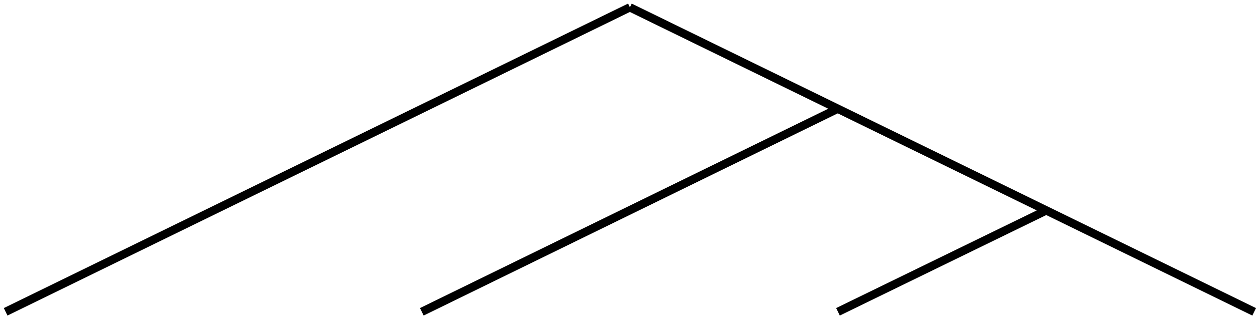
Recent Breakthroughs (and **Current Challenges**) in Computational Phylogenetics

Tandy Warnow

Department of Computer Science

University of Texas

Phylogeny (evolutionary tree)

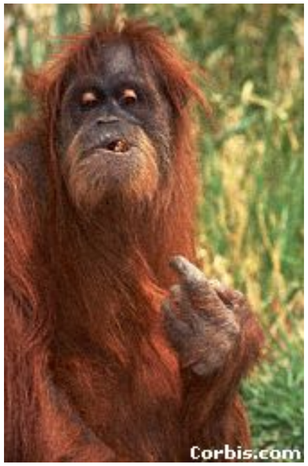


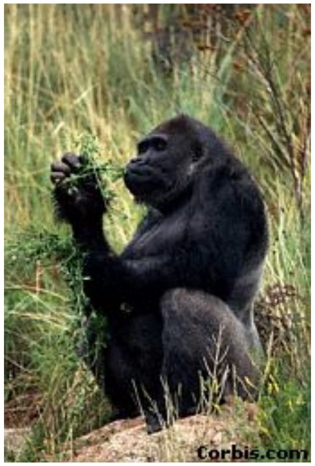
Orangutan

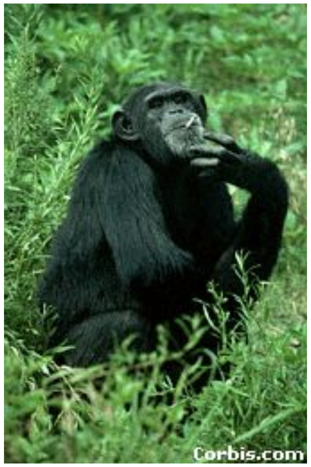
Gorilla

Chimpanzee

Human

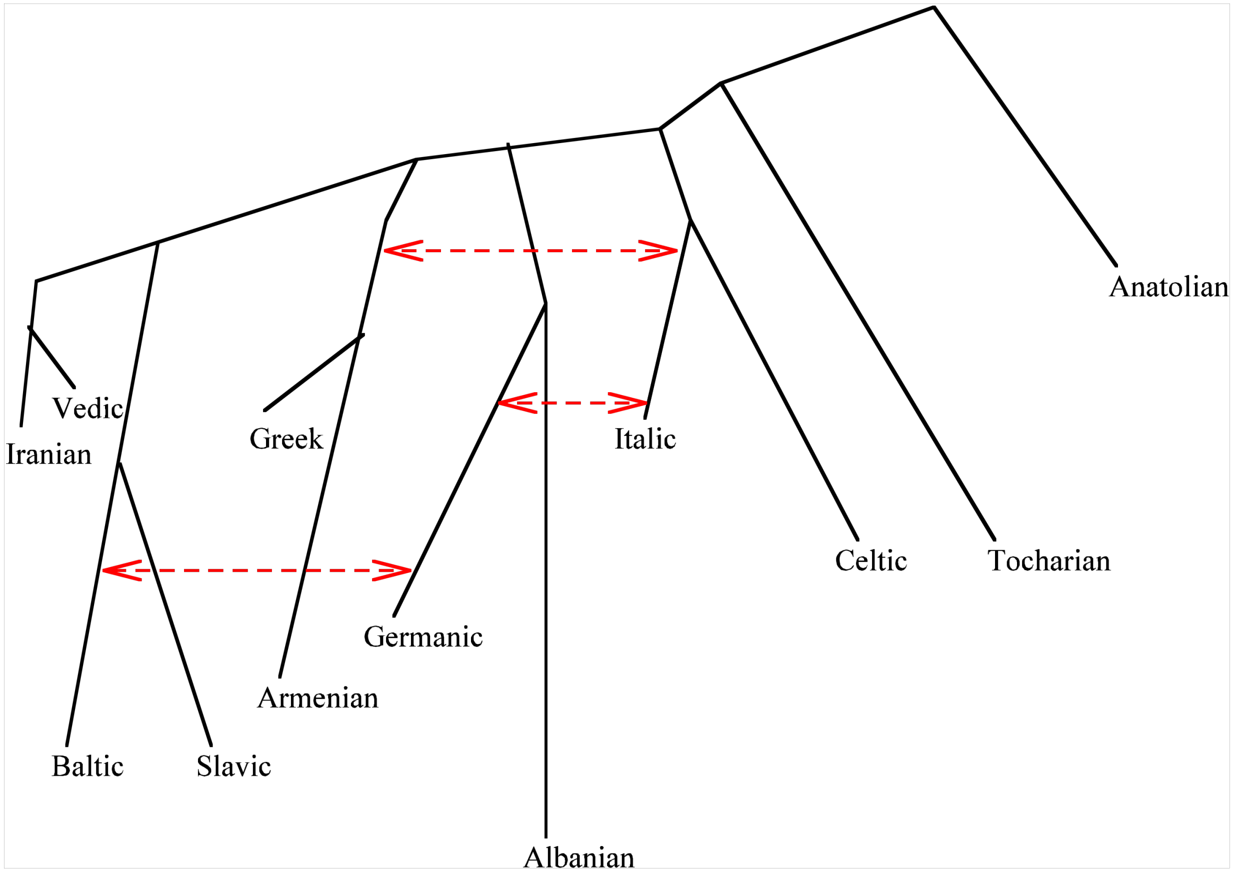






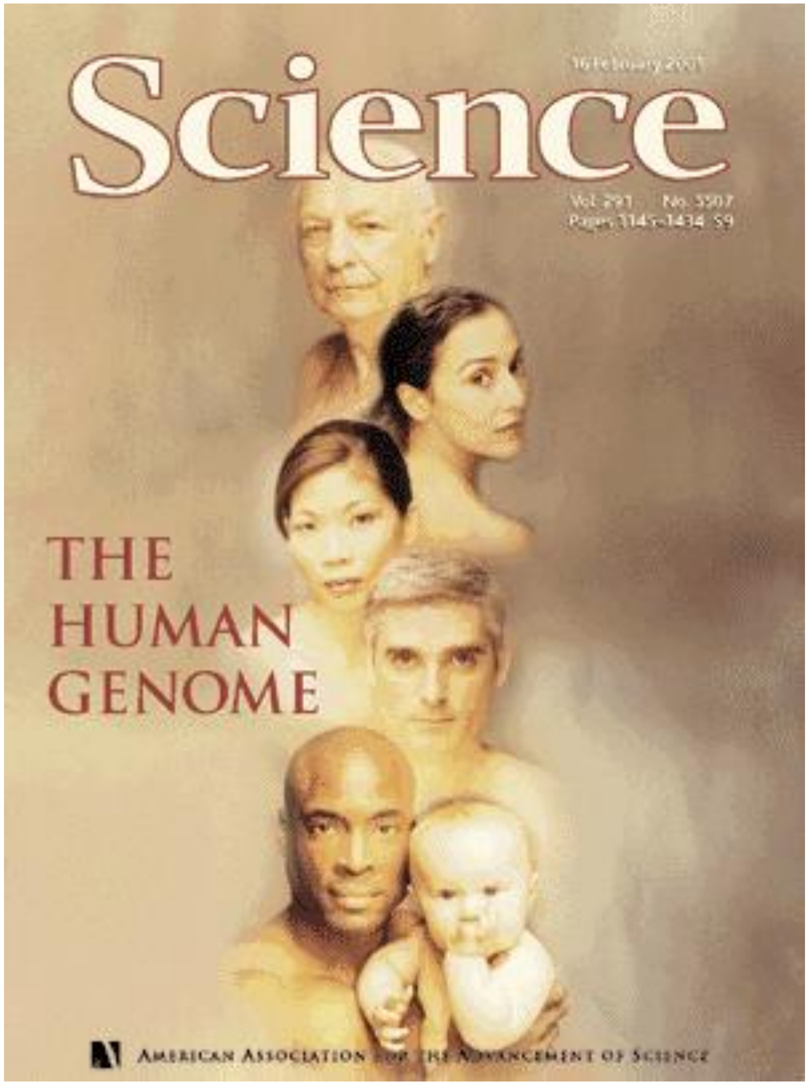
From the Tree of the Life Website,
University of Arizona

Indo-European Phylogeny
Nakhleh et al., Language, 2005

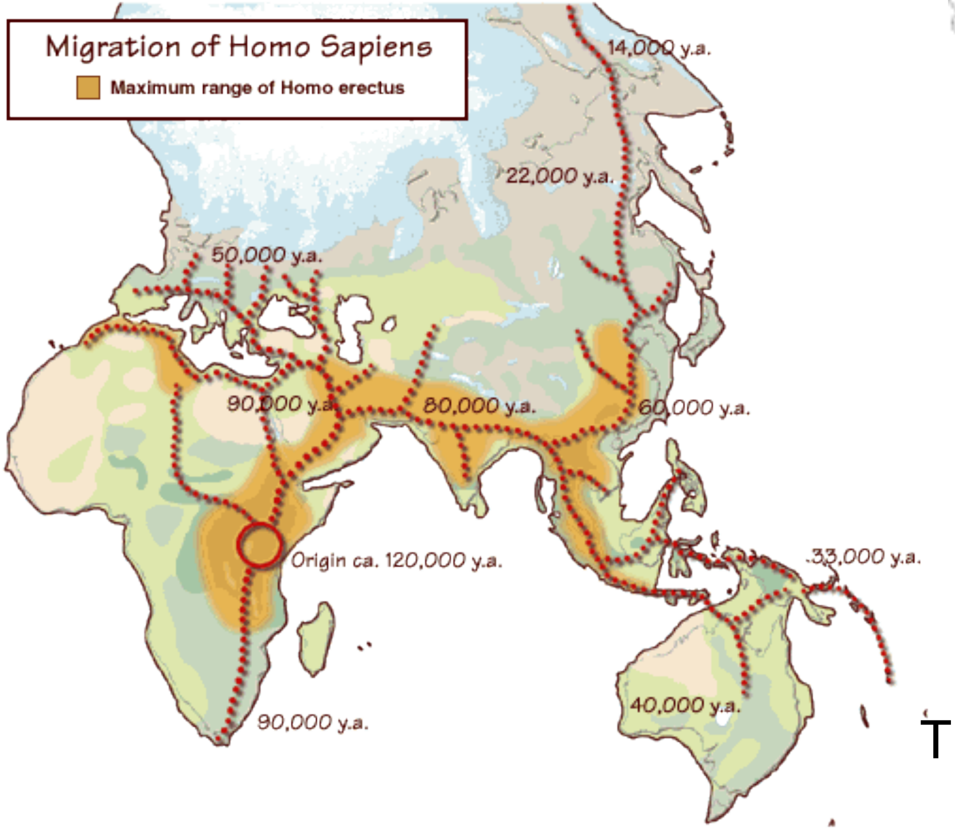


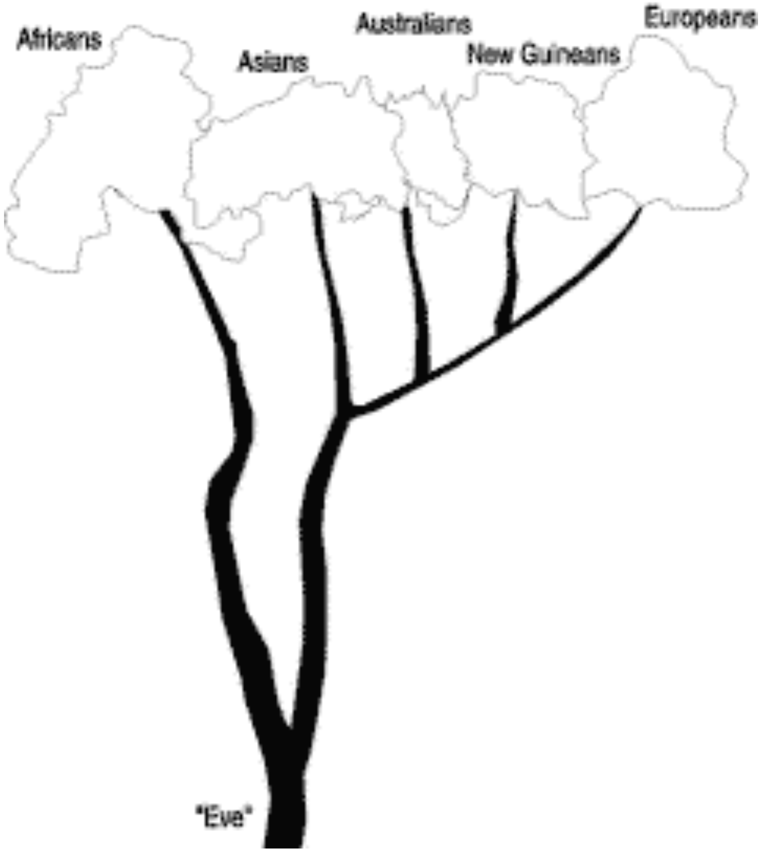
Genome Sequencing Projects:

Started with the Human Genome Project



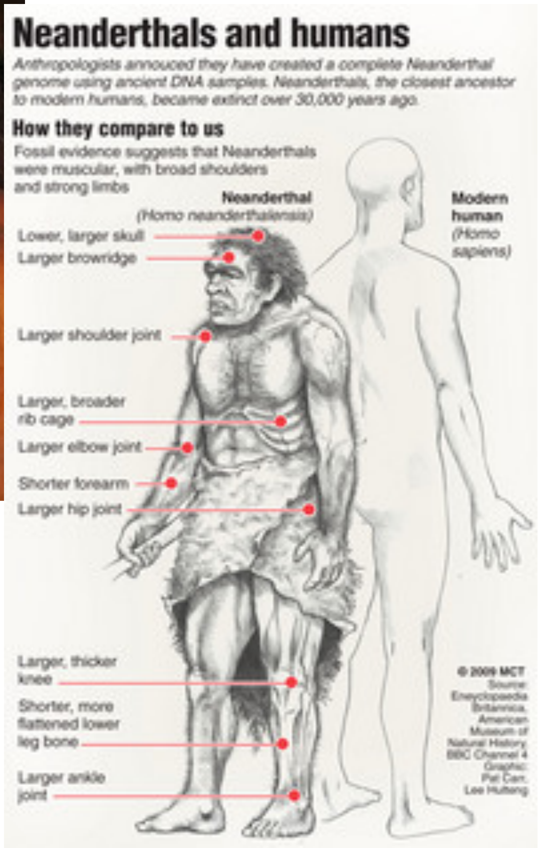
Where did humans come from, and how did they move throughout the globe?





The Mitochondrial “African Eve”

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



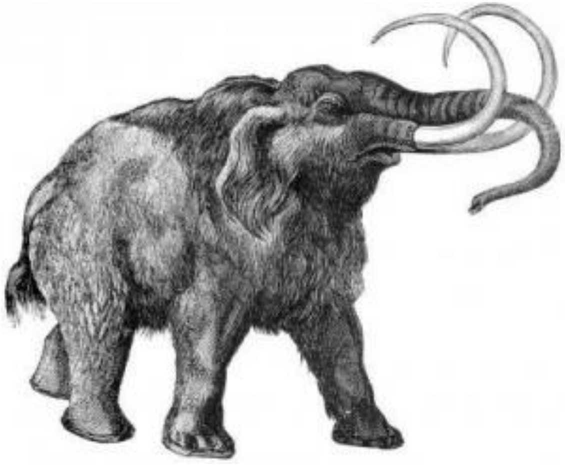
Neanderthals and humans

Anthropologists annouced they have created a complete Neanderthal genome using ancient DNA samples. Neanderthals, the closest ancestor to modern humans, became extinct over 30,000 years ago.

How they compare to us

Fossil evidence suggests that Neanderthals were muscular, with broad shoulders and strong limbs

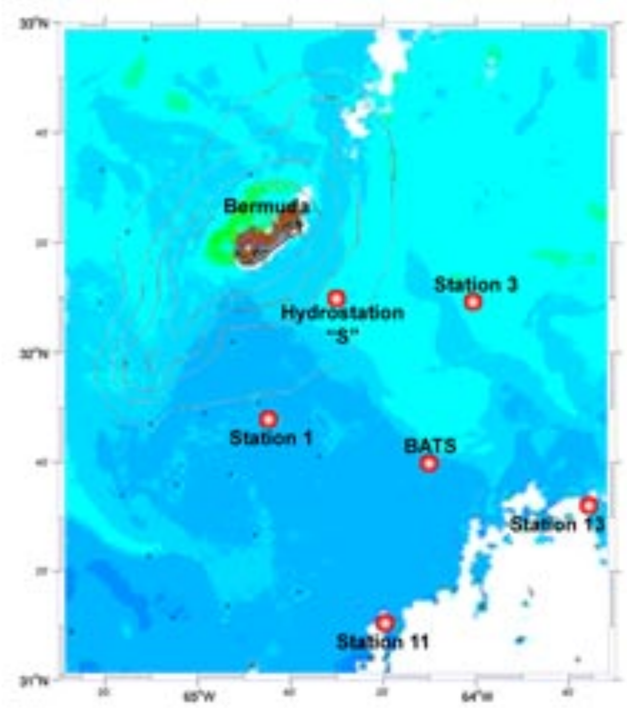
© 2009 MCT
Source: Encyclopaedia Britannica, American Museum of Natural History, BBC Channel 4
Graphic: Pat Carr, Lee Hulteng



Metagenomics:

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

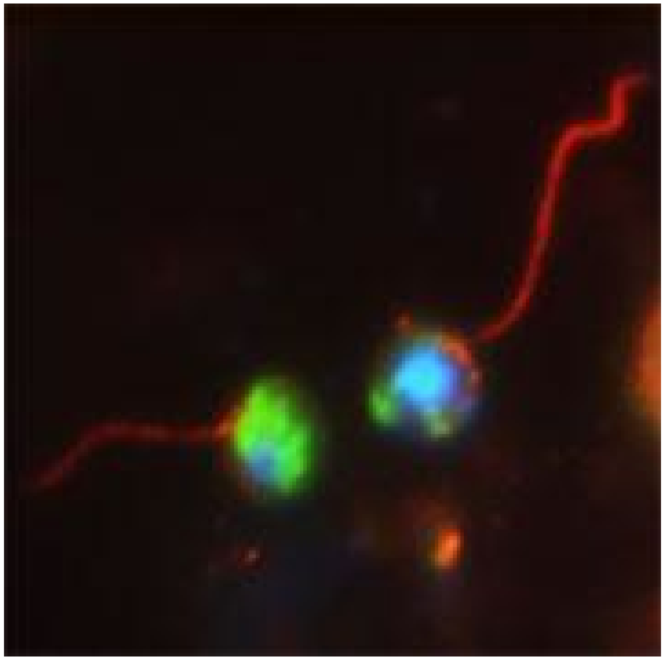
Scientists Discover One Million New Genes in Ocean Microbes



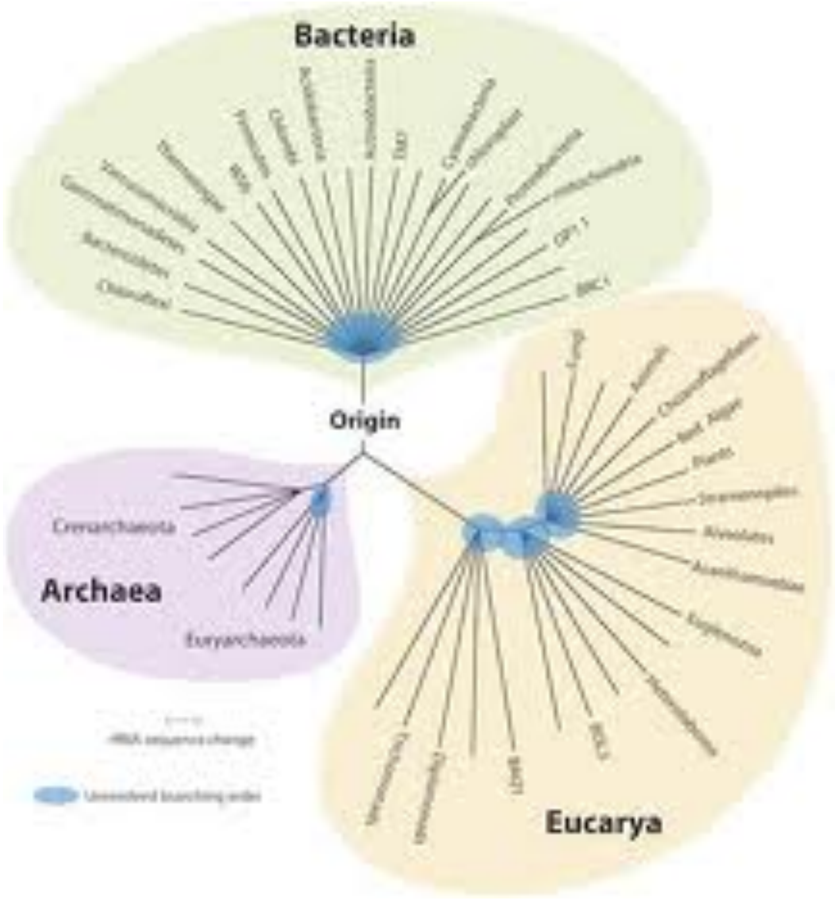
Metagenomic analyses: discovery of new species!

Two cryptomycota cells found in water samples collected from the University of Exeter pond.

Jones et al., Nature 2011.

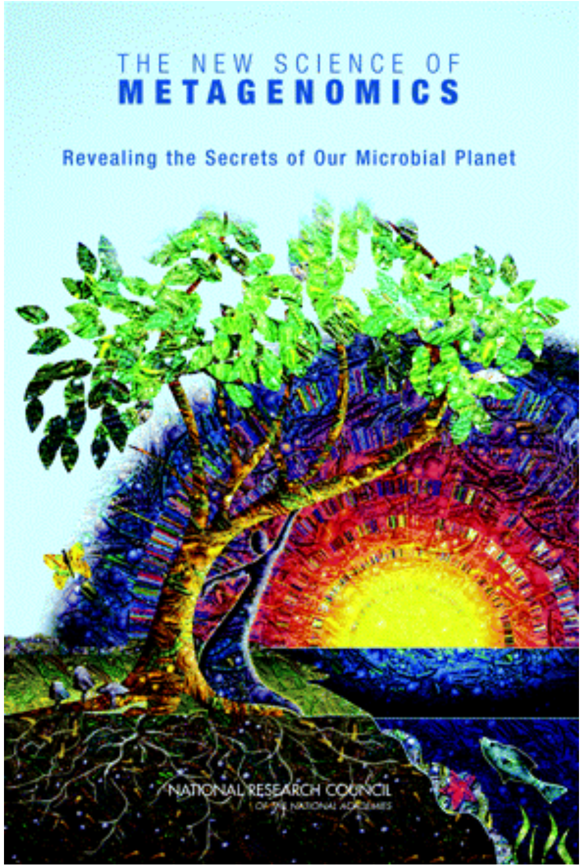


How did Life Evolve?



Computational Phylogenetics and Metagenomics

Courtesy of the Tree of Life project



THE NEW SCIENCE OF
METAGENOMICS

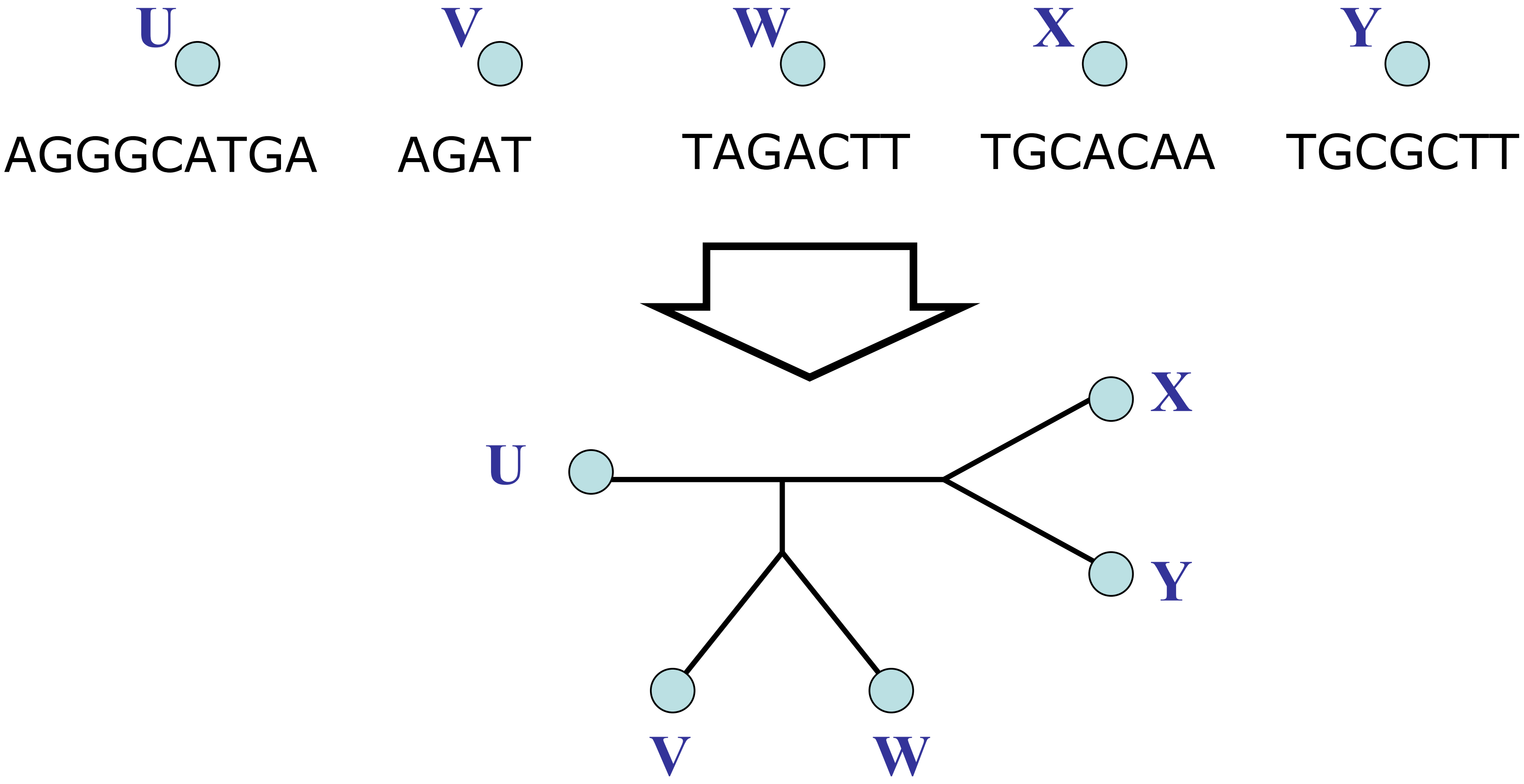
Revealing the Secrets of Our Microbial Planet

NATIONAL RESEARCH COUNCIL
OF THE NATIONAL ACADEMIES

Phylogenetic Estimation

Math/CS/Stat:

- NP-hard problems, finding good solutions can take months or years for single datasets
- Many optimal solutions for each analysis (data mining!)
- Mathematical modelling of evolutionary processes
- Probabilistic analysis of algorithms
- High performance computing
- Extensive simulation studies
- Real data analyses


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

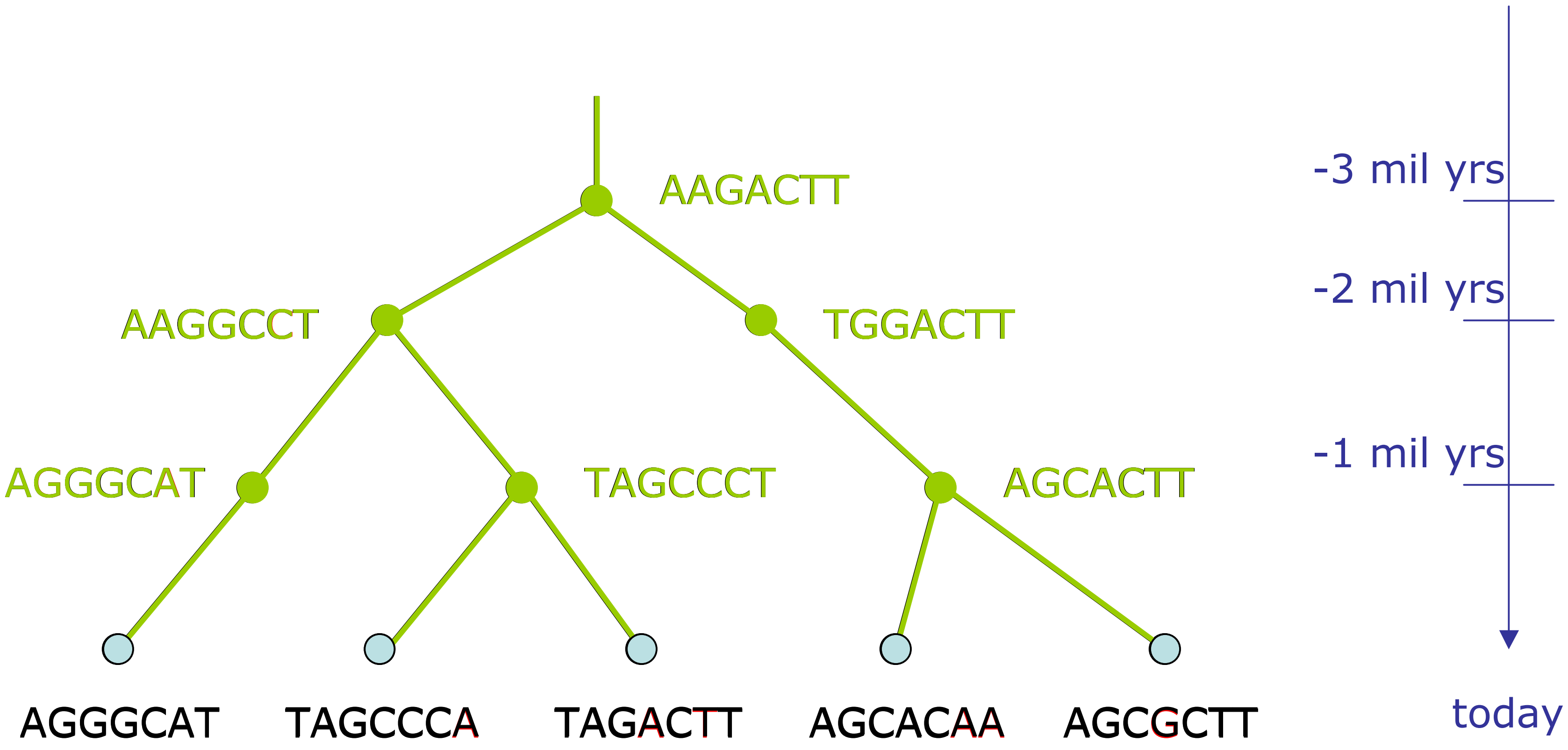
Today’s Talk

- SATé: Simultaneous alignment and tree estimation (Liu et al., Science 2009, and Systematic Biology 2012)
- DACTAL: divide-and-conquer trees (almost) without alignments (Nelesen et al., submitted)

Part I: SATé

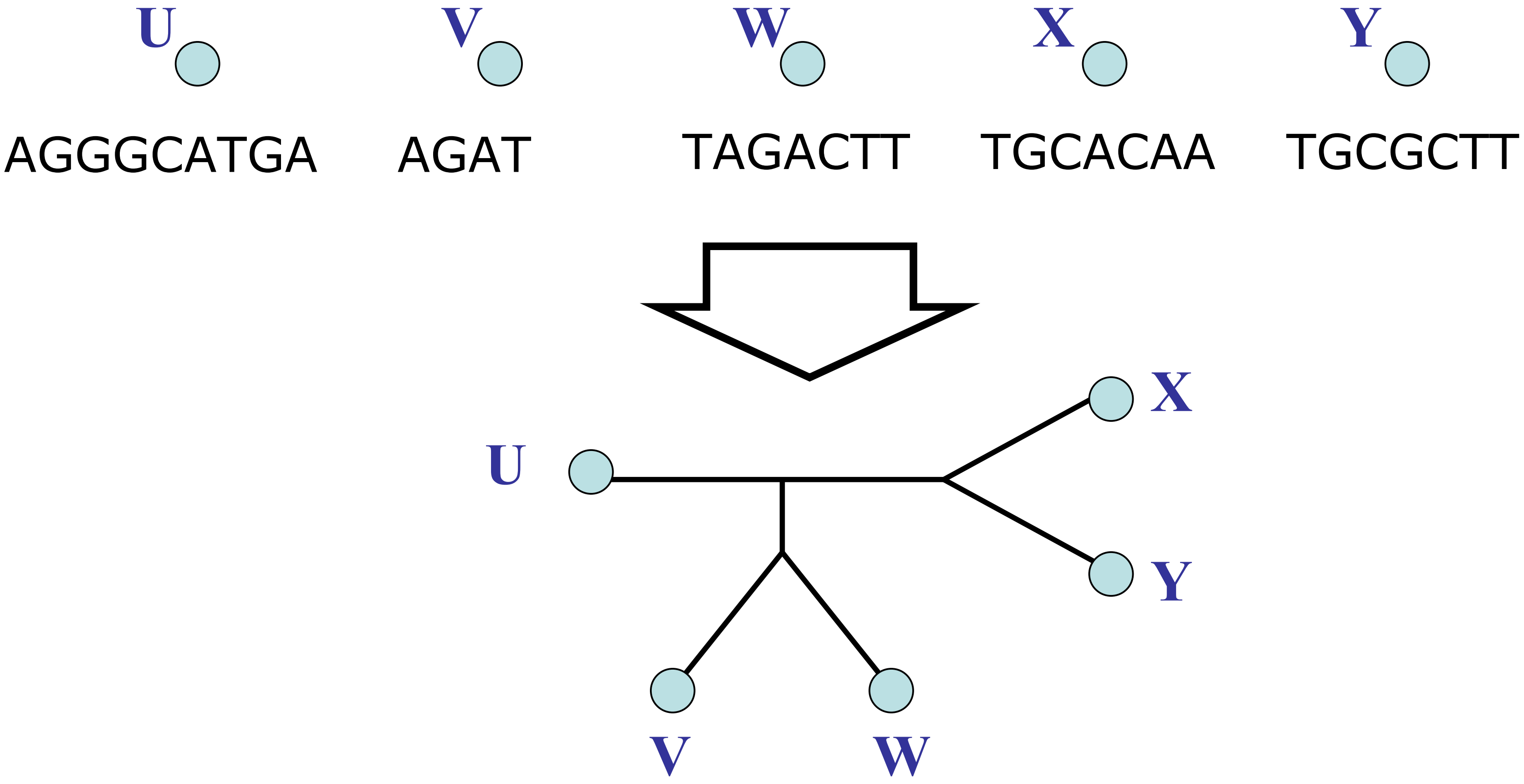
- Simultaneous alignment and tree estimation
- Liu et al., Science 2009, and Systematic Biology (in press)
- Software available at http://phylo.bio.ku.edu/software/sate/sate.html

DNA Sequence Evolution

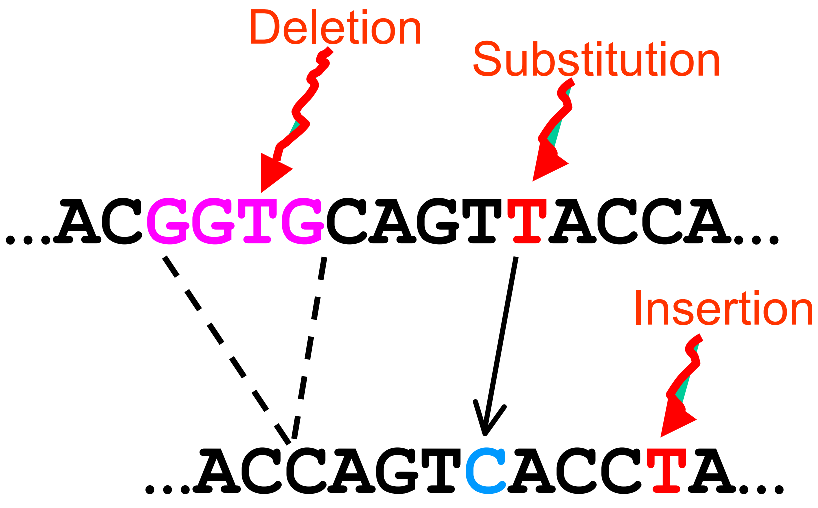


Standard Markov models of biomolecular sequence evolution

- Sequences evolve just with substitutions
- There are a finite number of states (four for DNA and RNA, 20 for aminoacids)
- Sites (i.e., positions) evolve identically and independently
- Numerical parameters describe the probability of substitutions of each type on each edge of the tree



AGGGCATGA AGAT TAGACTT TGCACAA TGCGCTT



Deletion

Substitution

...ACGGTGCAGTTACCA...

Insertion

...ACCAGTCACCTA...

```
...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
```

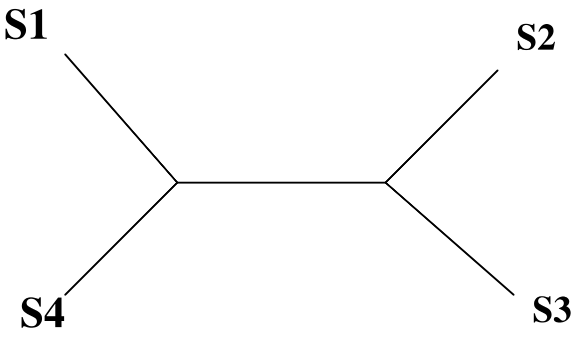
→

```
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

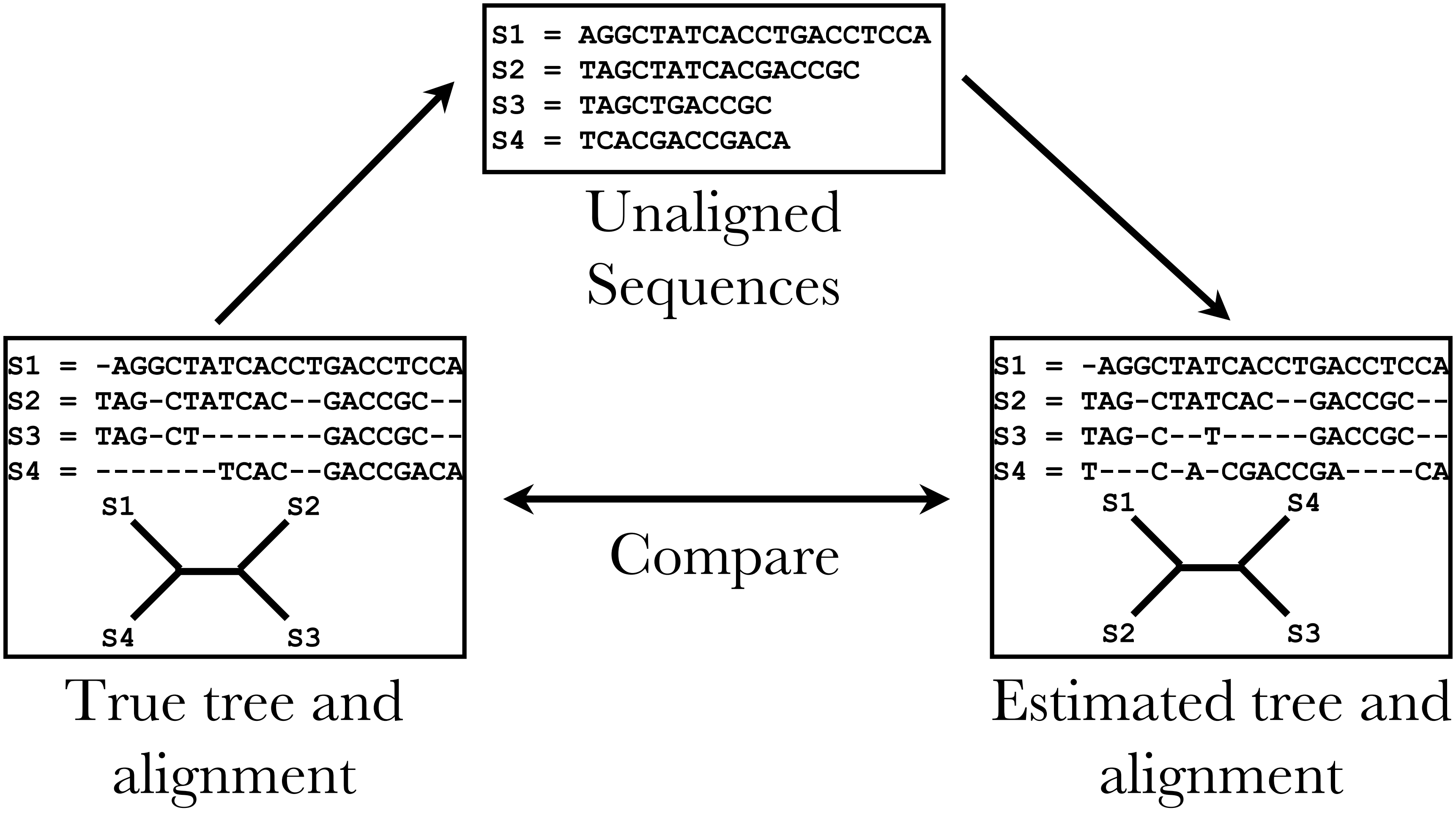
- Clustal
- POY (and POY*)
- Probcons (and Probtree)
- Probalign
- MAFFT
- Muscle
- Di-align
- T-Coffee
- Prank (PNAS 2005, Science 2008)
- Opal (ISMB and Bioinf. 2007)
- *FSA (PLoS Comp. Bio. 2009)*
- *Infernal (Bioinf. 2009)*
- Etc.

Phylogeny methods

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

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

Simulation Studies



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

Unaligned Sequences

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

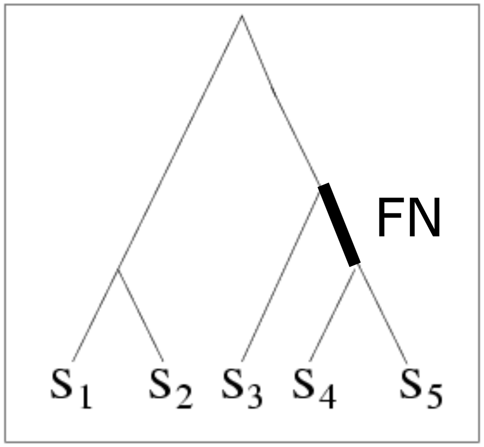
True tree and alignment

Compare

```
S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAC--GACCGC--
S3 = TAG-C--T-----GACCGC--
S4 = T---C-A-CGACCGA----CA
```

Estimated tree and alignment

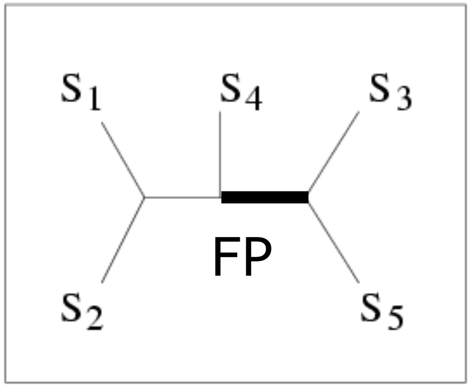
Quantifying Error



TRUE TREE

S_1	ACAATTAGAAC
S_2	ACCCTTAGAAC
S_3	ACCATTCCAAC
S_4	ACCAGACCAAC
S_5	ACCAGACCGGA

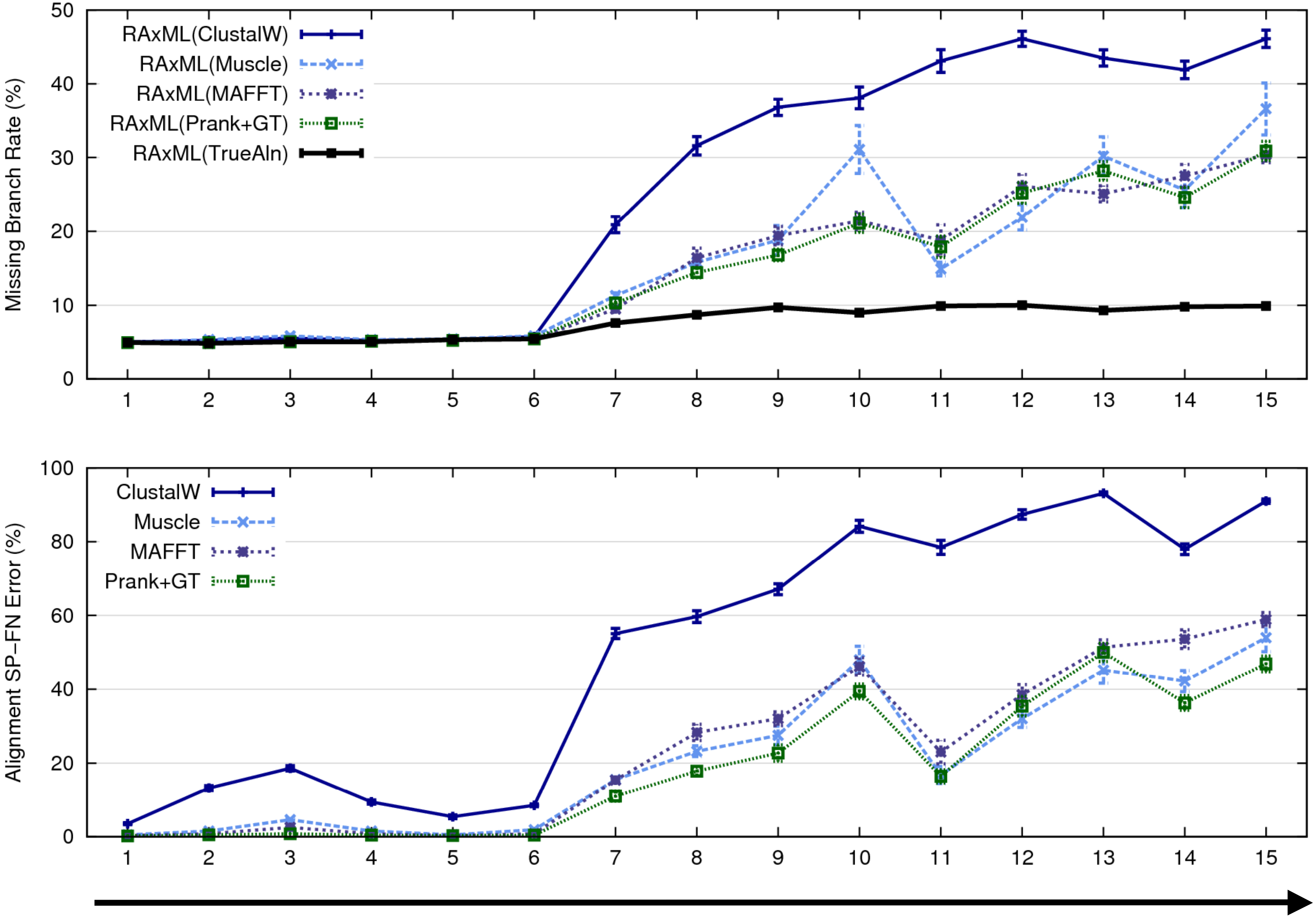
DNA SEQUENCES



INFERRED TREE

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

50% error rate



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)

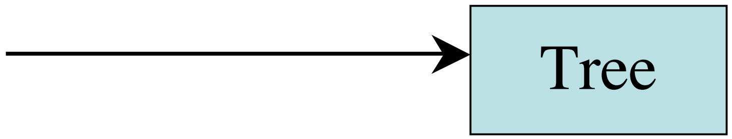
SATé

Input: set of molecular sequences (DNA, RNA, or amino-acids)

Output: alignment A and maximum likelihood tree T on the alignment

SATé Algorithm

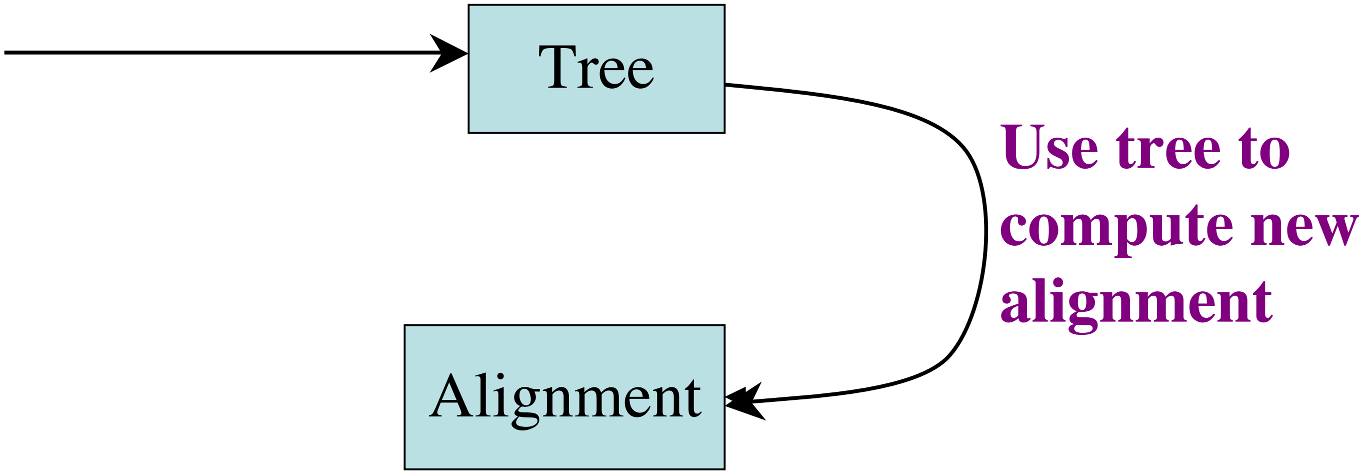
Obtain initial alignment
and estimated ML tree



Tree

SATé Algorithm

Obtain initial alignment
and estimated ML tree

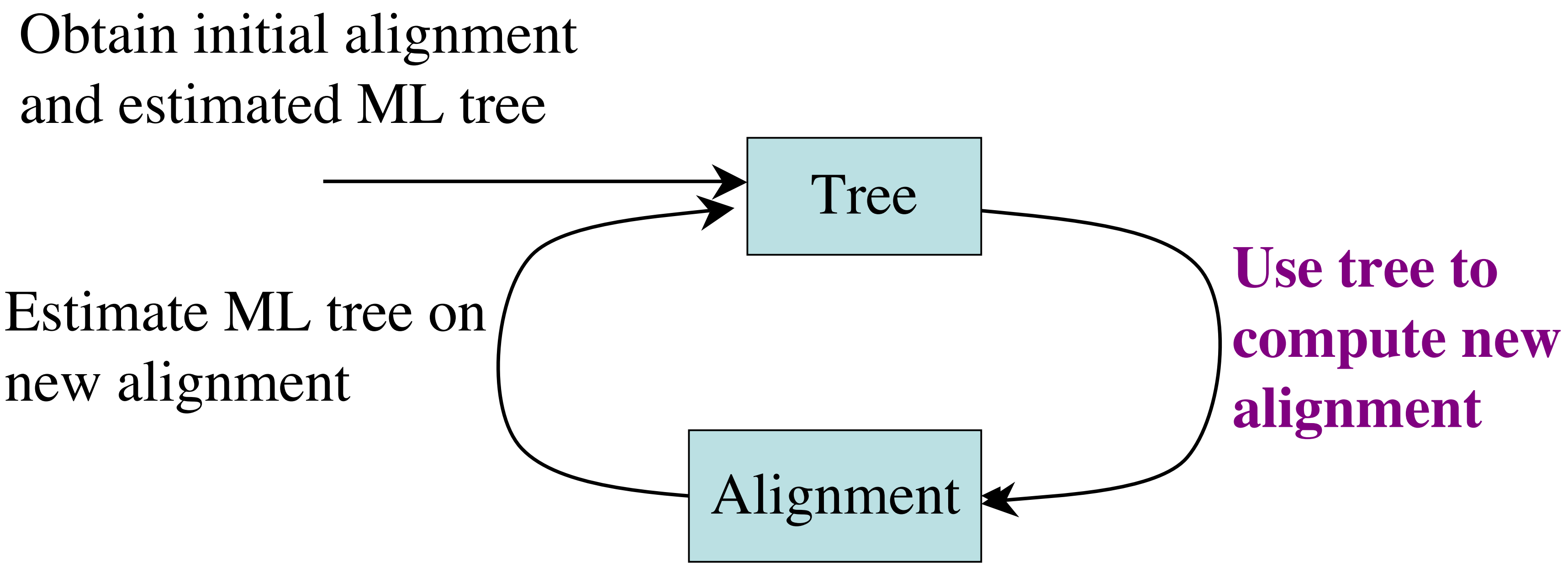


Tree

Use tree to
compute new
alignment

Alignment

SATé Algorithm



Obtain initial alignment and estimated ML tree

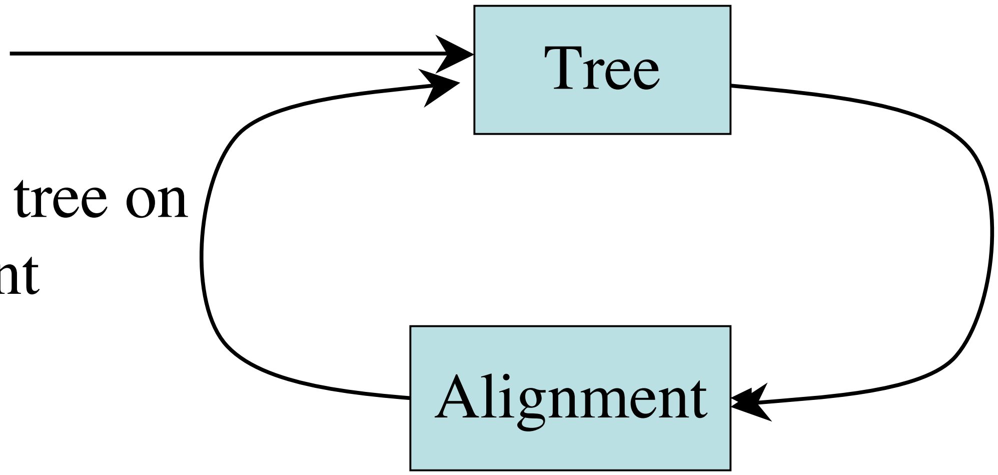
Tree

Use tree to compute new alignment

Alignment

Estimate ML tree on new alignment

SATé Algorithm



Obtain initial alignment and estimated ML tree

Tree

Use tree to compute new alignment

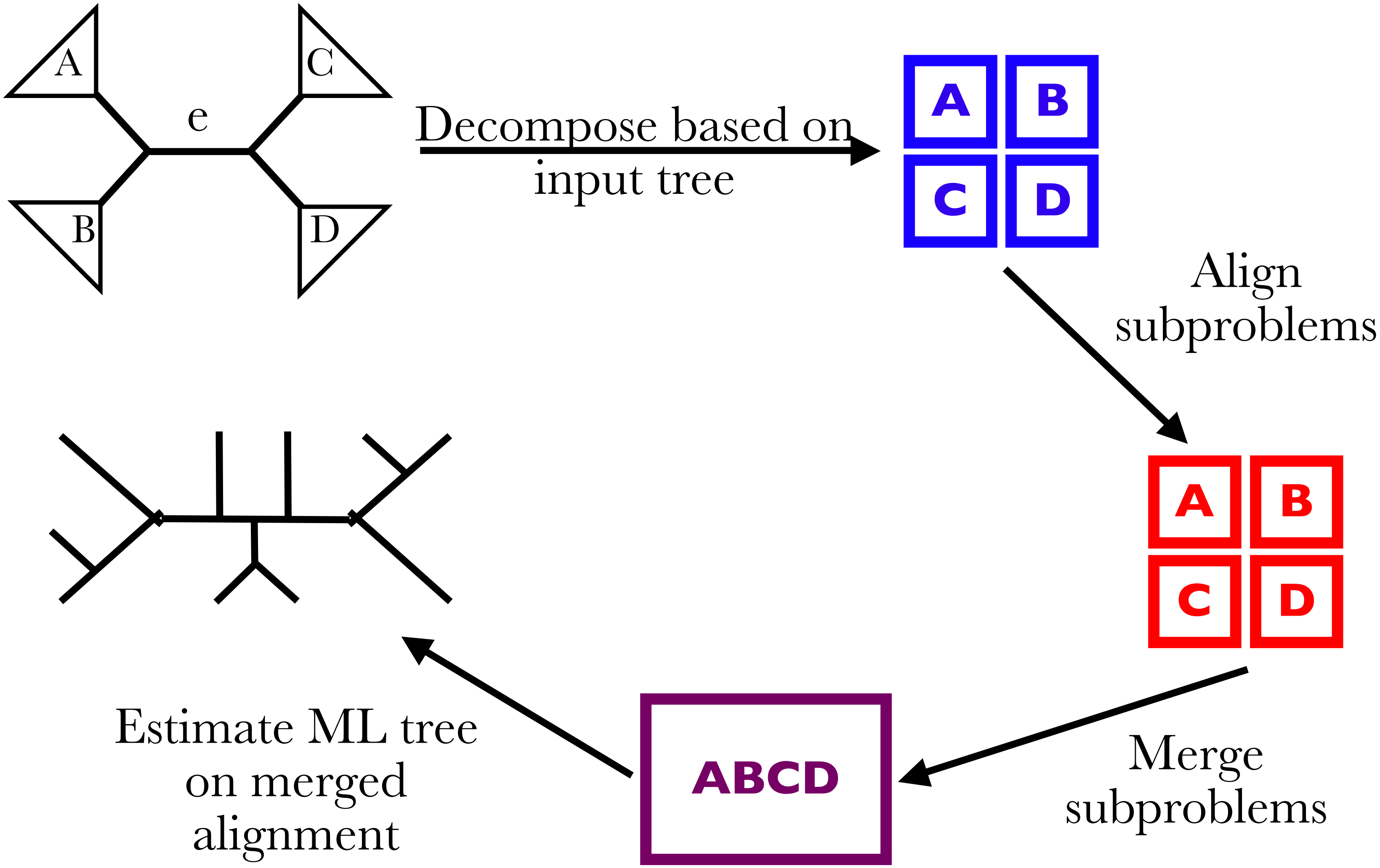
Alignment

Estimate ML tree on 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)

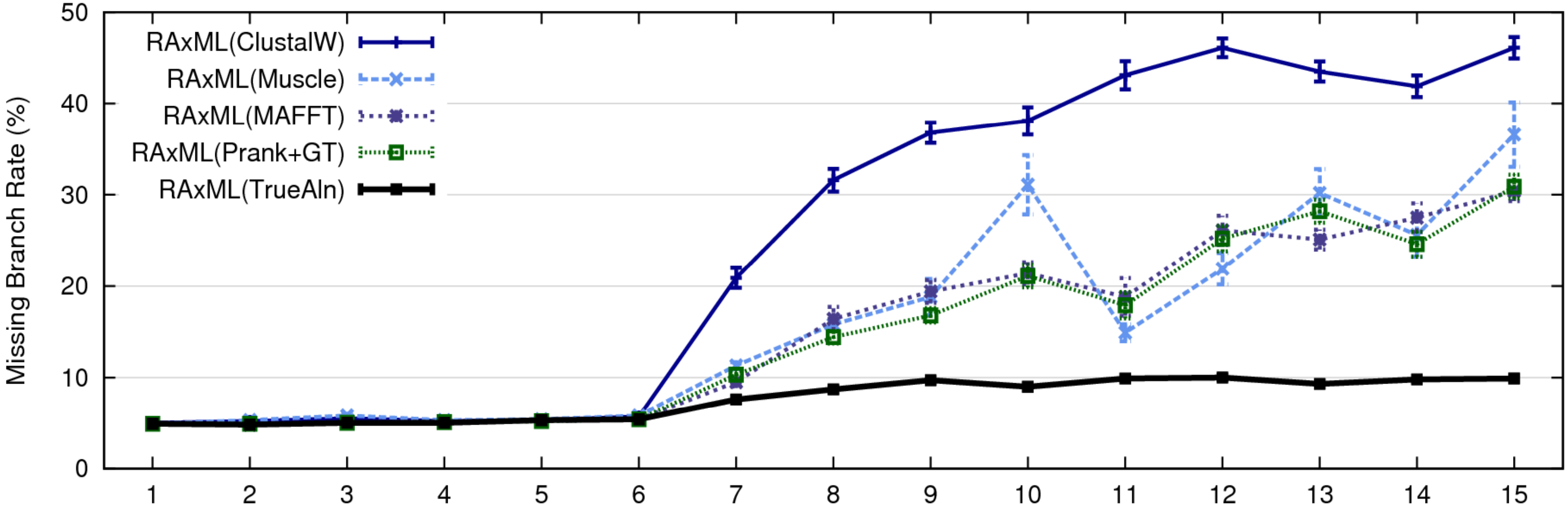


Decompose based on input tree

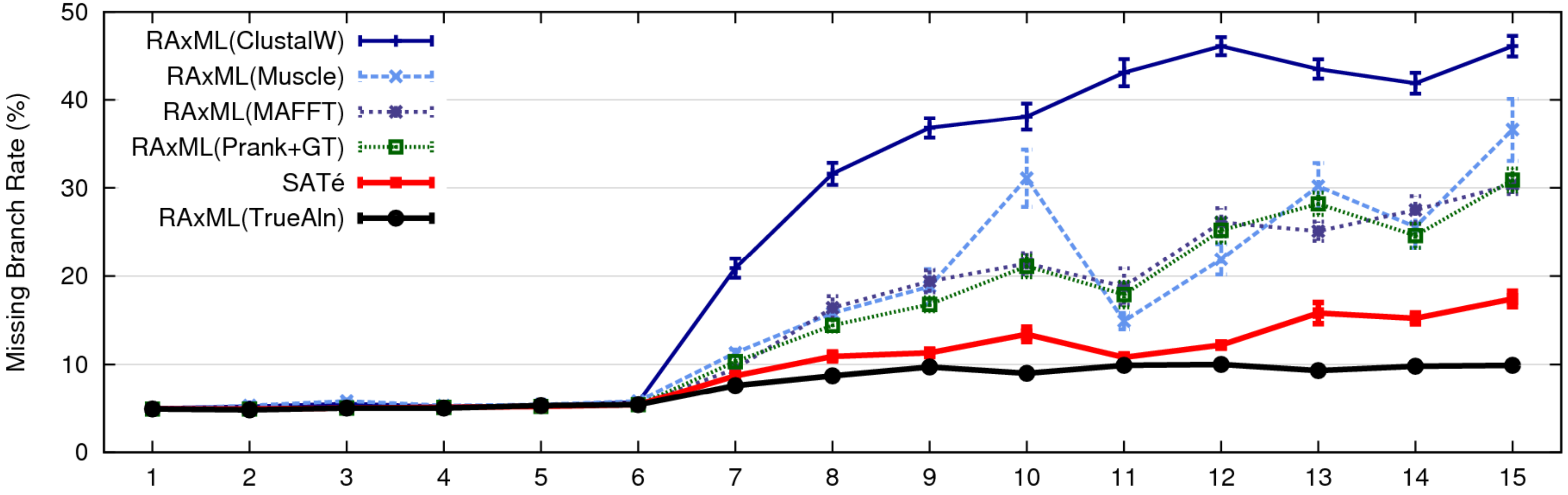
Align subproblems

Merge subproblems

Estimate ML tree on merged alignment



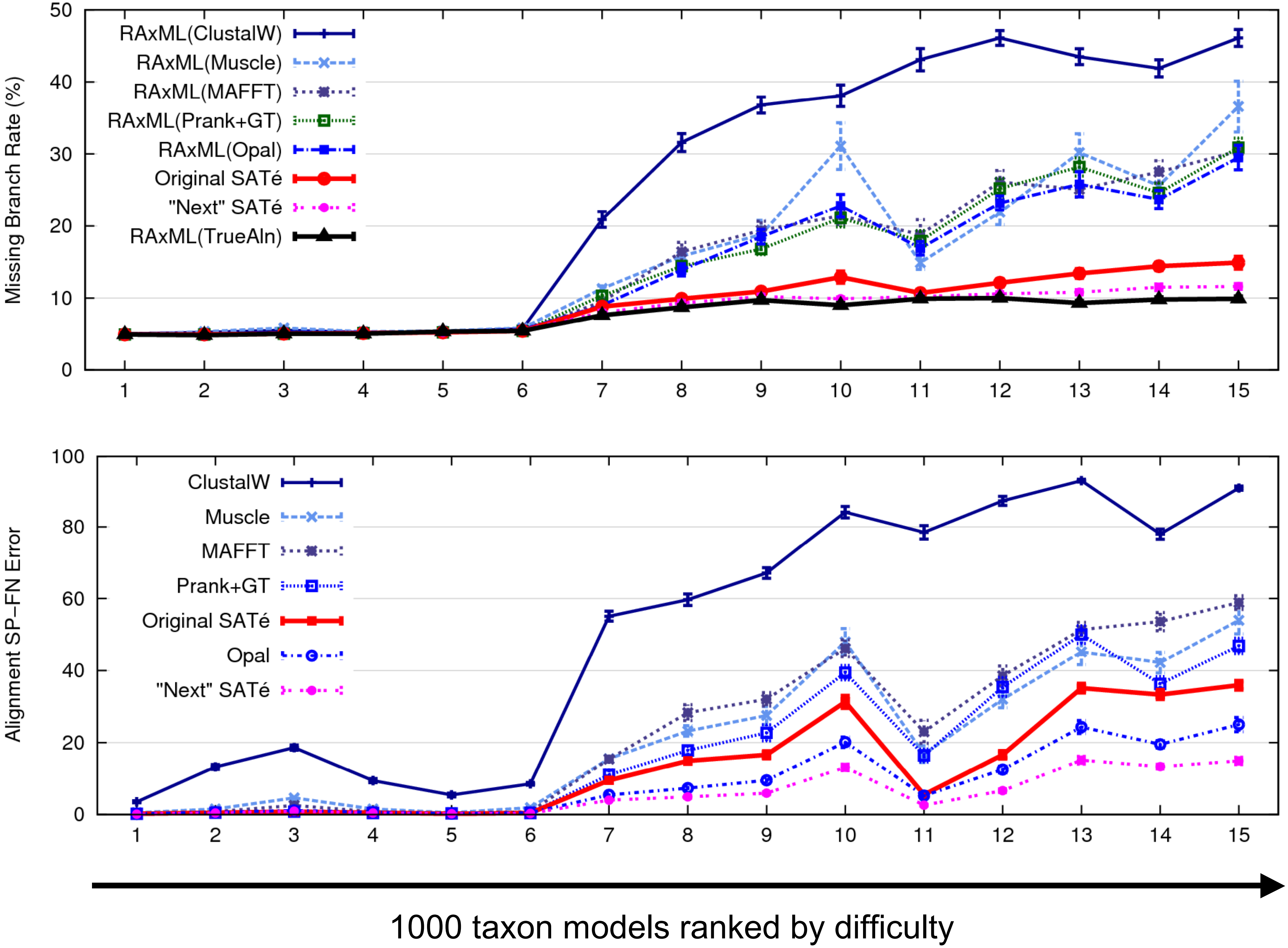
1000 taxon models, ordered by difficulty



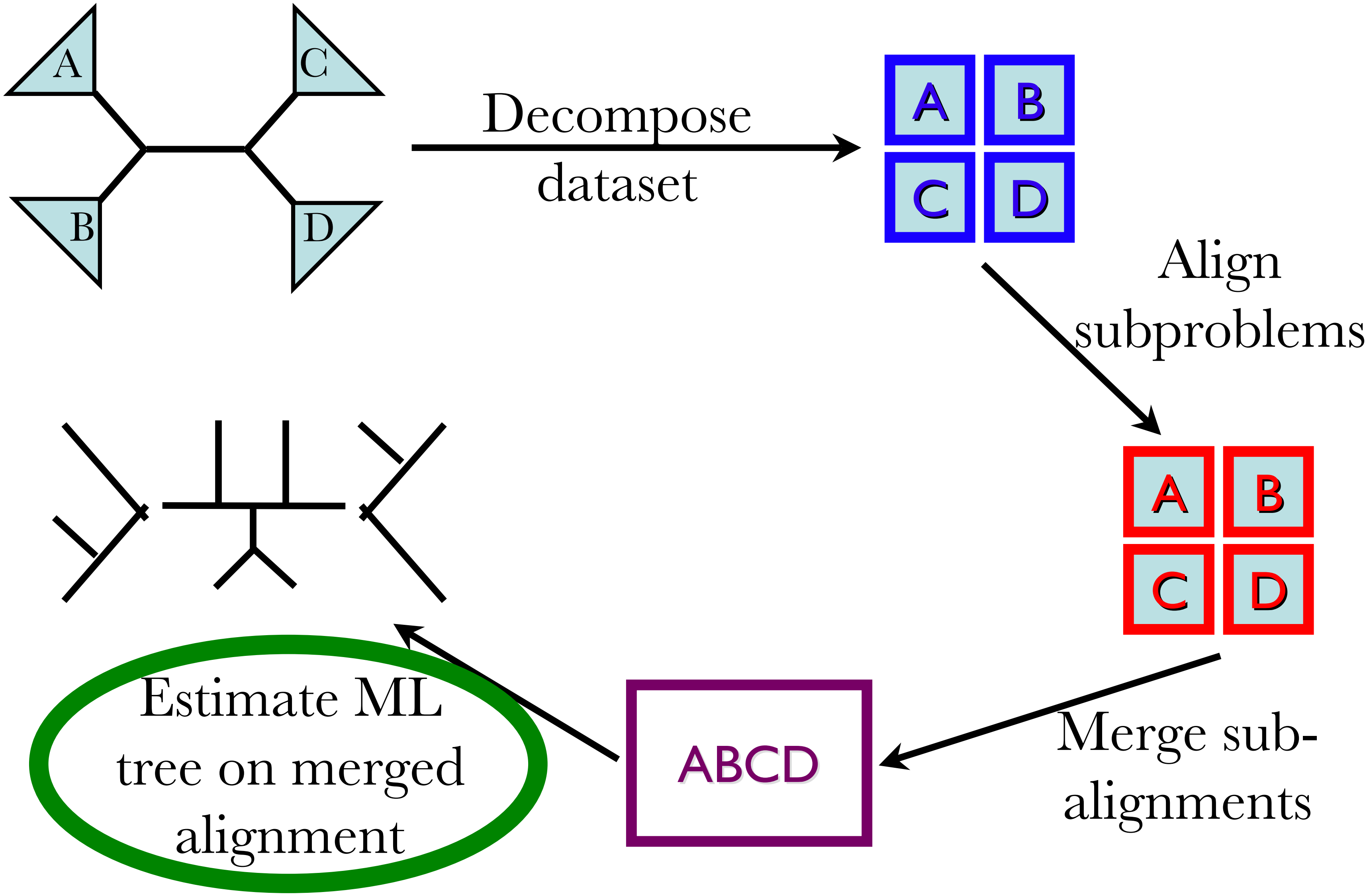
1000 taxon models, ordered by difficulty

24 hour SATé analysis, on desktop machines

(Similar improvements for biological datasets)


Missing Branch Rate (%)
0
10
20
30
40
50
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
RAxML(ClustalW)
RAxML(Muscle)
RAxML(MAFFT)
RAxML(Prank+GT)
RAxML(Opal)
Original SATé
"Next" SATé
RAxML(TrueAln)
Alignment SP–FN Error
0
20
40
60
80
100
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
ClustalW
Muscle
MAFFT
Prank+GT
Original SATé
Opal
"Next" SATé
1000 taxon models ranked by difficulty

Limitations of SATé-I and -II



Decompose dataset

Align subproblems

Merge sub-alignments

Estimate ML tree on merged alignment

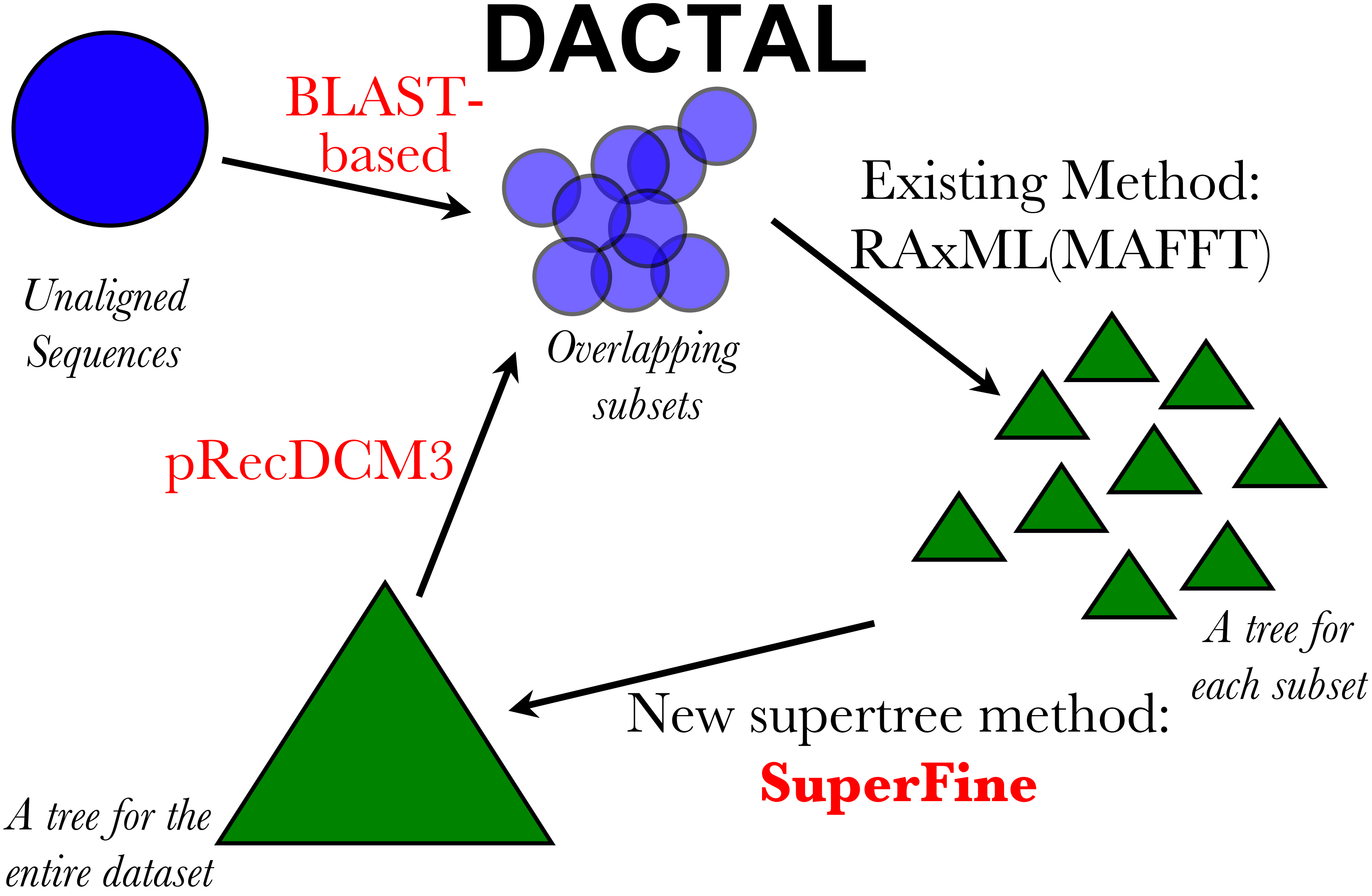
Part II: DACTAL

(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)

DACTAL



Unaligned Sequences

BLAST-based

Overlapping subsets

Existing Method: RAxML(MAFFT)

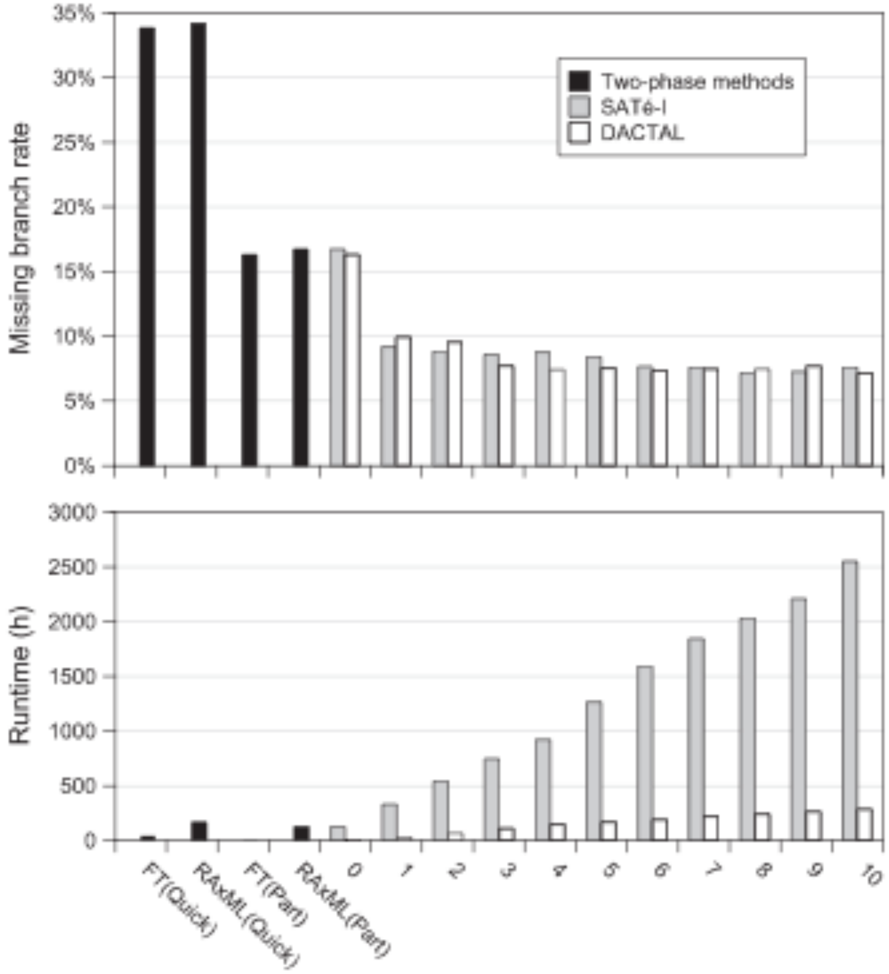
A tree for each subset

New supertree method: **SuperFine**

A tree for the entire dataset

pRecDCM3

DACTAL vs. SATé



16S.T dataset with 7350 seqs from the Comparative RNA website (Gutell).

DACTAL and SATé have comparable accuracy, but DACTAL is much faster.

Average of 3 Largest CRW Datasets

CRW: Comparative RNA database,

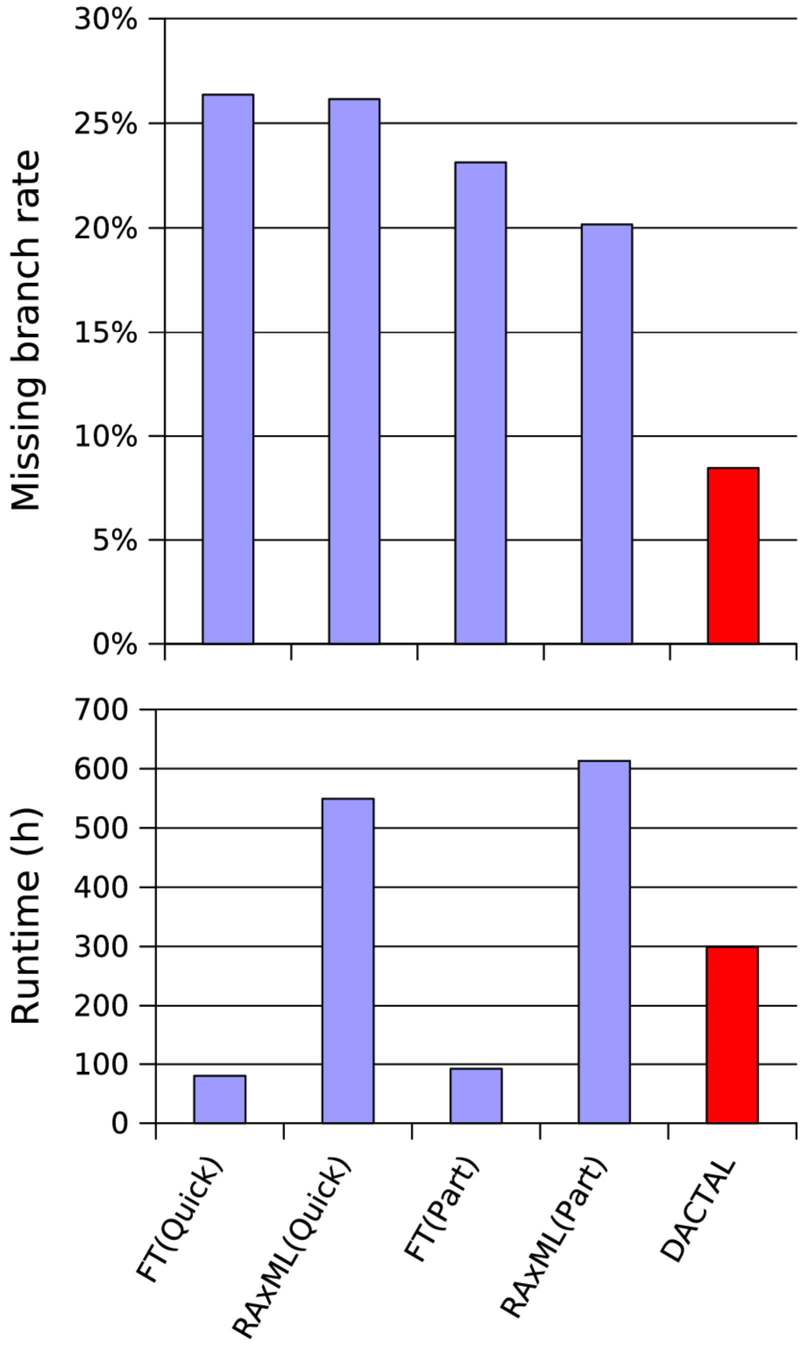
Three 16S datasets with 6,323 to 27,643 sequences

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 can analyze datasets that SATé cannot (and is faster on the datasets both can analyze)
- DACTAL and SATé are very robust to starting trees and other algorithmic parameters

Current Challenges

- **Calculating ultra-large alignments:** The re-alignment step in SATé is polynomial time but still too slow on large datasets.
- **Calculating ultra-large trees:** We have not tested DACTAL on datasets with more than 28,000 sequences.
- **Analyzing metagenomic data:** How do we identify species from short metagenomic reads? How do we do this efficiently? Current datasets are Huge! (300,000,000 reads)

Research Projects

Please come see me if you are interested in a research project in my lab.

- Metagenomics
- Phylogenomics
- Historical Linguistics
- Ultra-large alignment and phylogeny estimation

tandy@cs.utexas.edu

http://www.cs.utexas.edu/users/tandy