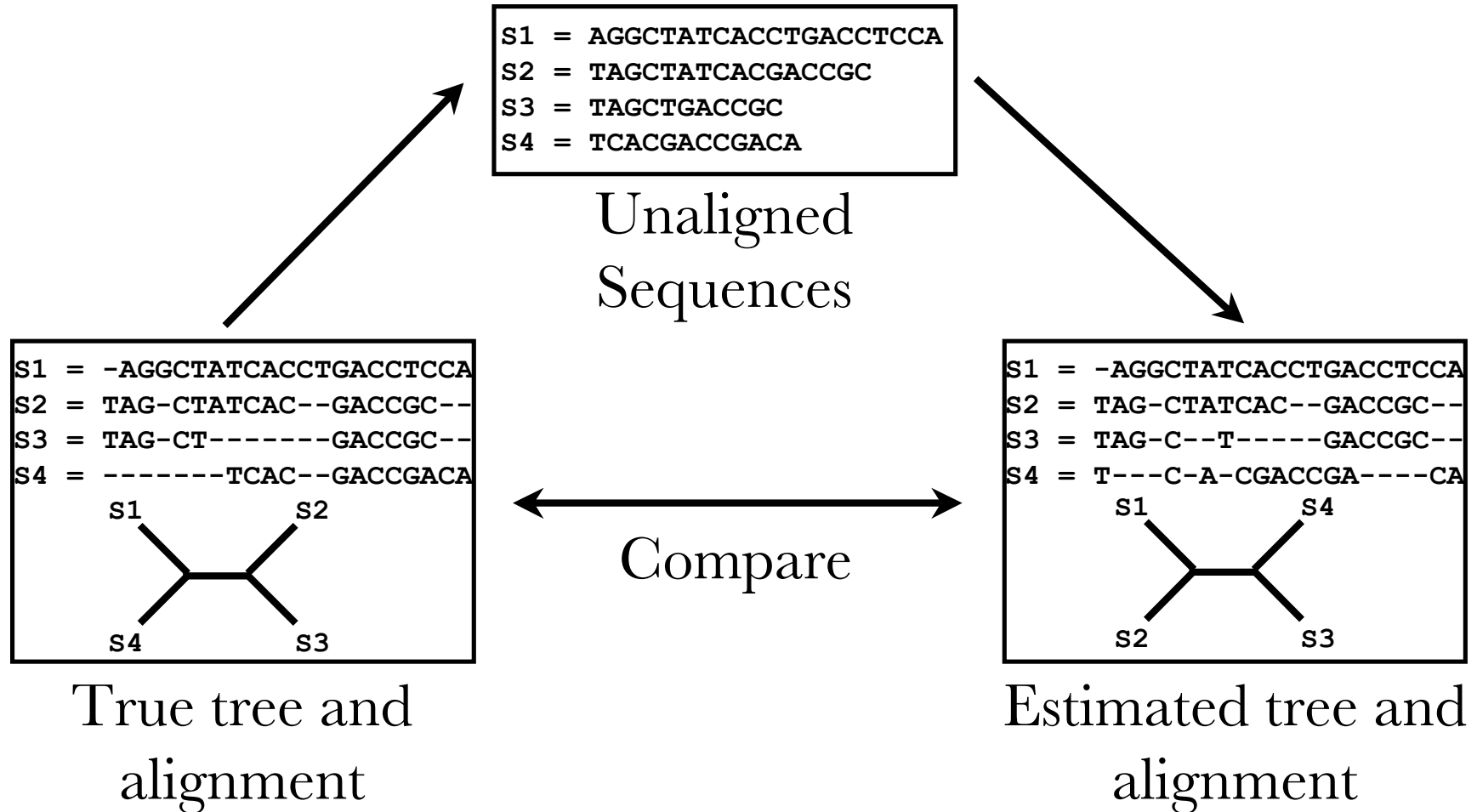
The neighbor joining method has high error rates on large trees
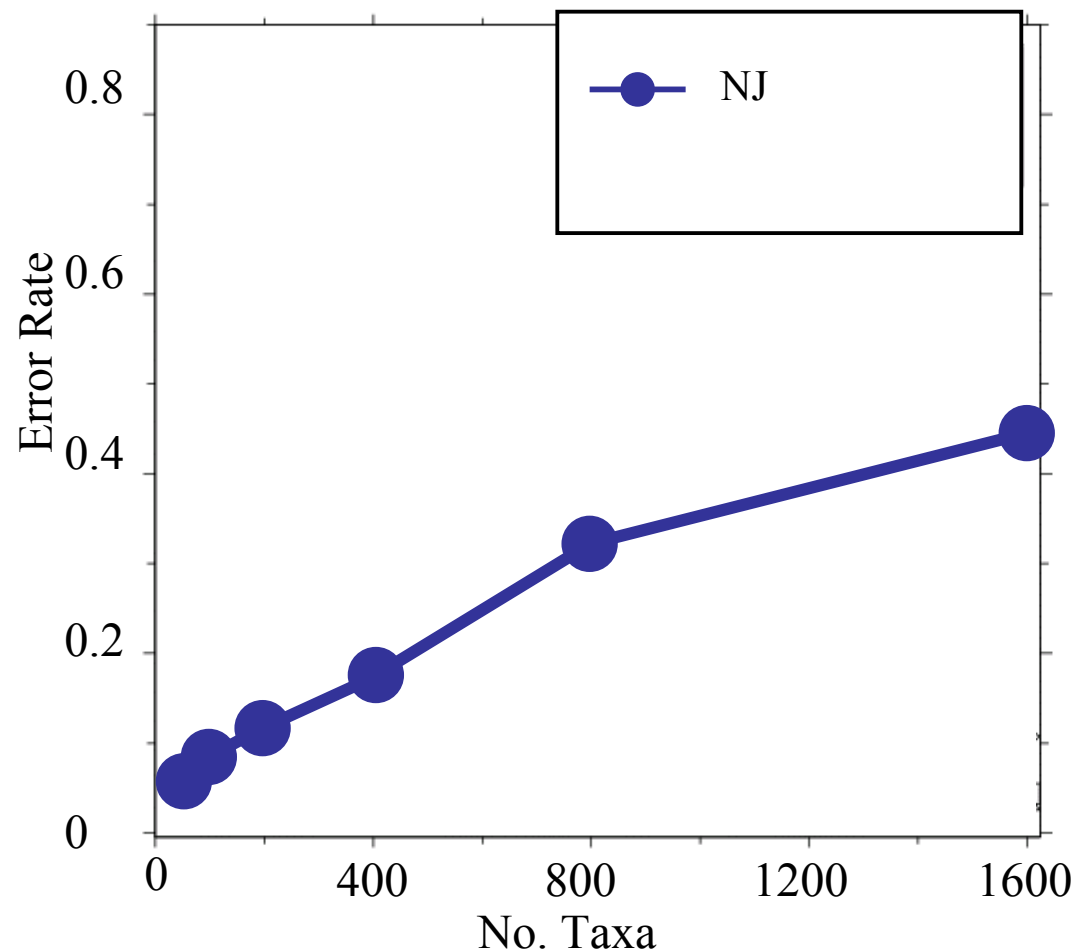
Simulation study based upon fixed edge lengths, K2P model of evolution, sequence lengths fixed to 1000 nucleotides.

Error rates reflect proportion of incorrect edges in inferred trees.

[Nakhleh et al. ISMB 2001]

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)

Major Challenges

- Current phylogenetic datasets contain hundreds to thousands of taxa, with multiple genes.

- Future datasets will be substantially larger (e.g., iPlant plans to construct a tree on 500,000 plant species)

- *Current methods have poor accuracy or cannot run on large datasets.*

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 (2011)
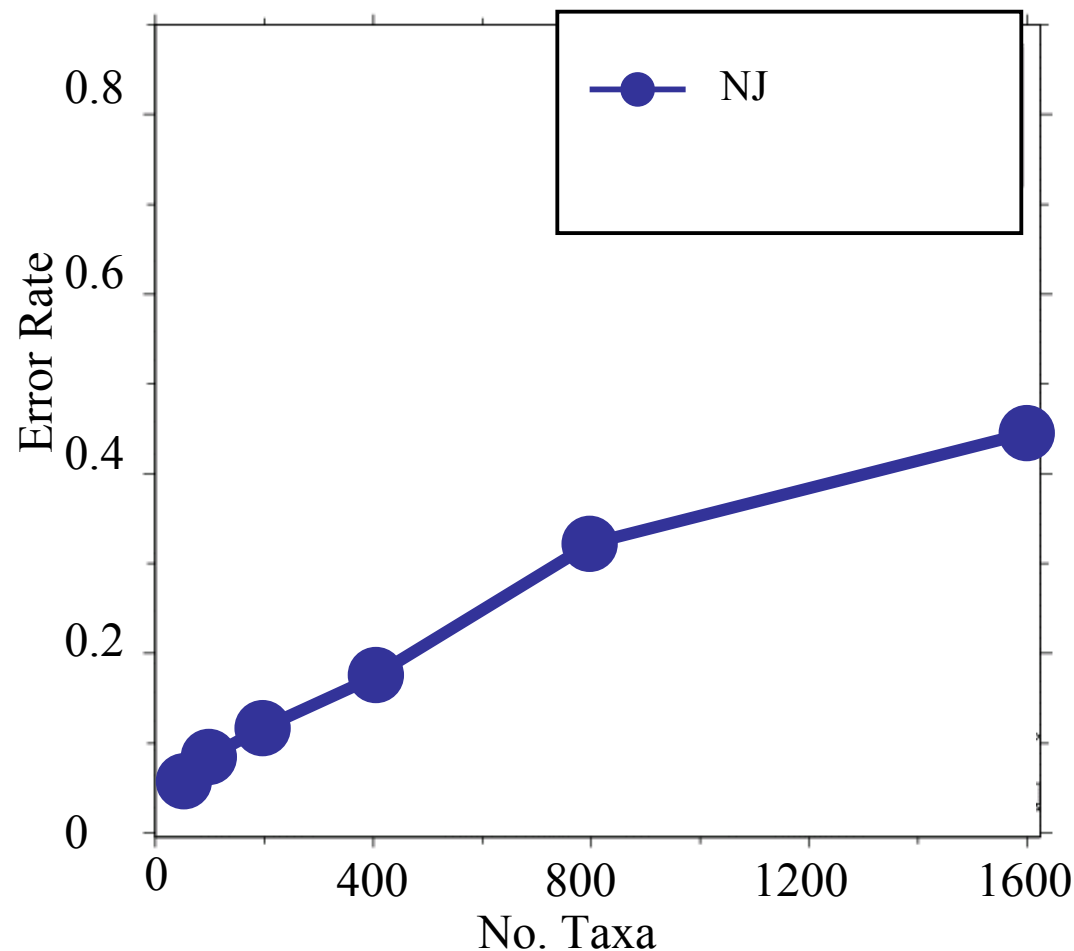- SEPP-boosting for metagenomic analyses (2011)

Disk-Covering Methods (DCMs)
(starting in 1998)
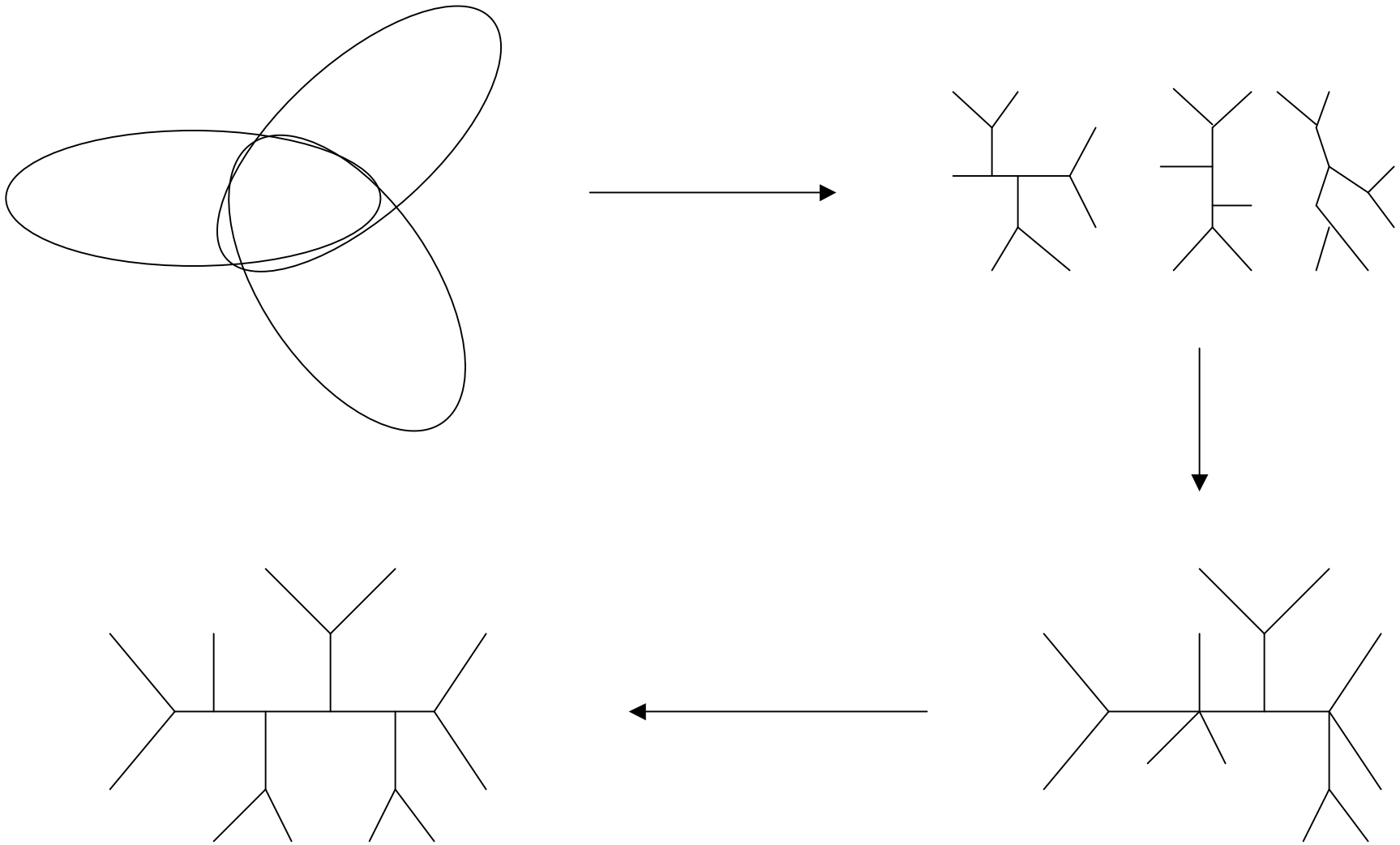
- DCMs "boost" the performance of phylogeny reconstruction methods.
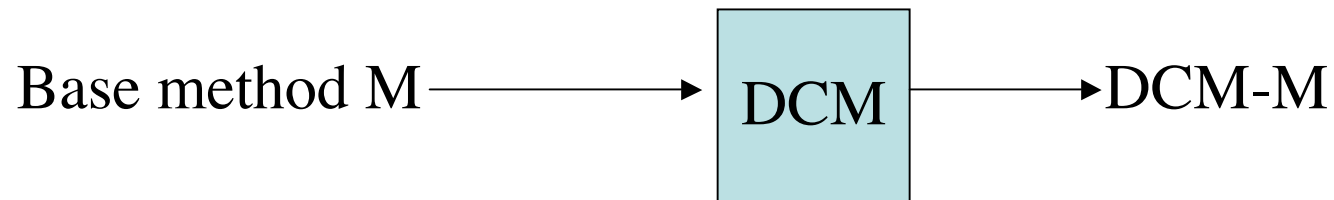
Base method M $\longrightarrow$ DCM $\longrightarrow$ DCM-M

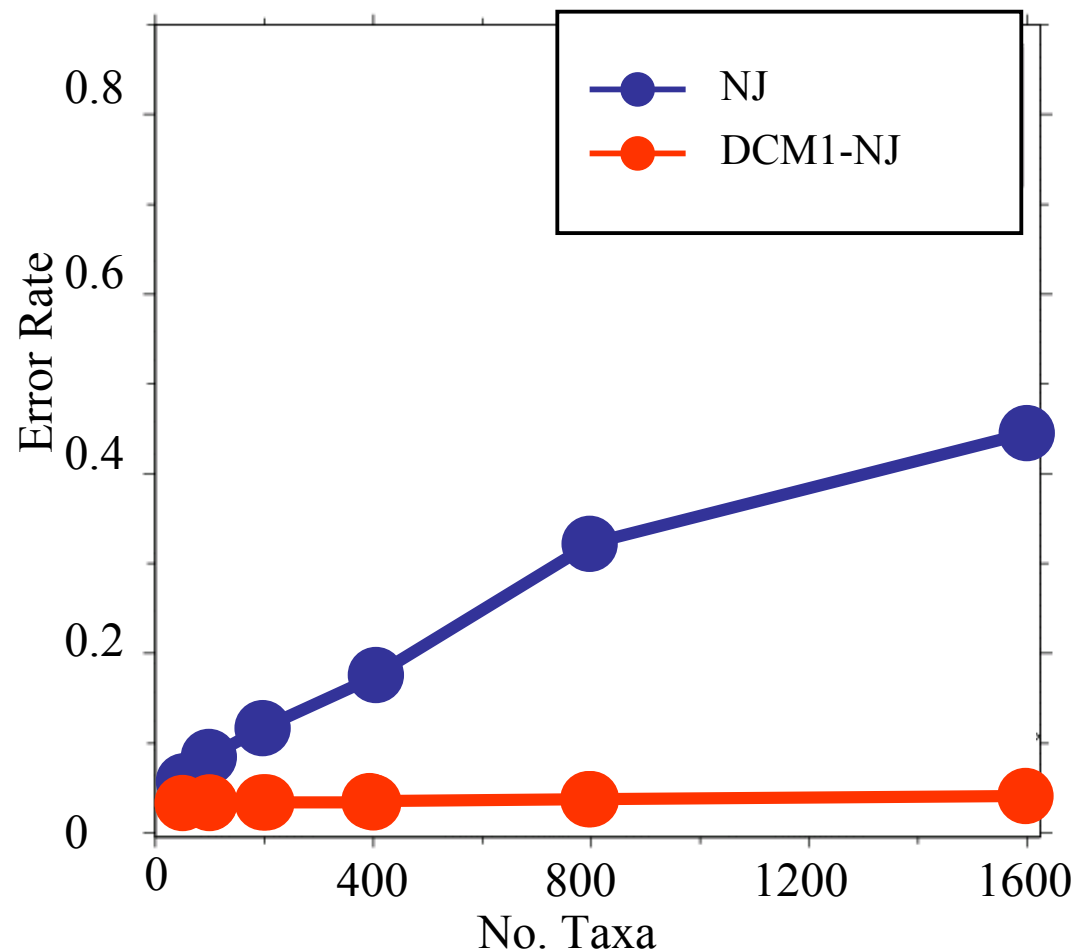# The neighbor joining method has high error rates on large trees

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]

DCM1-boosting distance-based methods
[Nakhleh et al. ISMB 2001]

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

Today's Talk

- **SATé**: Simultaneous Alignment and Tree Estimation (Liu et al., Science 2009, and Liu et al. Systematic Biology, in press)

- **DACTAL**: Divide-and-Conquer Trees (Almost) without alignments (Nelesen et al., submitted)

- **SEPP:** SATé-enabled Phylogenetic Placement (Mirarab, Nguyen and Warnow, to appear, PSB 2012)

Part 1: SATé

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

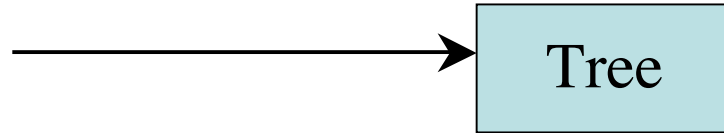Liu et al., Systematic Biology (in press)

Public software distribution (open source) through the University of Kansas, in use, world-wide
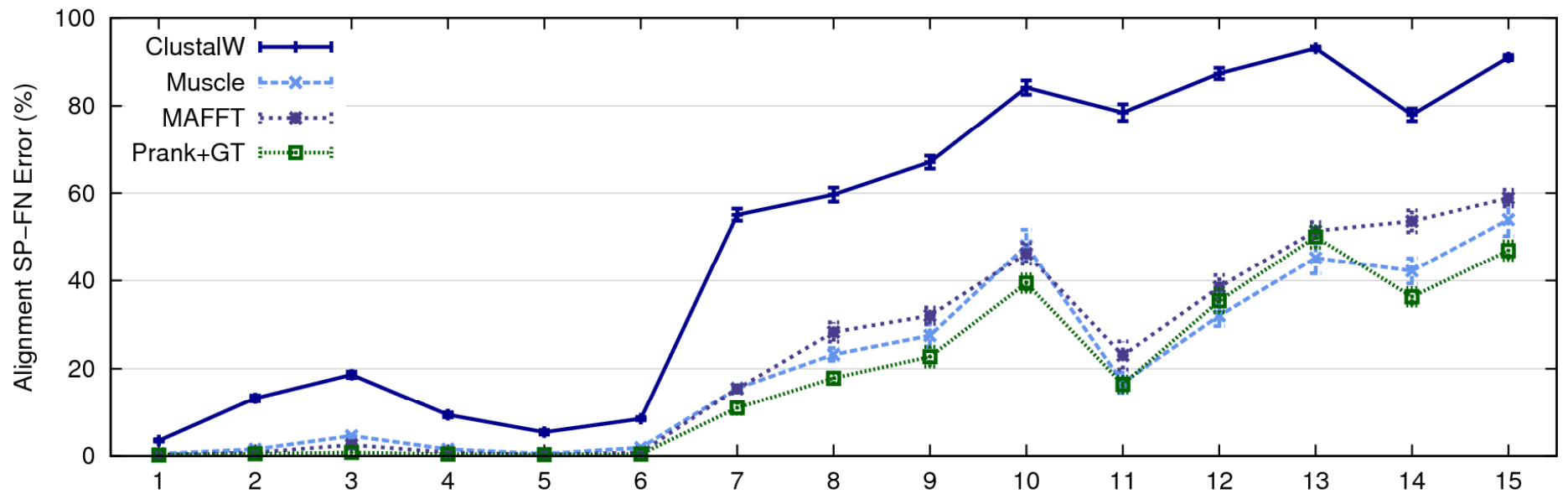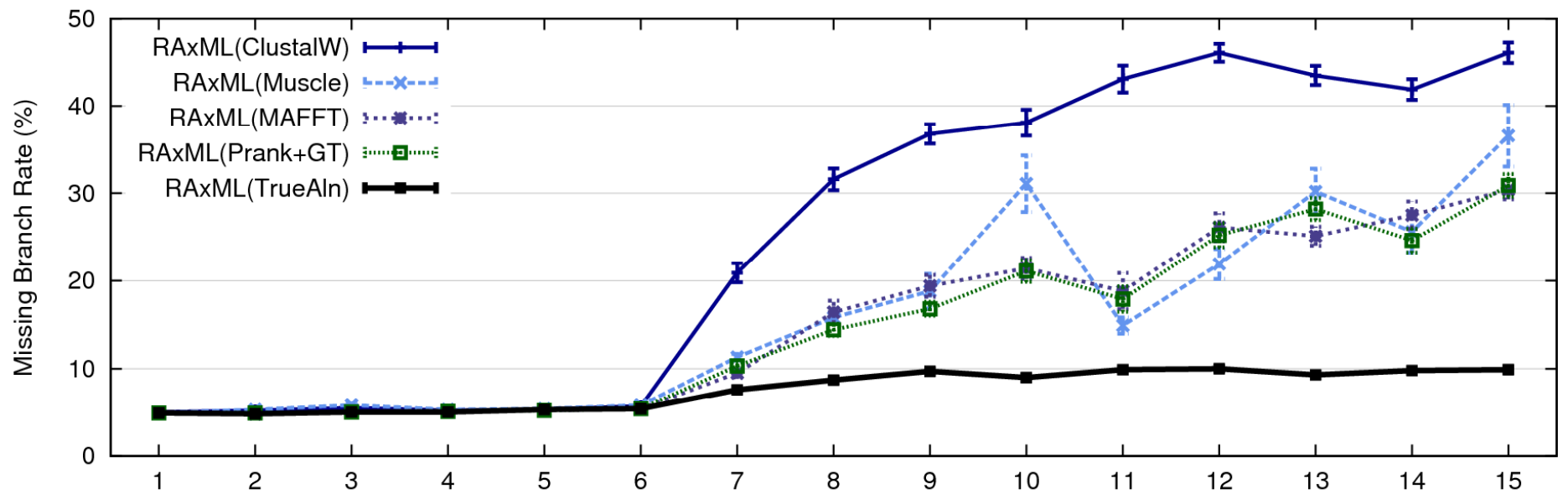
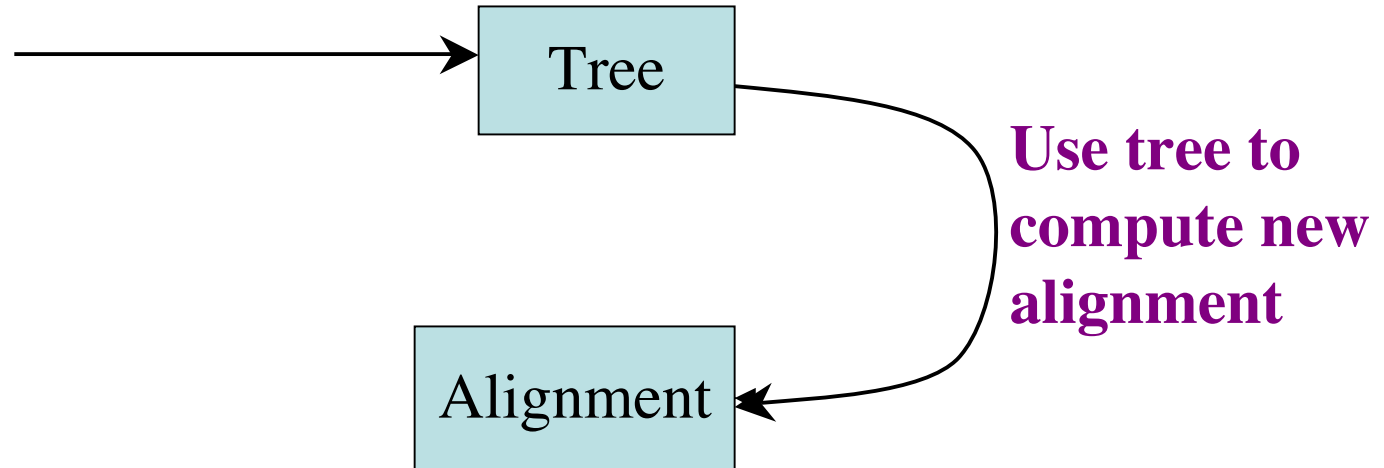1000 taxon models, ordered by difficulty (Liu et al., 2009)

SATé Algorithm

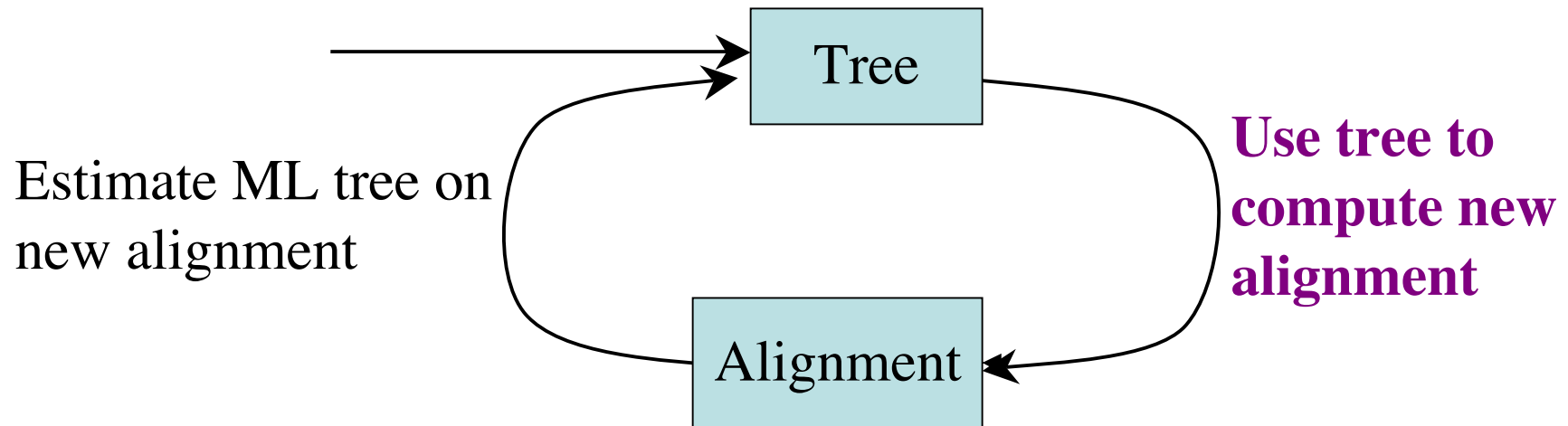Obtain initial alignment
and estimated ML tree

Tree

SATé Algorithm
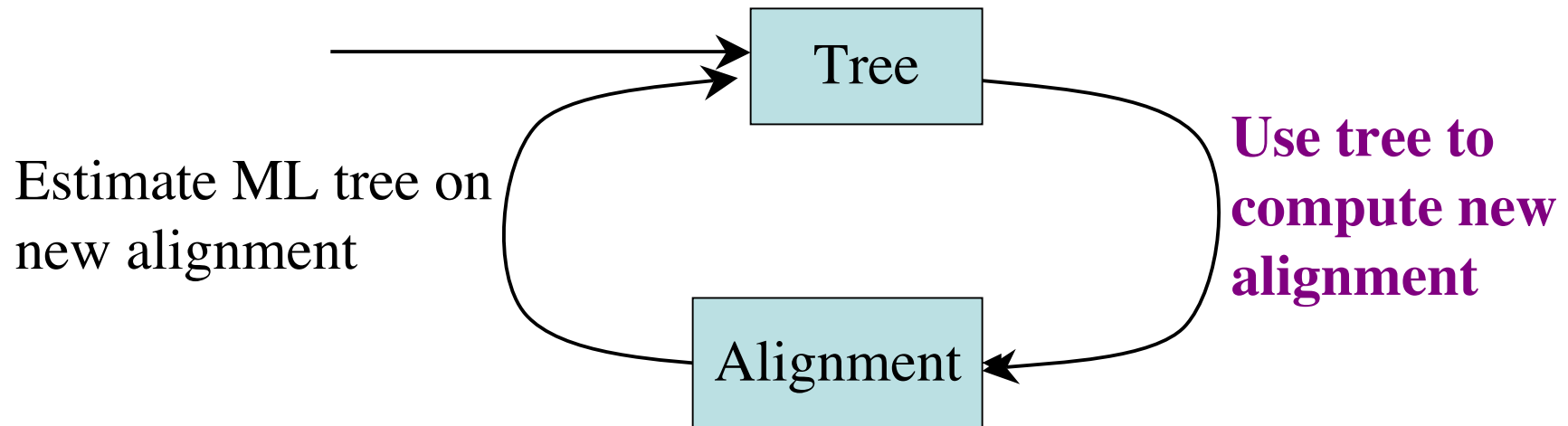
Obtain initial alignment
and estimated ML tree

Tree

Alignment

**Use tree to
compute new
alignment**

SATé Algorithm

Obtain initial alignment
and estimated ML tree

Estimate ML tree on
new alignment

Tree

Alignment

**Use tree to
compute new
alignment**

SATé Algorithm

Obtain initial alignment
and estimated ML tree

Estimate ML tree on
new alignment

Tree

Alignment

**Use tree to
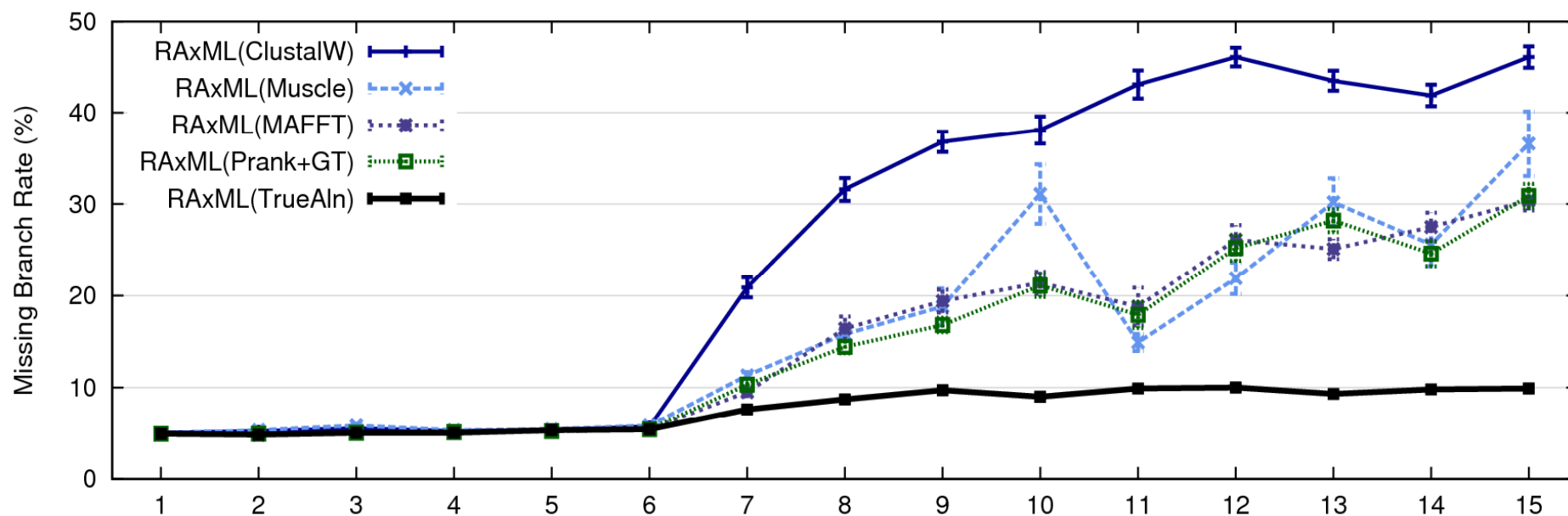compute new
alignment**

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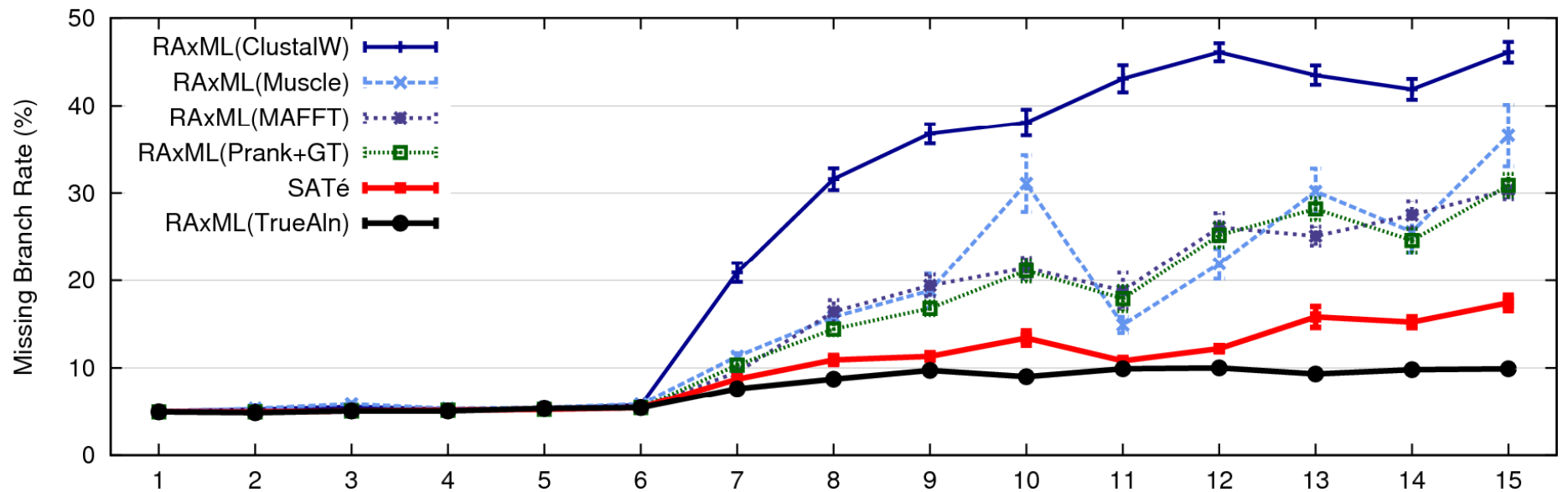a different decomposition

Repeat until termination condition (typically, 24 hours)
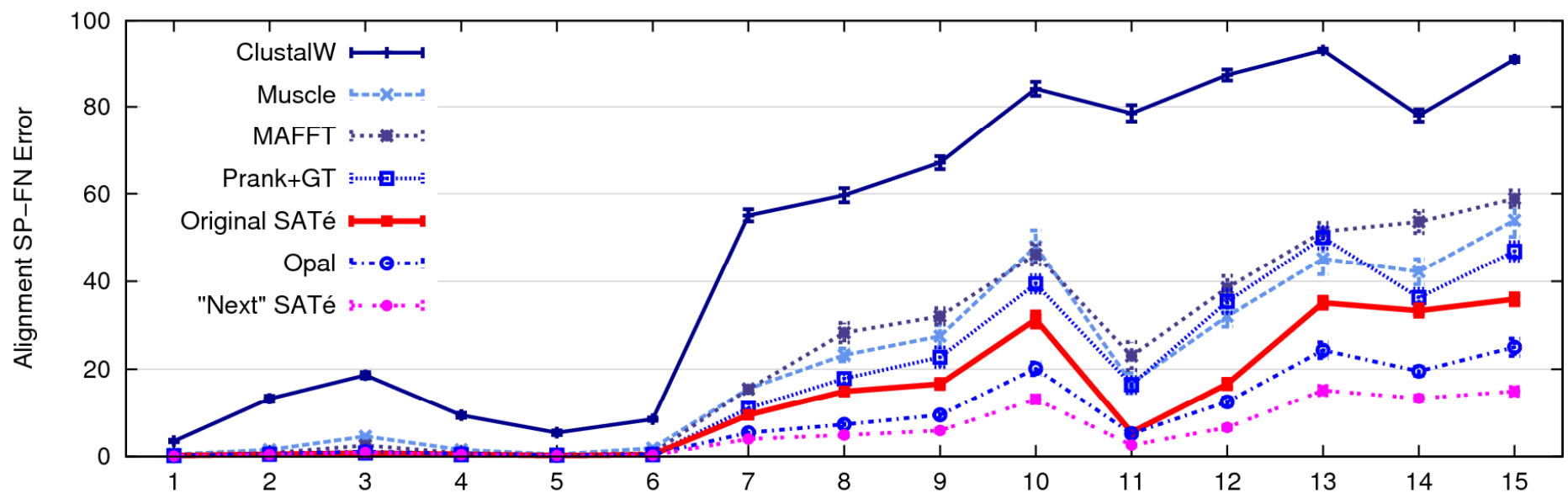
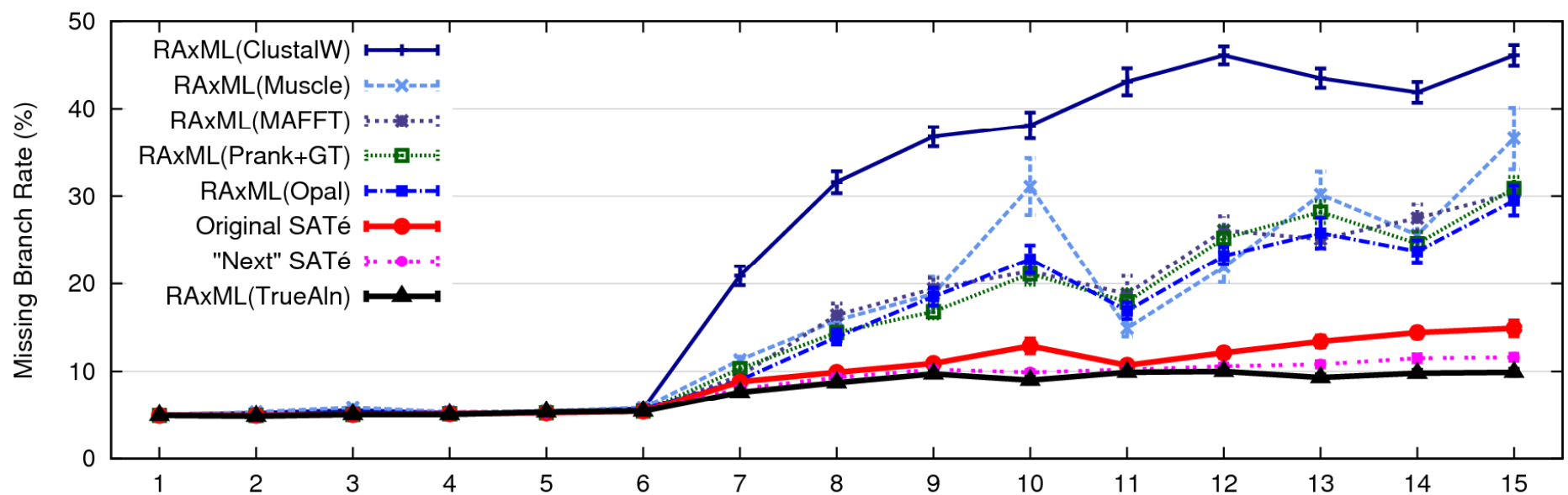# One SATé iteration (really 32 subsets)
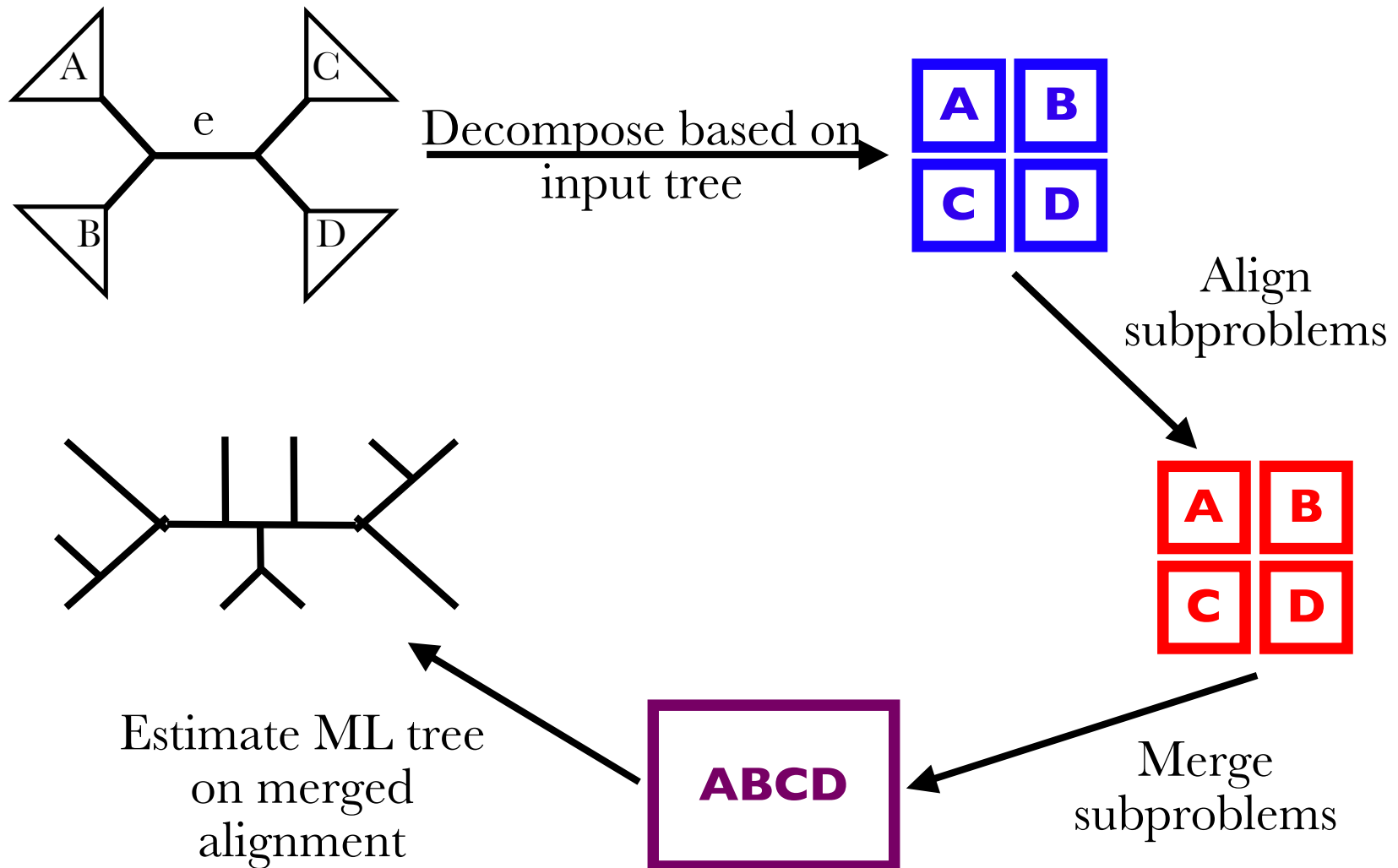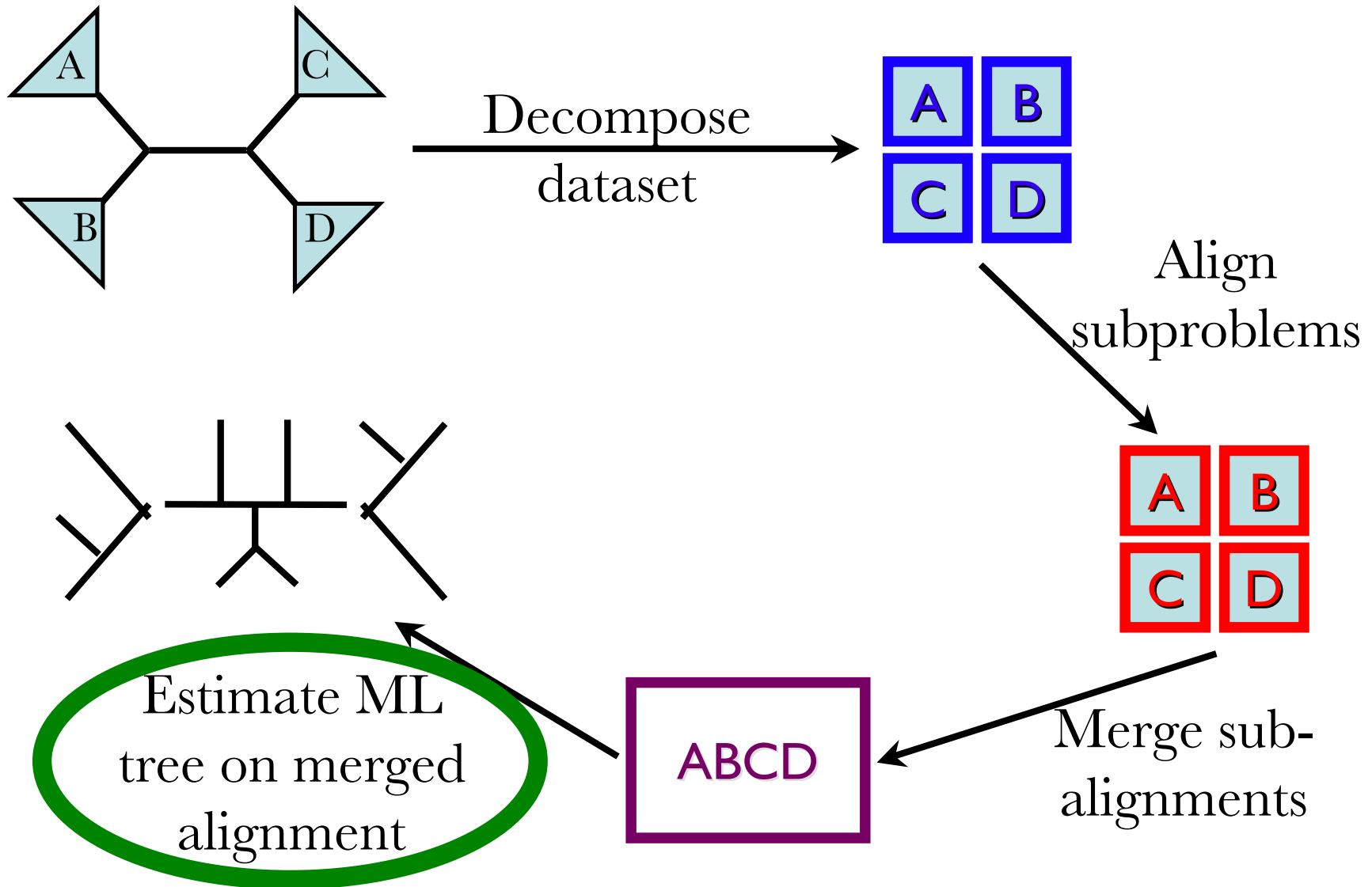
1000 taxon models, ordered by difficulty

1000 taxon models, ordered by difficulty

24 hour SATé analysis, on desktop machines

(Similar improvements for biological datasets)

1000 taxon models ranked by difficulty
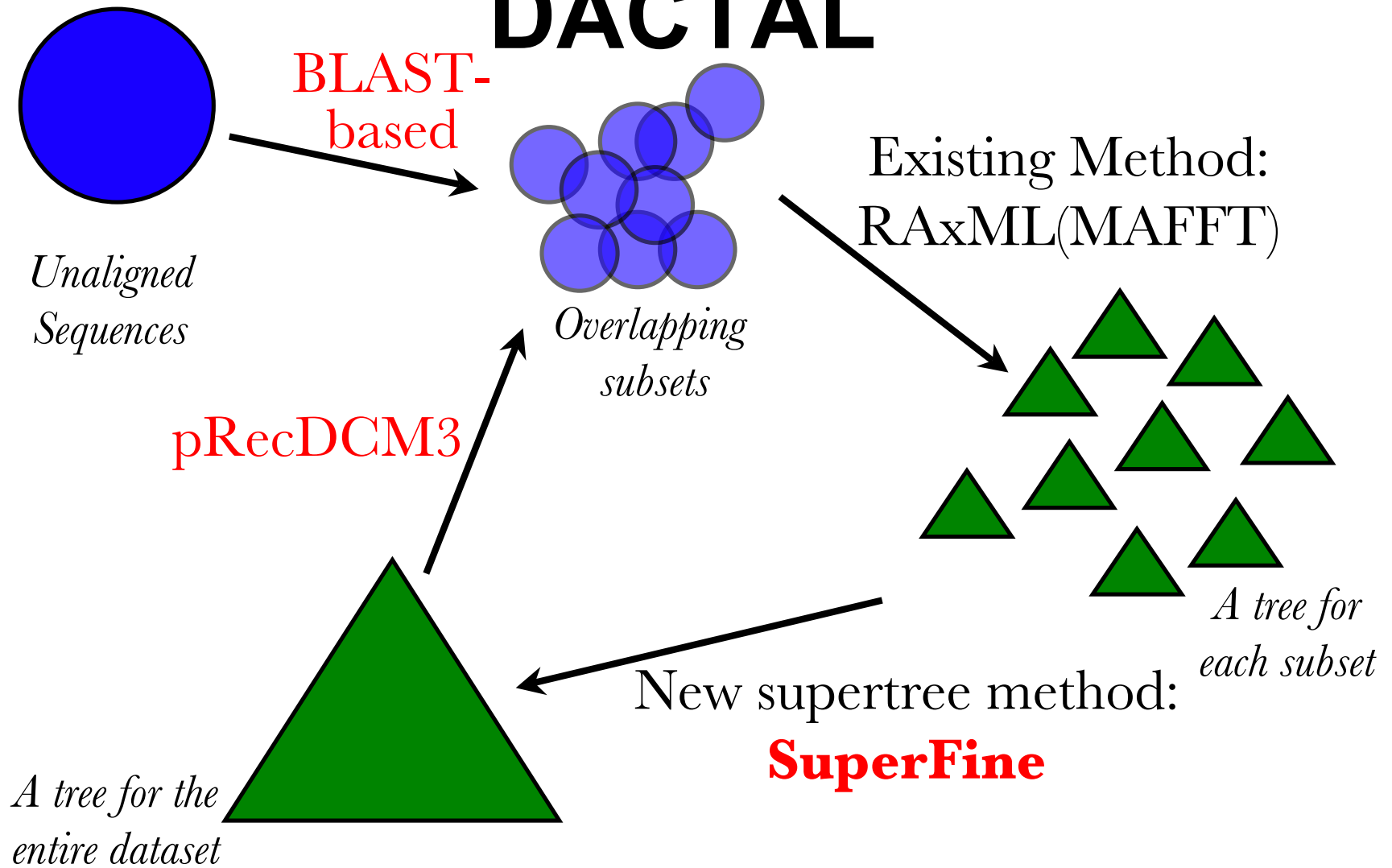
Limitations of SATé-I and -II

Part II: DACTAL
(Divide-And-Conquer Trees (Almost) without alignments)

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

(Nelesen, Liu, Wang, Linder, and Warnow, submitted)
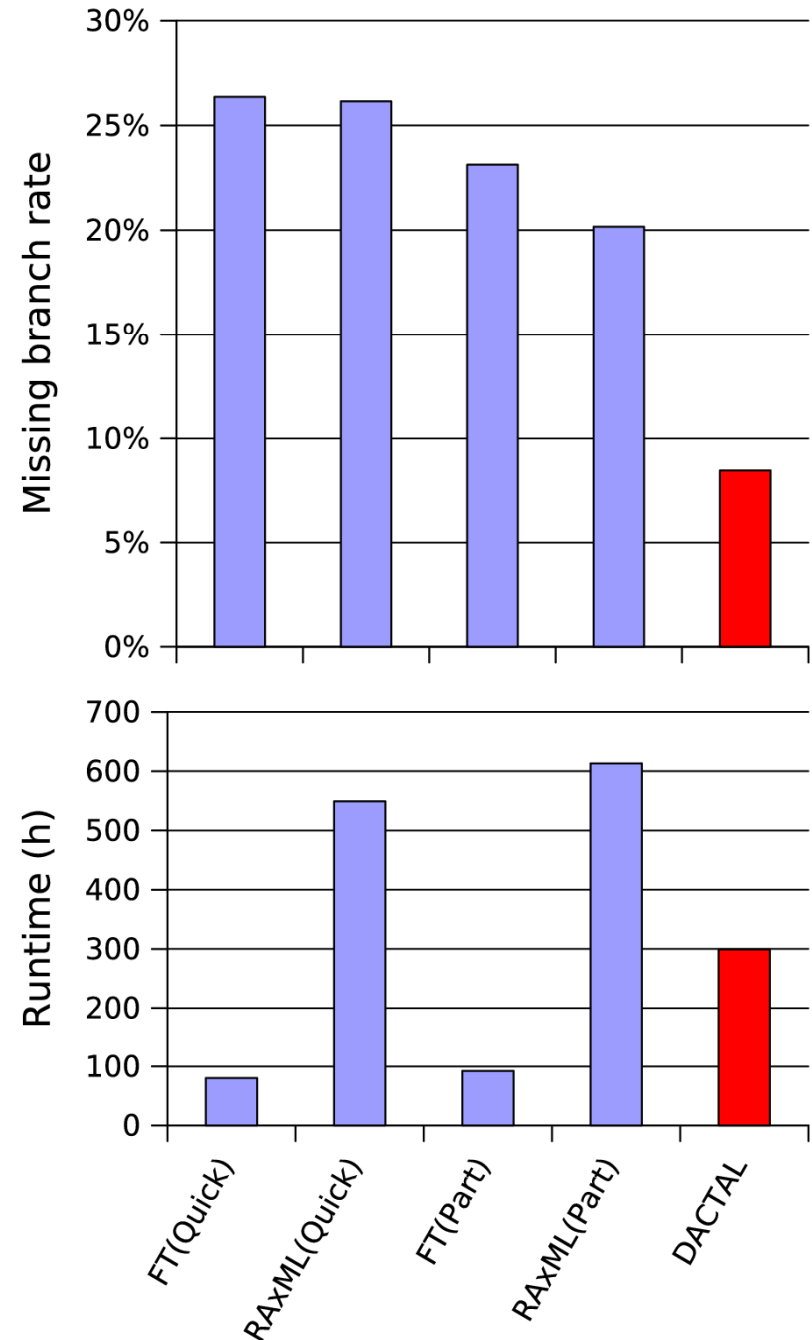
Average of 3 Largest CRW Datasets

CRW: Comparative RNA database,

Three 16S datasets with 6,323 to 27,643 sequences

Reference alignments based on secondary structure

Reference trees are 75% RAxML bootstrap trees

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

FastTree (FT) and RAxML are ML methods

Observations

- DACTAL gives more accurate trees than all other methods on the largest datasets

- DACTAL is much faster than SATé

- DACTAL is robust to starting trees and other algorithmic parameters

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

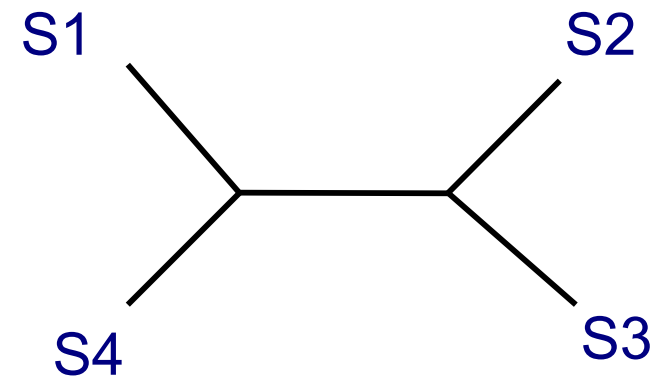- Place each query sequence into backbone tree, using extended alignment

Align Sequence

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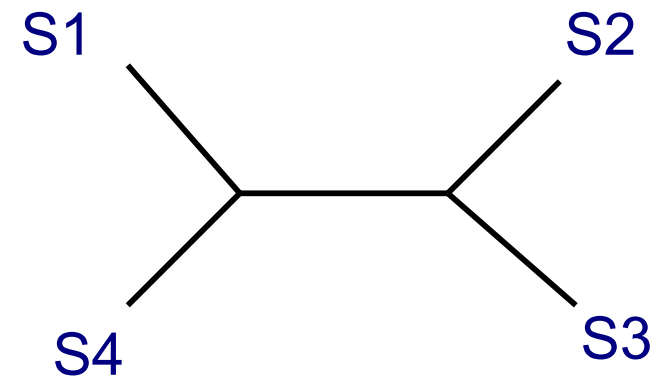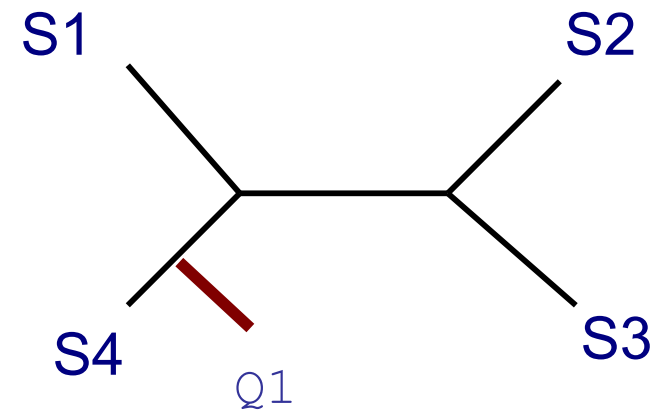
Align Sequence

```
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--------
```
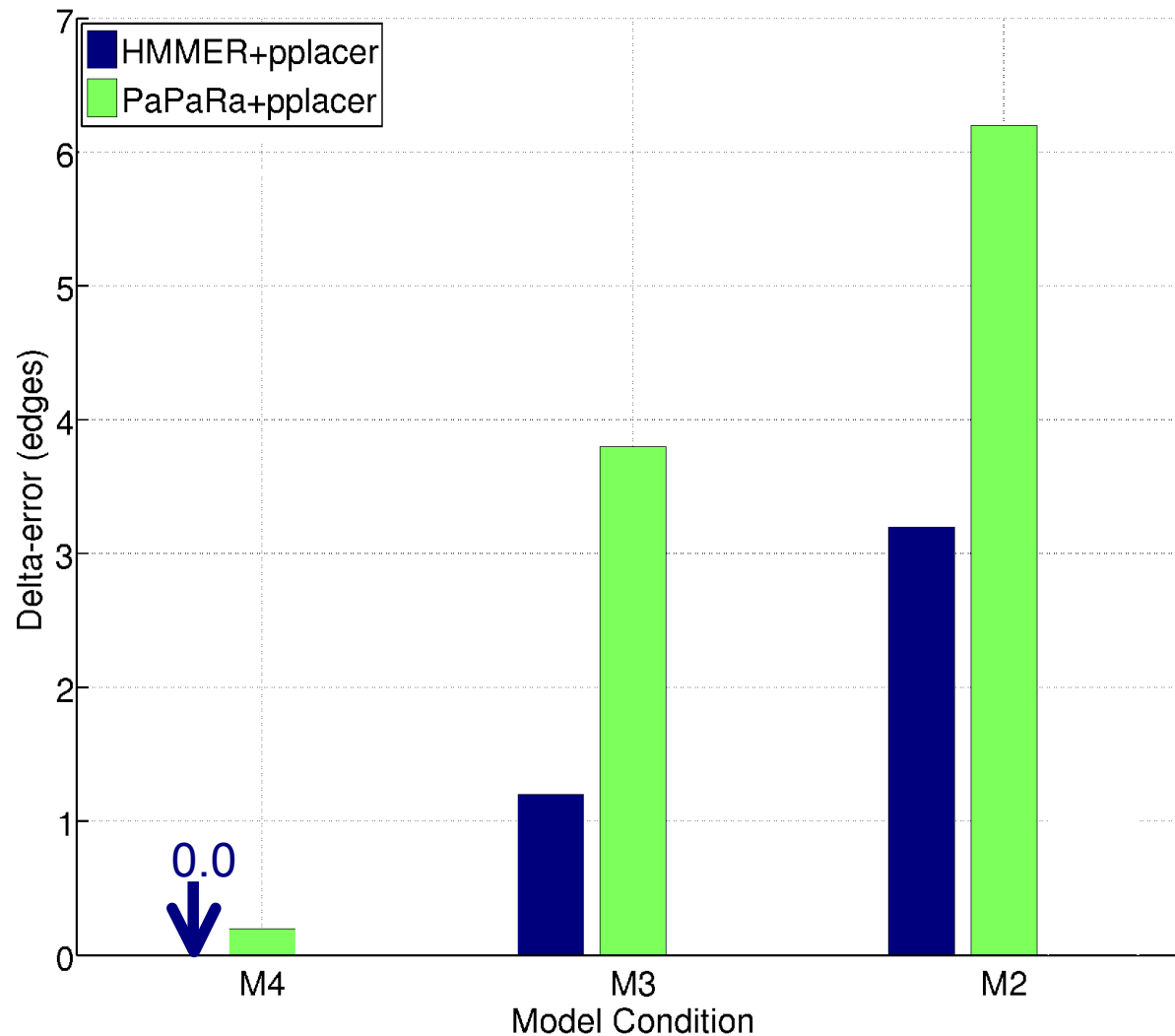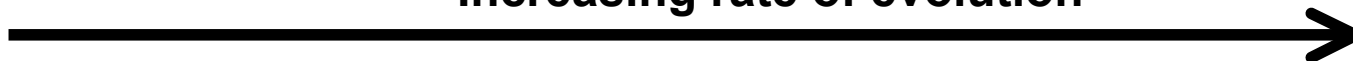
Phylogenetic Placement

- 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)
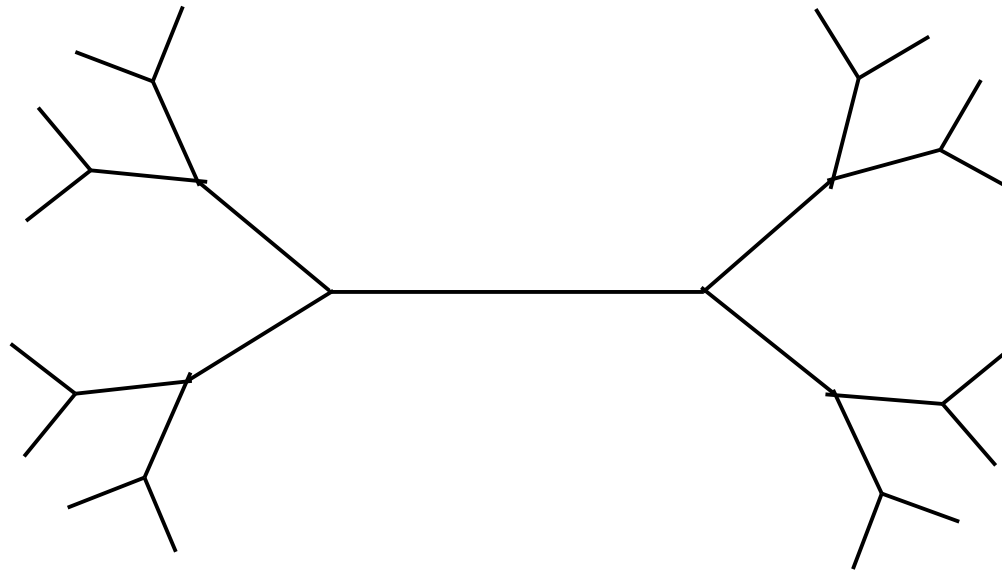
Note: pplacer and EPA use maximum likelihood

Insights from SATé
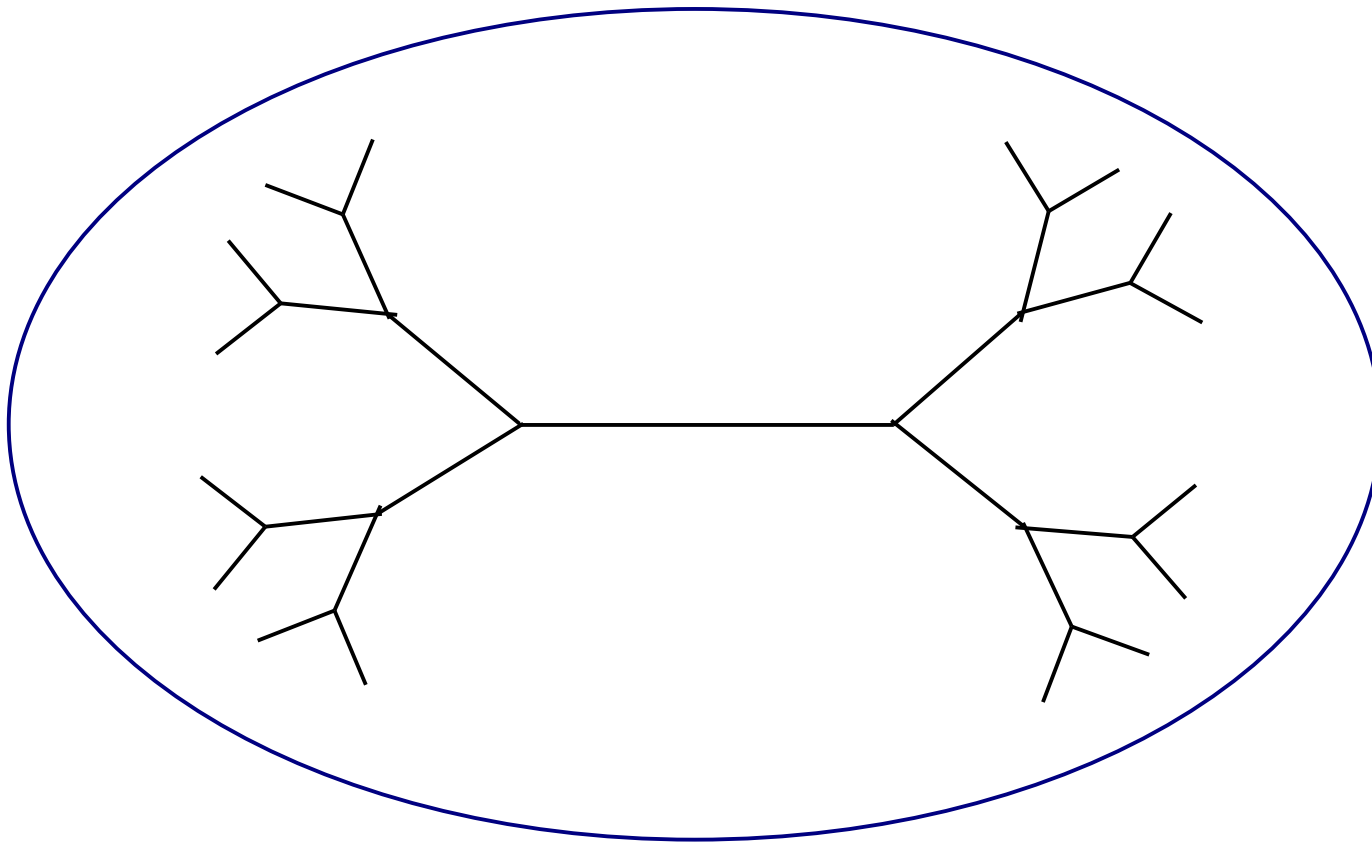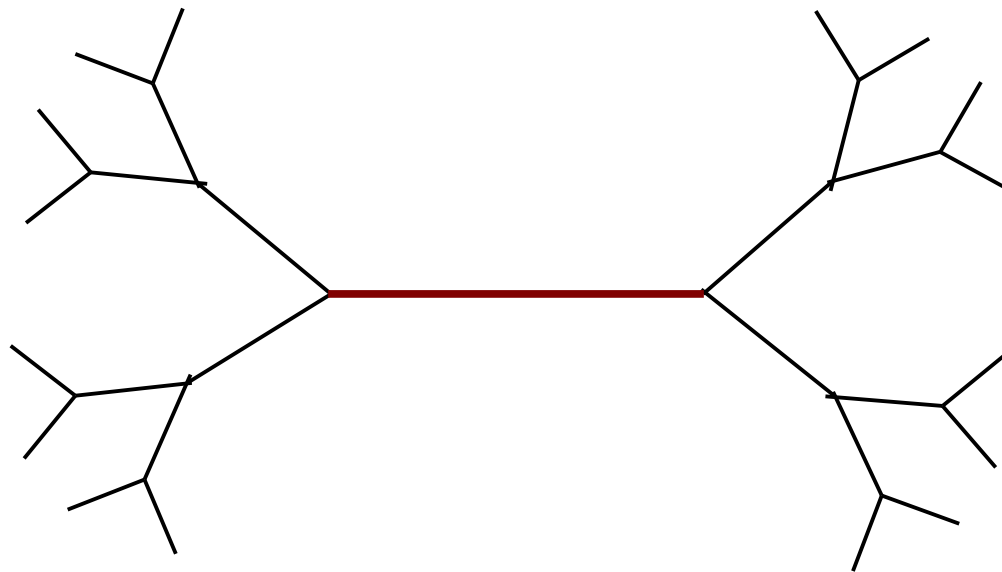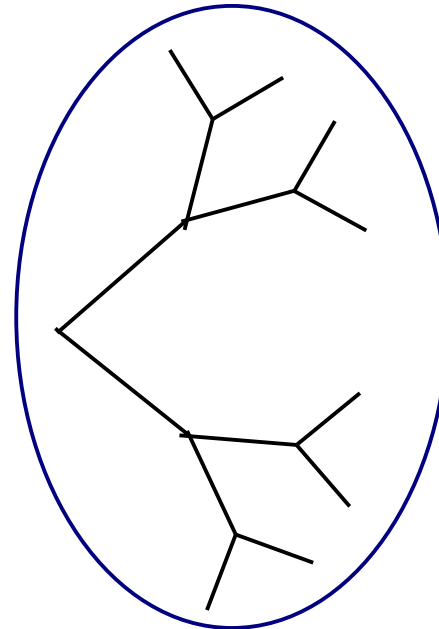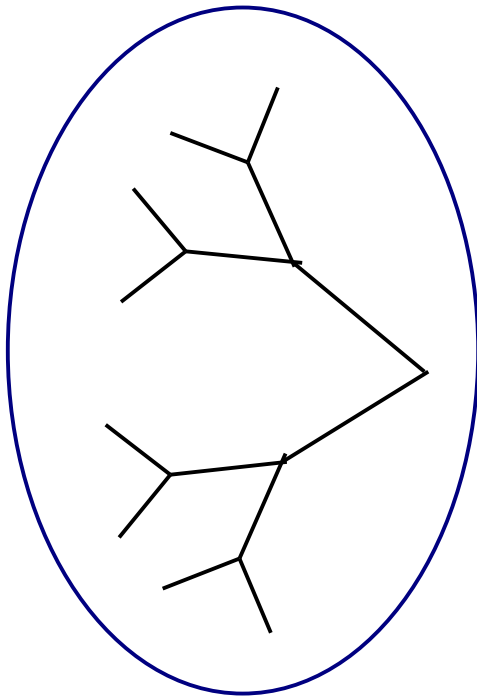
Insights from SATé

Insights from SATé

Insights from SATé

Insights from SATé
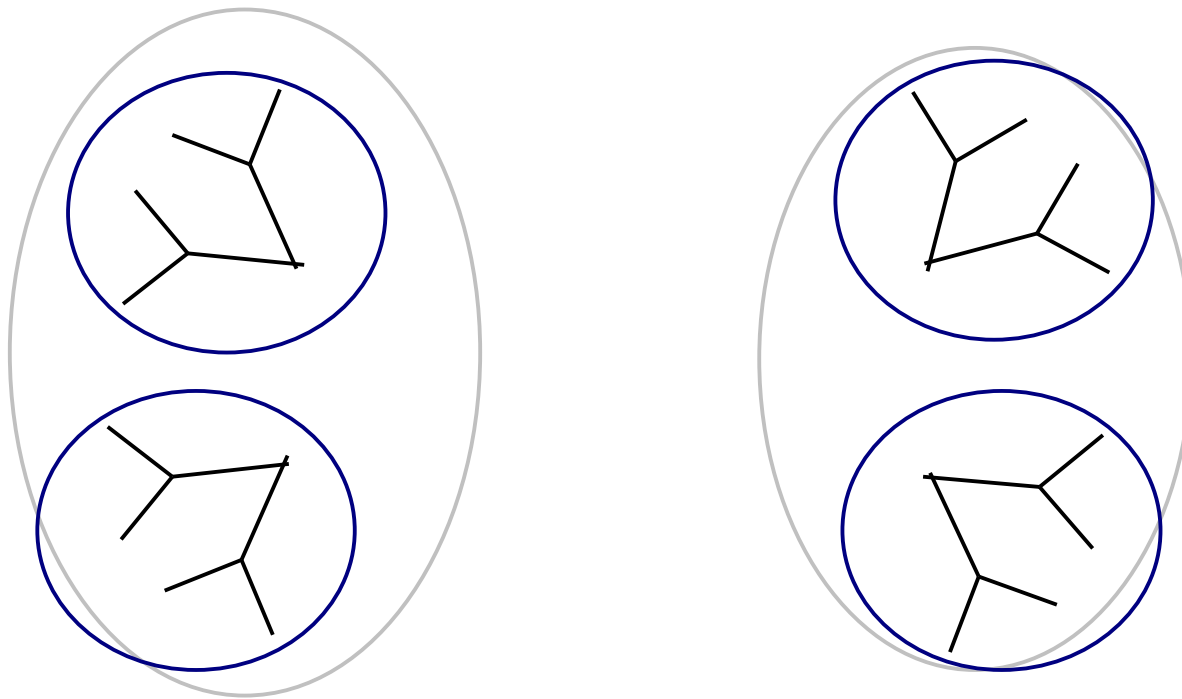
SEPP Parameter Exploration

- 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%) on Biological Data
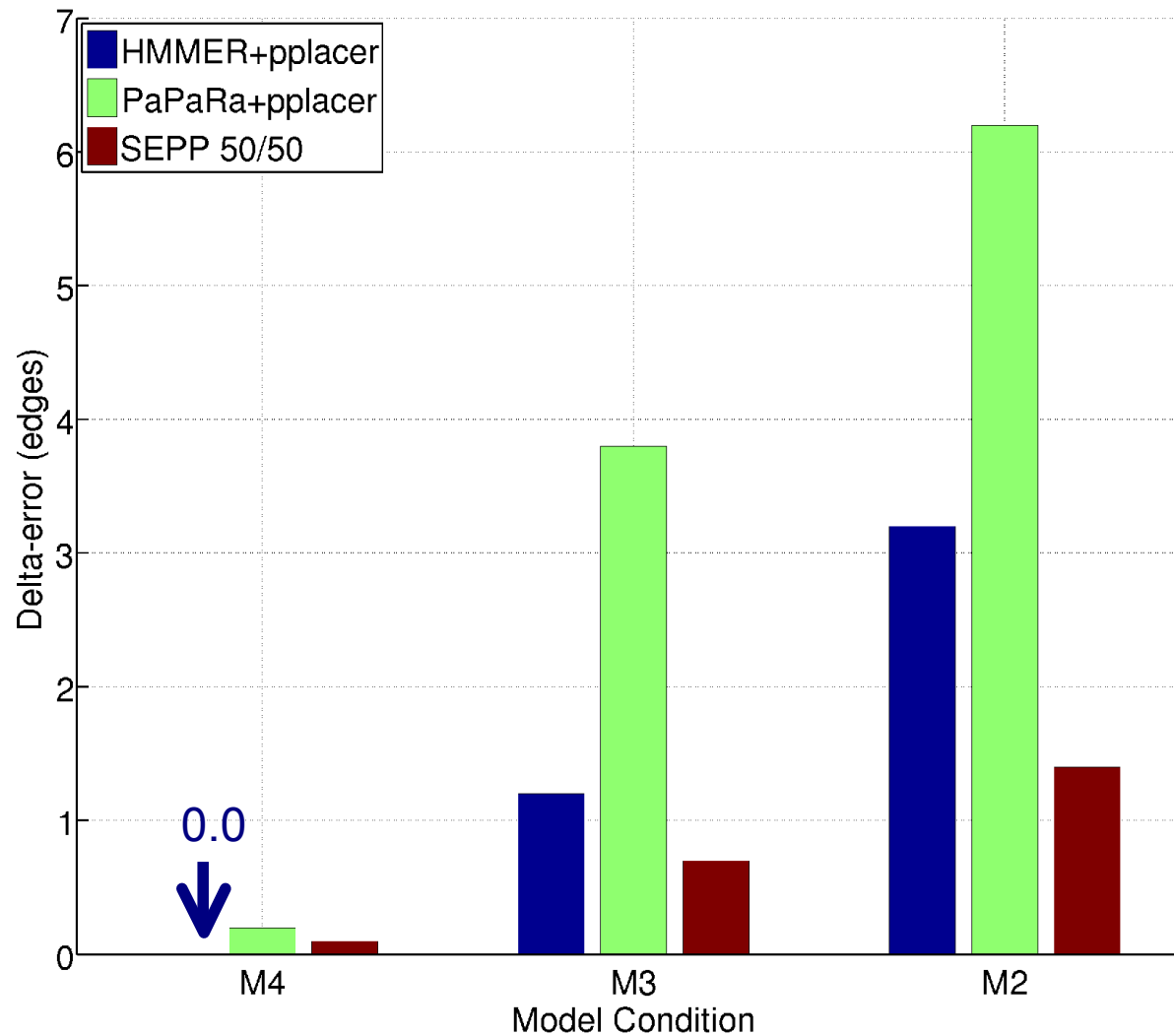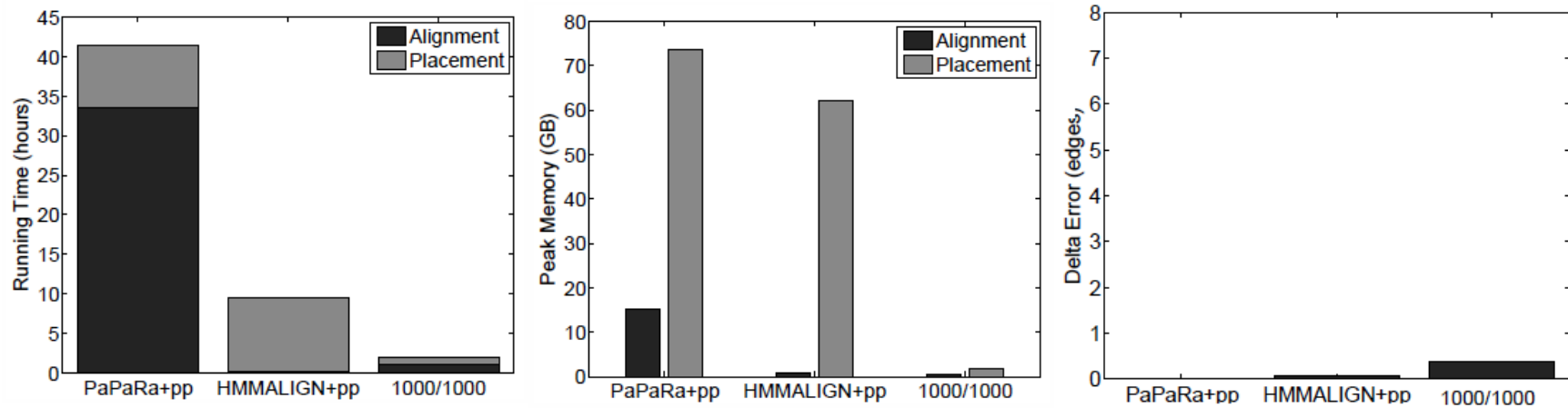
16S.B.ALL dataset, 13k curated backbone tree, 13k total fragments

SEPP (10%) on Biological Data
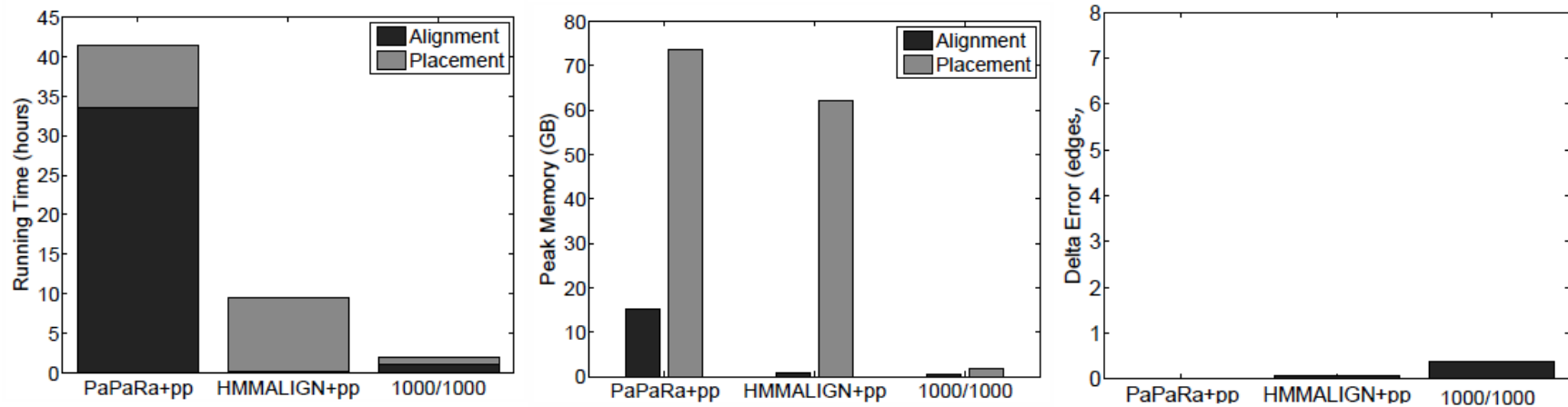
16S.B.ALL dataset, 13k curated backbone tree, 13k total fragments

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.

Acknowledgments

- Collaborators:

  - SATé: Kevin Liu, Serita Nelesen, Sindhu Raghavan, and Randy Linder

  - DACTAL: Serita Nelesen, Kevin Liu, Li-San Wang, and Randy Linder

  - SEPP: Siavash Mirarab and Nam Nguyen

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**

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**
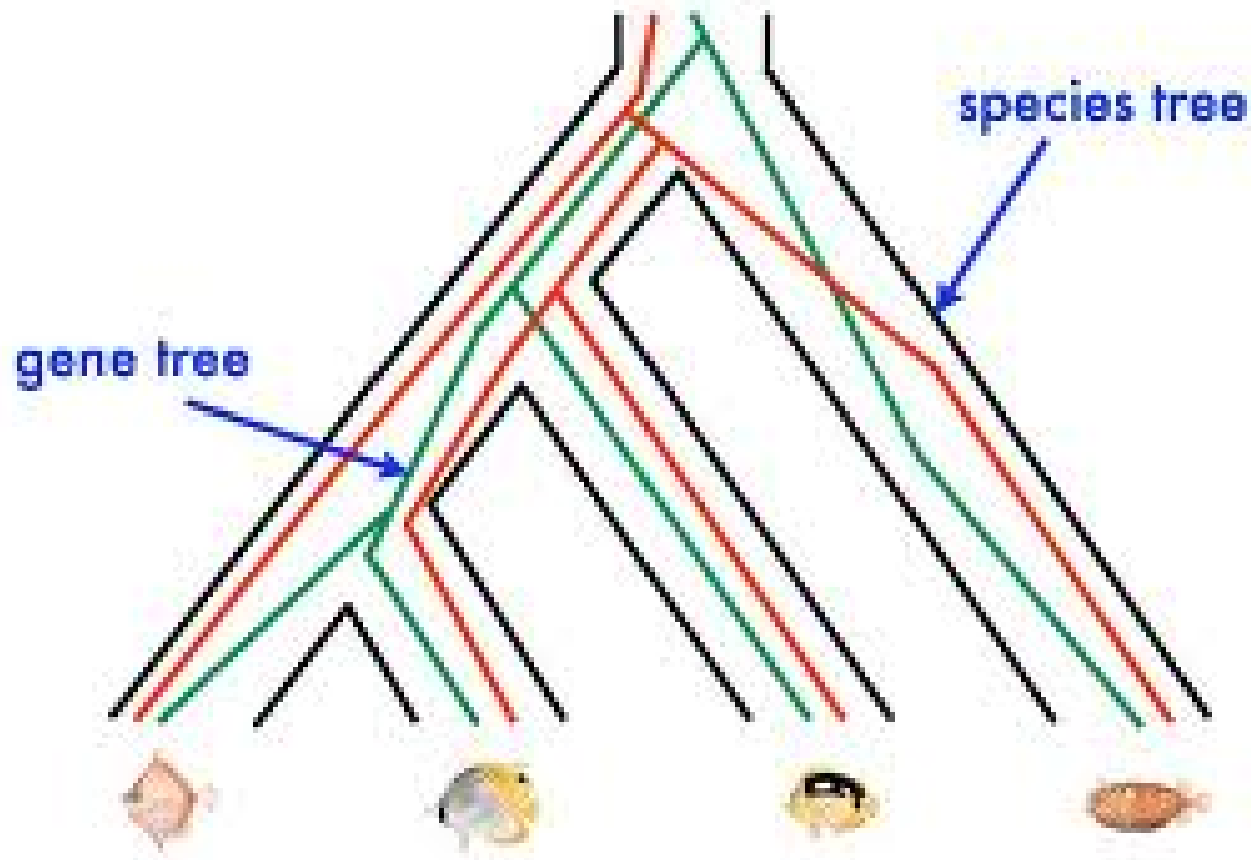- **Conifer evolution**

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.
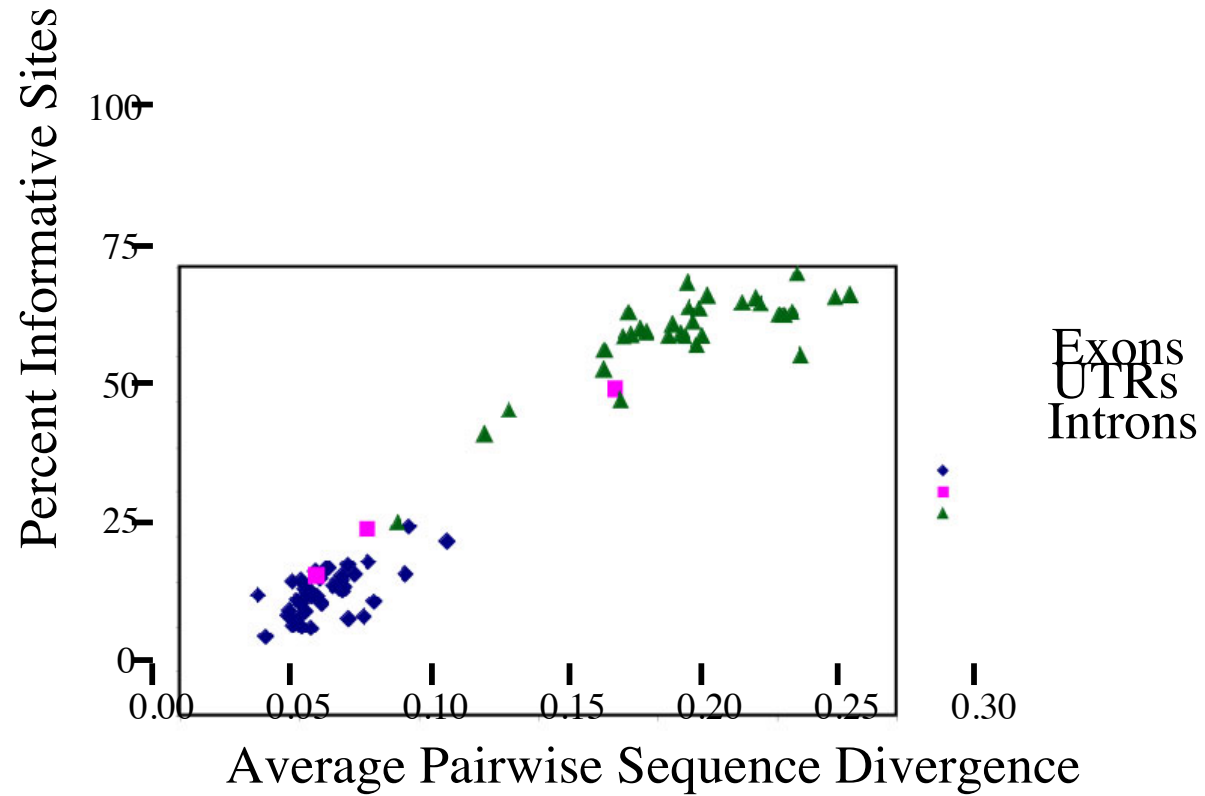
SATé-I vs. SATé-II

SATé-II

- Faster **and more accurate than** SATé-I

- **Longer analyses or use of ML to select tree/alignment pair slightly better results**
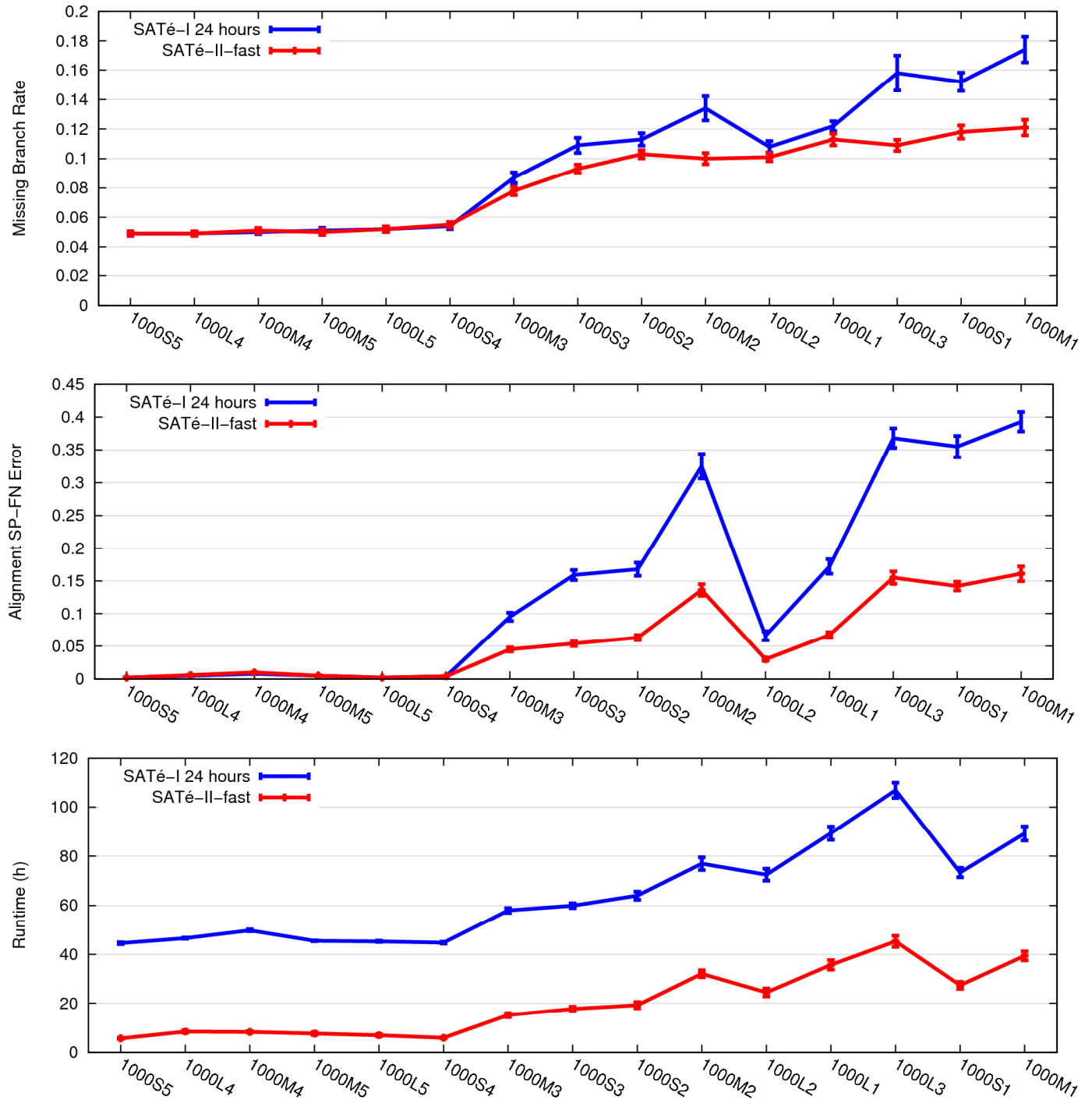
Divergence & Information Content

Analysis and figure provided by Mike Braun
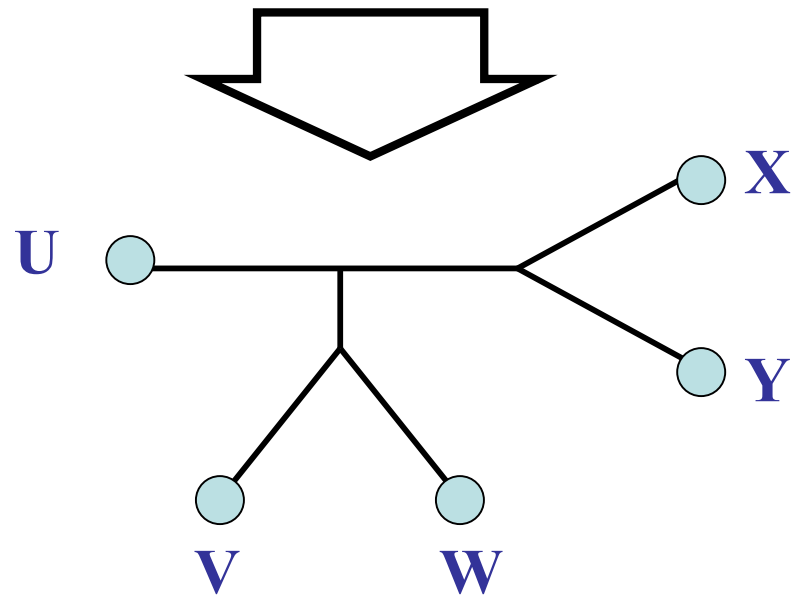Smithsonian Institution

Reticulate evolution

- Not all evolution is tree-like:
    - Horizontal gene transfer
    - Hybrid speciation
- How can we detect reticulate evolution?

Grading

- Homeworks 50%

- Final exam 30%

- Class participation 10%

- Class project 10%

U AGGGCAT   V TAGCCCA   W TAGACTT   X TGCACAA   Y TGCGCTT

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
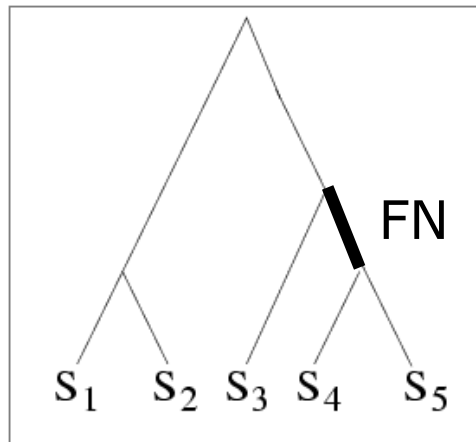- FastME
- UPGMA
- Quartet puzzling
- Etc.

RAxML: heuristic for large-scale ML optimization

Software

In use by research groups around the world

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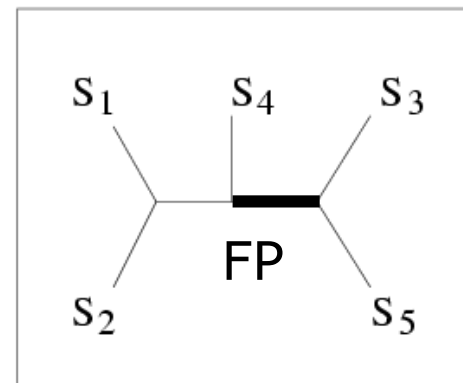- **Downloadable software for various platforms**
- **Easy-to-use GUI**
- **http://phylo.bio.ku.edu/software/sate/sate.html**

Quantifying Error

TRUE TREE

DNA SEQUENCES

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

50% error rate

INFERRED TREE

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.**

Biology: 21st Century Science!

"When the human genome was sequenced seven years ago, scientists knew that most of the major scientific discoveries of the 21st century would be in biology."

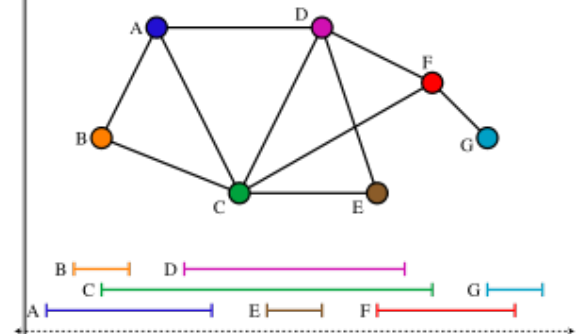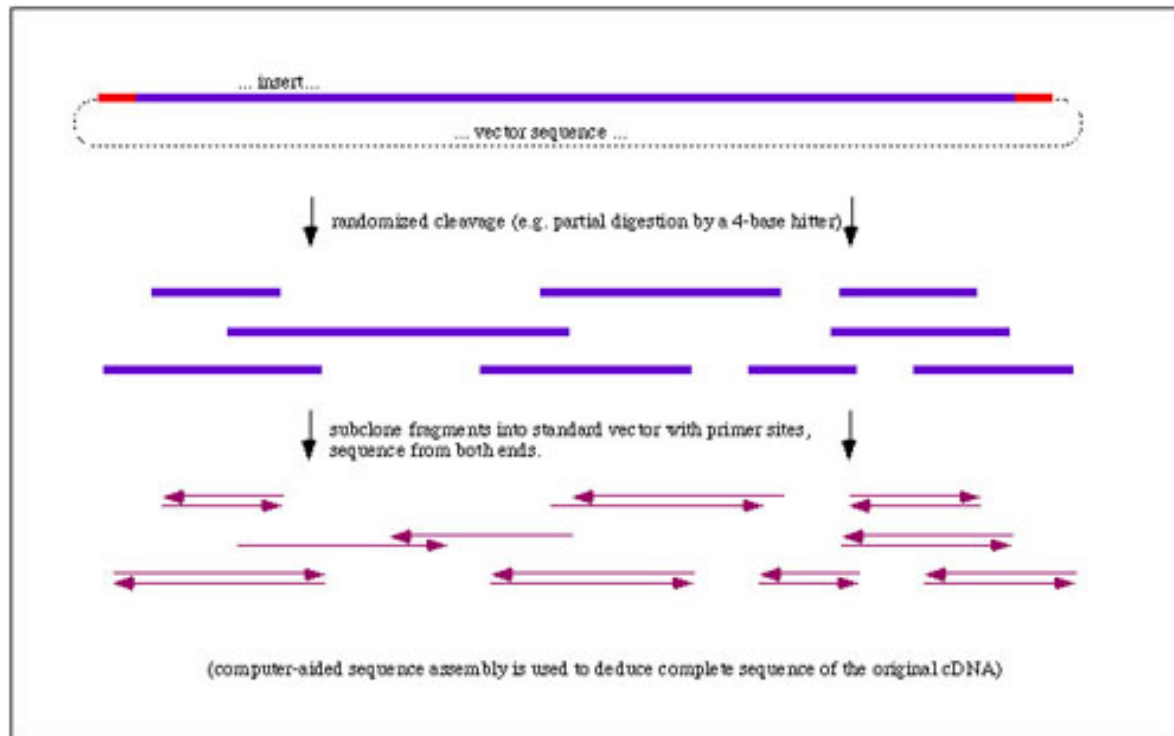January 1, 2008, guardian.co.uk

Genome Sequencing Projects:

Started with the Human Genome Project

Whole Genome Sequencing:

Graph Algorithms and Combinatorial Optimization!

(computer-aided sequence assembly is used to deduce complete sequence of the original cDNA)

Other Genome Projects! (Neandertals, Wooly Mammoths, and more ordinary creatures…)

Phylogeny (evolutionary tree)

Orangutan   Gorilla   Chimpanzee   Human
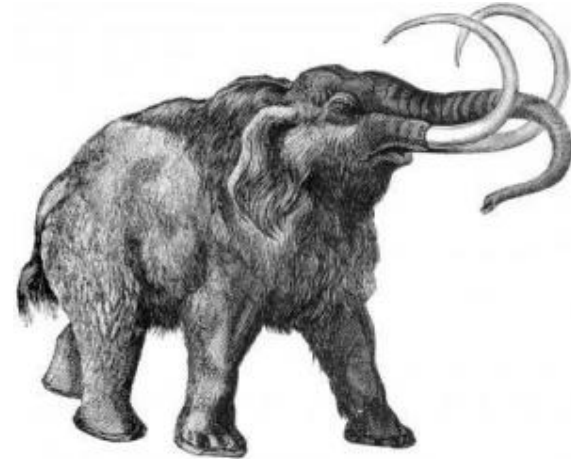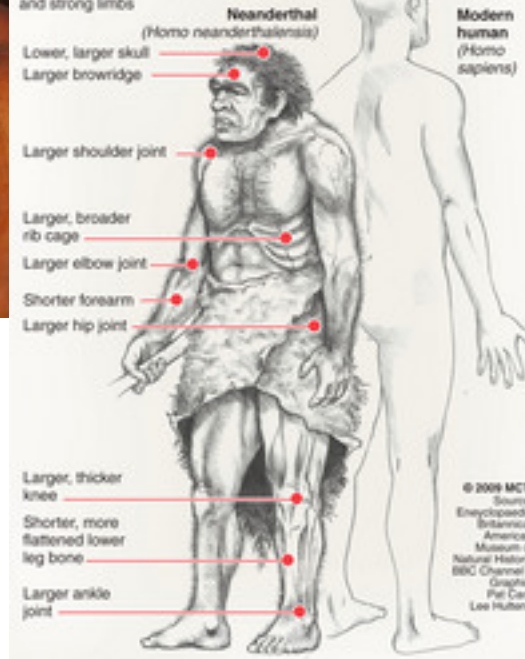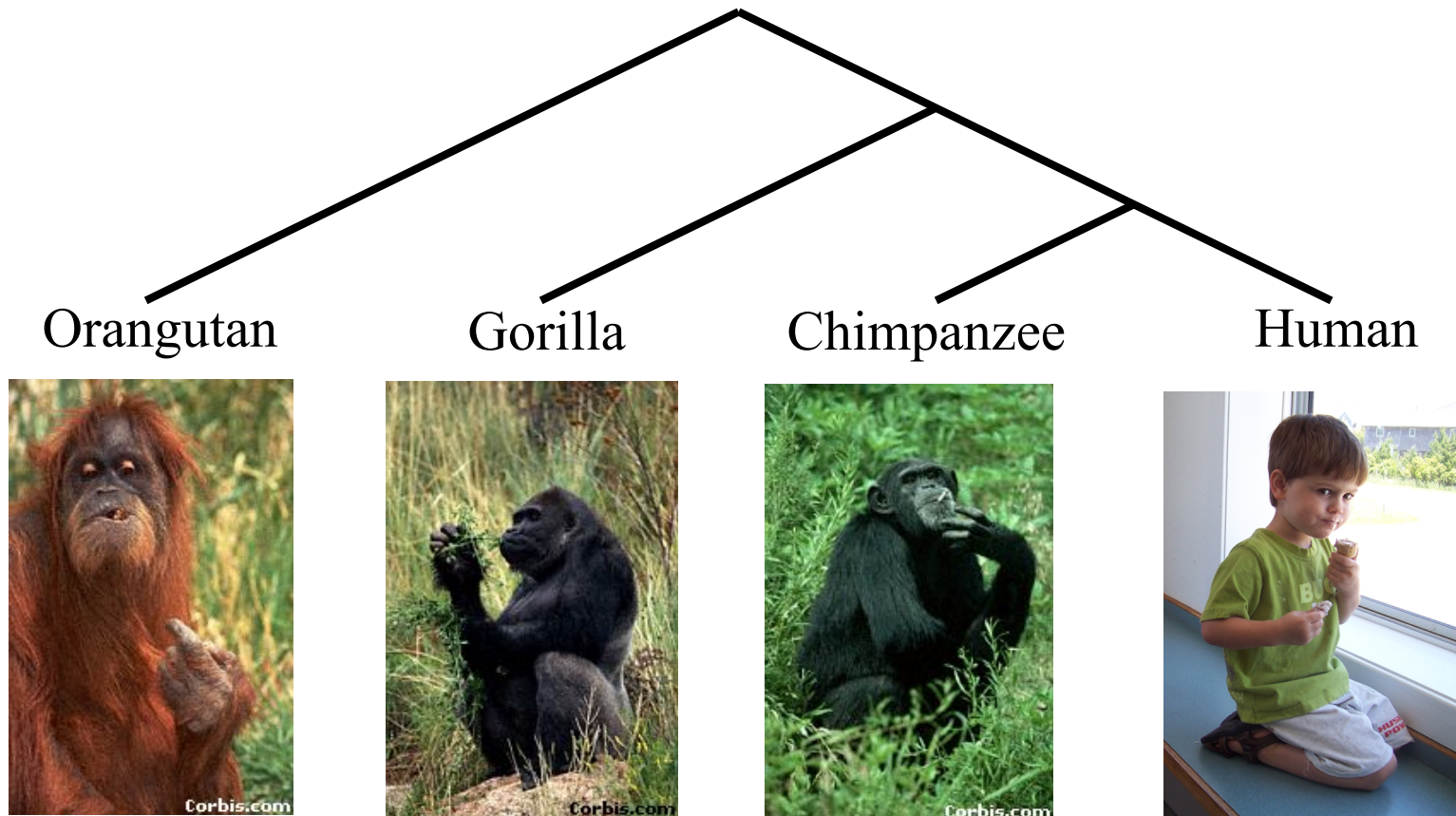
From the Tree of the Life Website,
University of Arizona

Metagenomics

- Input: set of sequences
- Output: a tree on the set of sequences, indicating the species identification of each sequence

- Issue: the sequences are not globally alignable, and there are often thousands (or more) of the sequences