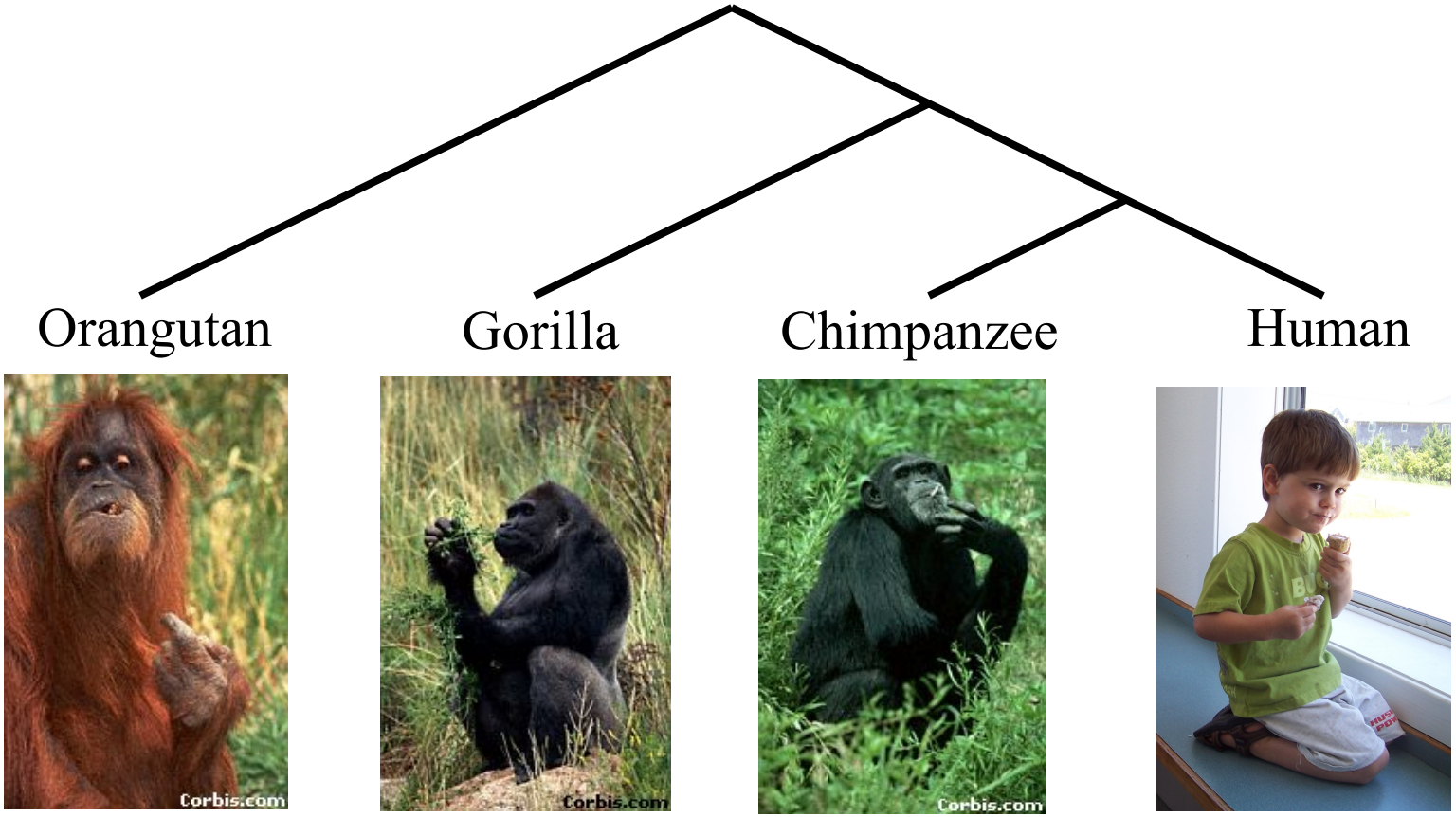
Computational Challenges in Constructing the Tree of Life

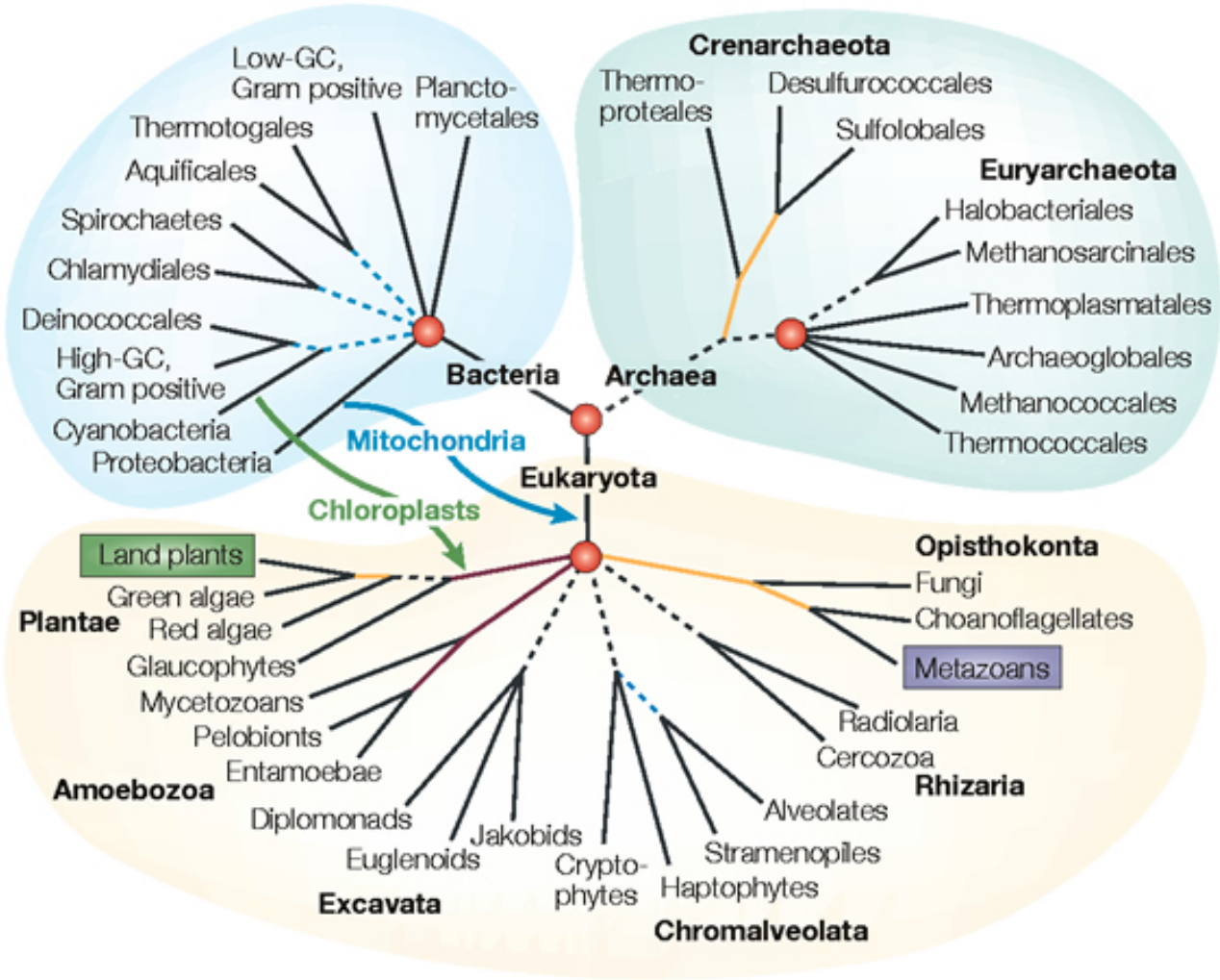
Tandy Warnow
Founder Professor of Engineering
The University of Illinois at Urbana-Champaign
http://tandy.cs.illinois.edu

Phylogeny
(evolutionary tree)



From the Tree of the Life Website,
University of Arizona

The “Tree of Life”

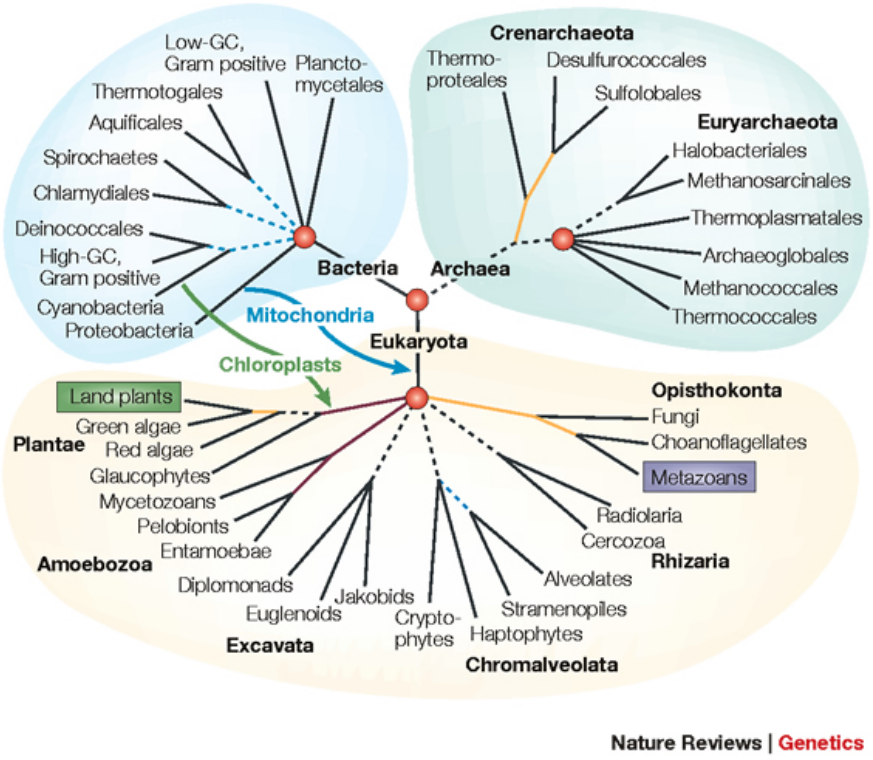


Nature Reviews | Genetics

“Archaea Tree”

courtesy of EMSL @ PNNL

Phylogenies and Applications



Basic Biology:

How did life evolve?

Applications of phylogenies to:

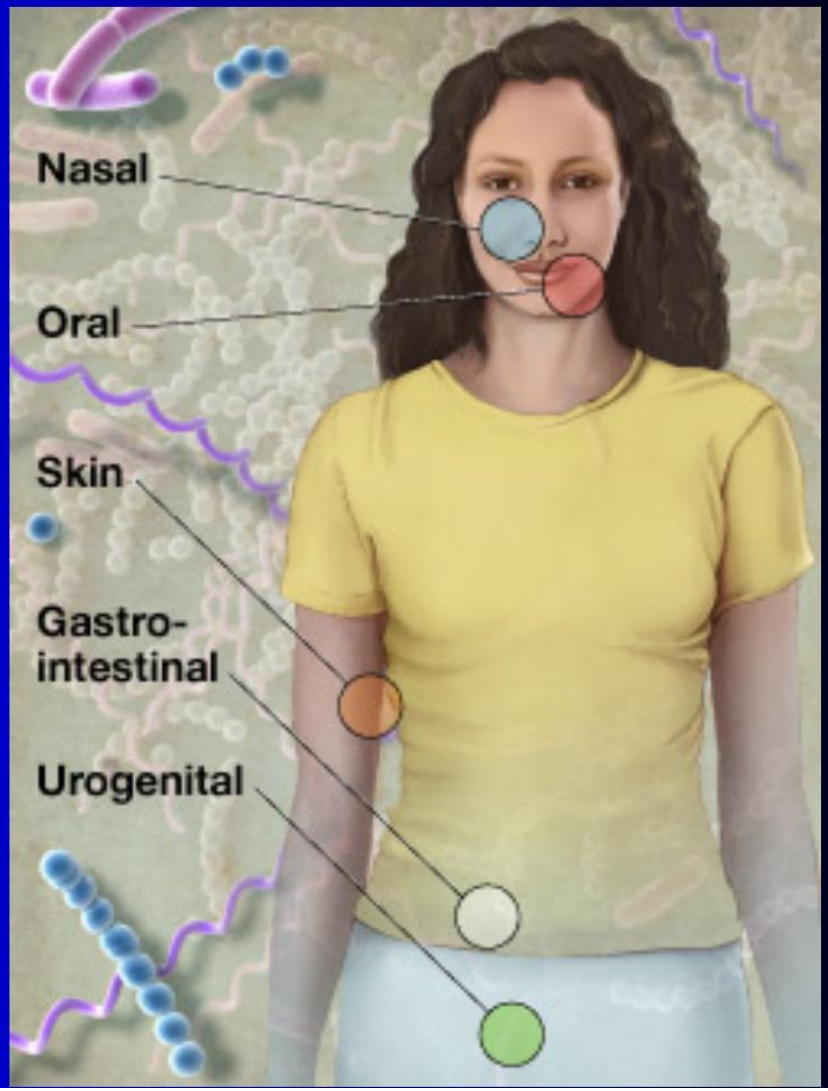
protein structure and function
population genetics
human migrations
metagenomics

"Nothing in biology makes sense except in the light of evolution"

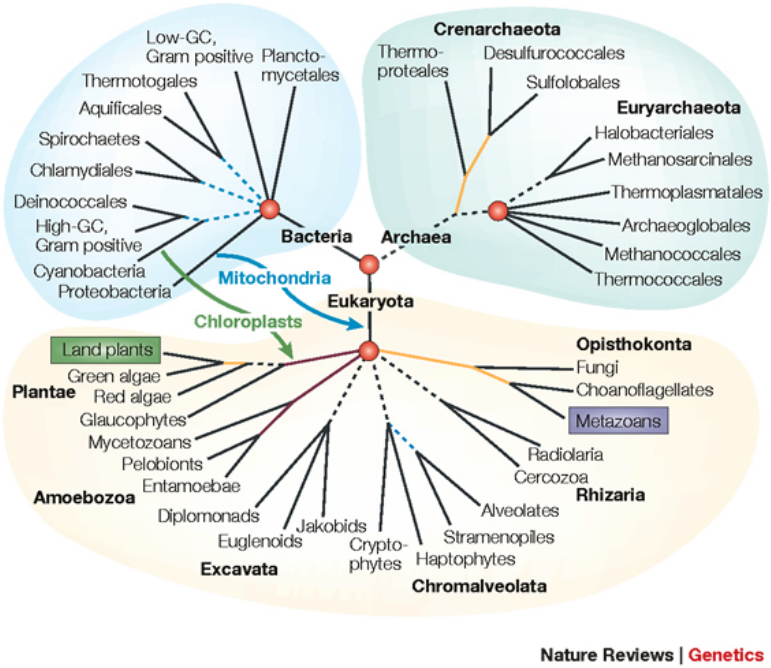
-- Dobhzansky (1973)

The NIH Human Microbiome Project

25,000 human genes,
1,000,000 bacterial genes



Phylogenomics = Species trees from whole genomes



Nature Reviews | Genetics


Does DNA act as a telephone line? p. 1264
Mutations enhancing leukemia development p. 1291 & 1272
A stable gold support gives sharper resolution p. 1317
Science
$10
12 DECEMBER 2014
sciencemag.org
AAAS
SPECIAL ISSUE
Avian genomes
Sequencing across the bird species tree
p. 1308

All aTwitter over an Internet study *p. 683*

The extragalactic background's uneven glow *pp. 696 & 732*

A cellular target for human norovirus *pp. 700 & 755*

Science

$10
7 NOVEMBER 2014
sciencemag.org

AAAS

Insect phylogeny resolved

Molecular insights into insect origins and evolution

p. 763

Phylogenomic pipeline

- Select taxon set and markers
- Gather and screen sequence data, possibly identify orthologs
- For each gene:
  - Compute multiple sequence alignment
  - Construct phylogenetic tree
- Compute species tree or network:
  - Combine the estimated gene trees, OR
  - Estimate a tree from a concatenation of the multiple sequence alignments
- Use species tree with branch support and dates to understand biology

But everything is NP-hard!

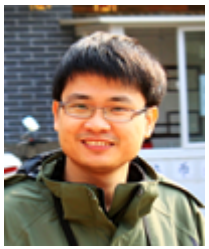
- Select taxon set and markers
- Gather and screen sequence data, possibly identify orthologs
- For each gene:
  - Compute multiple sequence alignment
  - Construct phylogenetic tree
- Compute species tree or network:
  - Combine the estimated gene trees, OR
  - Estimate a tree from a concatenation of the multiple sequence alignments
- Use species tree with branch support and dates to understand biology

Avian Phylogenomics Project

Erich Jarvis, HHMI

MTP Gilbert, Copenhagen

Guojie Zhang, BGI



Siavash Mirarab, Texas

Tandy Warnow, Texas and UIUC

- Approx. 50 species, whole genomes
- 14,000 loci
- Multi-national team (100+ investigators)
- 8 papers published in special issue of Science 2014

Only 48 species, but tree estimation took ~300 CPU years on multiple supercomputers and used 1Tb of memory

Jarvis, Mirarab, et al., examined 48 bird species using 14,000 loci from whole genomes. Two trees were presented.

1. A single dataset maximum likelihood concatenation analysis used ~300 CPU years and 1Tb of distributed memory, using TACC and other supercomputers around the world.

2. However, *every locus had a different tree* – suggestive of "incomplete lineage sorting" – and the noisy genome-scale data required the development of a new method, "statistical binning".

RESEARCH ARTICLE

Whole-genome analyses resolve early branches in the tree of life of modern birds

Erich D. Jarvis,[1*†] Siavash Mirarab,[2*] Andre J. Aberer,[3] Bo Li,[4,5,6] Peter Houde,[7] Cai Li,[4,6] Simon Y. W. Ho,[8] Brant C. Faircloth,[9,10] Benoit Nabholz,[11] Jason T. Howard,[1] Alexander Suh,[12] Claudia C. Weber,[12] Rute R. da Fonseca,[6] Jianwen Li,[4] Fang Zhang,[4] Hui Li,[4] Long Zhou,[4] Nitish Narula,[7,13] Liang Liu,[14] Ganesh Ganapathy,[1] Bastien Boussau,[15] Md. Shamsuzzoha Bayzid,[2] Volodymyr Zavidovych,[1] Sankar Subramanian,[16] Toni Gabaldón,[17,18,19] Salvador Capella-Gutiérrez,[17,18] Jaime Huerta-Cepas,[17,18] Bhanu Rekepalli,[20] Kasper Munch,[21] Mikkel Schierup,[21] Bent Lindow,[6] Wesley C. Warren,[22] David Ray,[23,24,25] Richard E. Green,[26] Michael W. Bruford,[27] Xiangjiang Zhan,[27,28] Andrew Dixon,[29] Shengbin Li,[30] Ning Li,[31] Yinhua Huang,[31] Elizabeth P. Derryberry,[32,33] Mads Frost Bertelsen,[34] Frederick H. Sheldon,[33] Robb T. Brumfield,[33] Claudio V. Mello,[35,36] Peter V. Lovell,[35] Morgan Wirthlin,[35] Maria Paula Cruz Schneider,[36,37] Francisco Prosdocimi,[36,38] José Alfredo Samaniego,[6] Amhed Missael Vargas Velazquez,[6] Alonzo Alfaro-Núñez,[6] Paula F. Campos,[6] Bent Petersen,[39] Thomas Sicheritz-Ponten,[39] An Pas,[40] Tom Bailey,[41] Paul Scofield,[42] Michael Bunce,[43] David M. Lambert,[16] Qi Zhou,[44] Polina Perelman,[45,46] Amy C. Driskell,[47] Beth Shapiro,[26] Zijun Xiong,[4] Yongli Zeng,[4] Shiping Liu,[4] Zhenyu Li,[4] Binghang Liu,[4] Kui Wu,[4] Jin Xiao,[4] Xiong Yinqi,[4] Qiuemei Zheng,[4] Yong Zhang,[4] Huanming Yang,[48] Jian Wang,[48] Linnea Smeds,[12] Frank E. Rheindt,[49] Michael Braun,[50] Jon Fjeldsa,[51] Ludovic Orlando,[6] F. Keith Barker,[52] Knud Andreas Jønsson,[51,53,54] Warren Johnson,[55] Klaus-Peter Koepfli,[56] Stephen O'Brien,[57,58] David Haussler,[59] Oliver A. Ryder,[60] Carsten Rahbek,[51,54] Eske Willerslev,[6] Gary R. Graves,[51,61] Travis C. Glenn,[62] John McCormack,[63] Dave Burt,[64] Hans Ellegren,[12] Per Alström,[65,66] Scott V. Edwards,[67] Alexandros Stamatakis,[3,68] David P. Mindell,[69] Joel Cracraft,[70] Edward L. Braun,[71] Tandy Warnow,[2,72†] Wang Jun,[48,73,74,75,76†] M. Thomas P. Gilbert,[6,43†] Guojie Zhang[4,77†]

To better determine the history of modern birds, we performed a genome-scale phylogenetic analysis of 48 species representing all orders of Neoaves using phylogenomic methods created to handle genome-scale data. We recovered a highly resolved tree that confirms

The second tree was computed by combining estimated gene trees, and used only 5 CPU years (serial time), and was embarrassingly parallel.

Jarvis, Mirarab, et al., examined 48 bird species using 14,000 loci from whole genomes. Two trees were presented.

1. A single dataset maximum likelihood concatenation analysis used ~300 CPU years and 1Tb of distributed memory, using TACC and other supercomputers around the world.

2. However, *every locus had a different tree* – suggestive of "incomplete lineage sorting" – and the noisy genome-scale data required the development of a new method, "statistical binning".

RESEARCH ARTICLE

Whole-genome analyses resolve early branches in the tree of life of modern birds

Erich D. Jarvis,[1*†] Siavash Mirarab,[2*] Andre J. Aberer,[3] Bo Li,[4,5,6] Peter Houde,[7] Cai Li,[4,6] Simon Y. W. Ho,[8] Brant C. Faircloth,[9,10] Benoit Nabholz,[11] Jason T. Howard,[1] Alexander Suh,[12] Claudia C. Weber,[12] Rute R. da Fonseca,[6] Jianwen Li,[4] Fang Zhang,[4] Hui Li,[4] Long Zhou,[4] Nitish Narula,[7,13] Liang Liu,[14] Ganesh Ganapathy,[1] Bastien Boussau,[15] Md. Shamsuzzoha Bayzid,[2] Volodymyr Zavidovych,[1] Sankar Subramanian,[16] Toni Gabaldón,[17,18,19] Salvador Capella-Gutiérrez,[17,18] Jaime Huerta-Cepas,[17,18] Bhanu Rekepalli,[20] Kasper Munch,[21] Mikkel Schierup,[21] Bent Lindow,[6] Wesley C. Warren,[22] David Ray,[23,24,25] Richard E. Green,[26] Michael W. Bruford,[27] Xiangjiang Zhan,[27,28] Andrew Dixon,[29] Shengbin Li,[30] Ning Li,[31] Yinhua Huang,[31] Elizabeth P. Derryberry,[32,33] Mads Frost Bertelsen,[34] Frederick H. Sheldon,[33] Robb T. Brumfield,[33] Claudio V. Mello,[35,36] Peter V. Lovell,[35] Morgan Wirthlin,[35] Maria Paula Cruz Schneider,[36,37] Francisco Prosdocimi,[36,38] José Alfredo Samaniego,[6] Amhed Missael Vargas Velazquez,[6] Alonzo Alfaro-Núñez,[6] Paula F. Campos,[6] Bent Petersen,[39] Thomas Sicheritz-Ponten,[39] An Pas,[40] Tom Bailey,[41] Paul Scofield,[42] Michael Bunce,[43] David M. Lambert,[16] Qi Zhou,[44] Polina Perelman,[45,46] Amy C. Driskell,[47] Beth Shapiro,[26] Zijun Xiong,[4] Yongli Zeng,[4] Shiping Liu,[4] Zhenyu Li,[4] Binghang Liu,[4] Kui Wu,[4] Jin Xiao,[4] Xiong Yinqi,[4] Qiuemei Zheng,[4] Yong Zhang,[4] Huanming Yang,[48] Jian Wang,[48] Linnea Smeds,[12] Frank E. Rheindt,[49] Michael Braun,[50] Jon Fjeldsa,[51] Ludovic Orlando,[6] F. Keith Barker,[52] Knud Andreas Jønsson,[51,53,54] Warren Johnson,[55] Klaus-Peter Koepfli,[56] Stephen O'Brien,[57,58] David Haussler,[59] Oliver A. Ryder,[60] Carsten Rahbek,[51,54] Eske Willerslev,[6] Gary R. Graves,[51,61] Travis C. Glenn,[62] John McCormack,[63] Dave Burt,[64] Hans Ellegren,[12] Per Alström,[65,66] Scott V. Edwards,[67] Alexandros Stamatakis,[3,68] David P. Mindell,[69] Joel Cracraft,[70] Edward L. Braun,[71] Tandy Warnow,[2,72†] Wang Jun,[48,73,74,75,76†] M. Thomas P. Gilbert,[6,43†] Guojie Zhang[4,77†]

To better determine the history of modern birds, we performed a genome-scale phylogenetic analysis of 48 species representing all orders of Neoaves using phylogenomic methods created to handle genome-scale data. We recovered a highly resolved tree that confirms

We used 100 CPU years (mostly on TACC) to develop and test this method.

RESEARCH ARTICLE SUMMARY

AVIAN GENOMICS

Statistical binning enables an accurate coalescent-based estimation of the avian tree

Siavash Mirarab, Md. Shamsuzzoha Bayzid, Bastien Boussau, Tandy Warnow*

INTRODUCTION: Reconstructing species trees for rapid radiations, as in the early diversification of birds, is complicated by biological processes such as incomplete lineage sorting (ILS) that can cause different parts of the genome to have different evolutionary histories. Statistical methods, based on the multispecies coalescent model and that combine gene trees, can be highly accurate even in the presence of massive ILS; however, these methods can produce species trees that are topologically far from the species tree when estimated gene trees have error. We have developed a statistical binning technique to address gene tree estimation error and have explored its use in genome-scale species tree estimation with MP-EST, a popular coalescent-based species tree estimation method.

ON OUR WEB SITE

Read the full article at http://dx.doi.org/10.1126/science.1250463

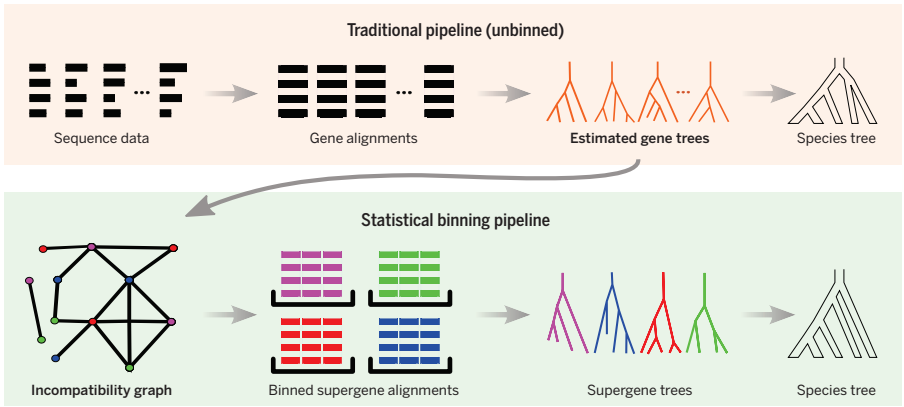
RATIONALE: In statistical binning, phylogenetic trees on different genes are estimated and then placed into bins, so that the differences between trees in the same bin can be explained by estimation error (see the figure). A new tree is then estimated for each bin by applying maximum likelihood to a concatenated alignment of the multiple sequence alignments of its genes, and a species tree is estimated using a coalescent-based species tree method from these supergene trees.

RESULTS: Under realistic conditions in our simulation study, statistical binning reduced the topological error of species trees estimated using MP-EST and enabled a coalescent-based analysis that was more accurate than concatenation even when gene tree estimation error was relatively high. Statistical binning also reduced the error in gene tree topology and species tree branch length estimation, especially when the phylogenetic signal in gene sequence alignments was low. Species trees estimated using MP-EST with statistical binning on four biological data sets showed increased concordance with the biological literature. When MP-EST was used to analyze 14,446 gene trees in the avian phylogenomics project, it produced a species tree that was discordant with the concatenation analysis and conflicted with prior literature. However, the statistical binning analysis produced a tree that was highly congruent with the concatenation analysis and was consistent with the prior scientific literature.

CONCLUSIONS: Statistical binning reduces the error in species tree topology and branch length estimation because it reduces gene tree estimation error. These improvements are greatest when gene trees have reduced bootstrap support, which was the case for the avian phylogenomics project. Because using unbinned gene trees can result in overestimation of ILS, statistical binning may be helpful in providing more accurate estimations of ILS levels in biological data sets. Thus, statistical binning enables highly accurate species tree estimations, even on genome-scale data sets. ■

The list of author affiliations is available in the full article online.
*Corresponding author. E-mail: warnow@illinois.edu
Cite this article as S. Mirarab *et al.*, *Science* 346, 1250463 (2014). DOI: 10.1126/science.1250463

Statistical binning technique



The statistical binning pipeline for estimating species trees from gene trees. Loci are grouped into bins based on a statistical test for combinabilty, before estimating gene trees.

Published by AAAS

1kp: Thousand Transcriptome Project

G. Ka-Shu Wong
U Alberta

J. Leebens-Mack
U Georgia

N. Wickett
Northwestern

N. Matasci
iPlant

T. Warnow,
UIUC

S. Mirarab,
UT-Austin

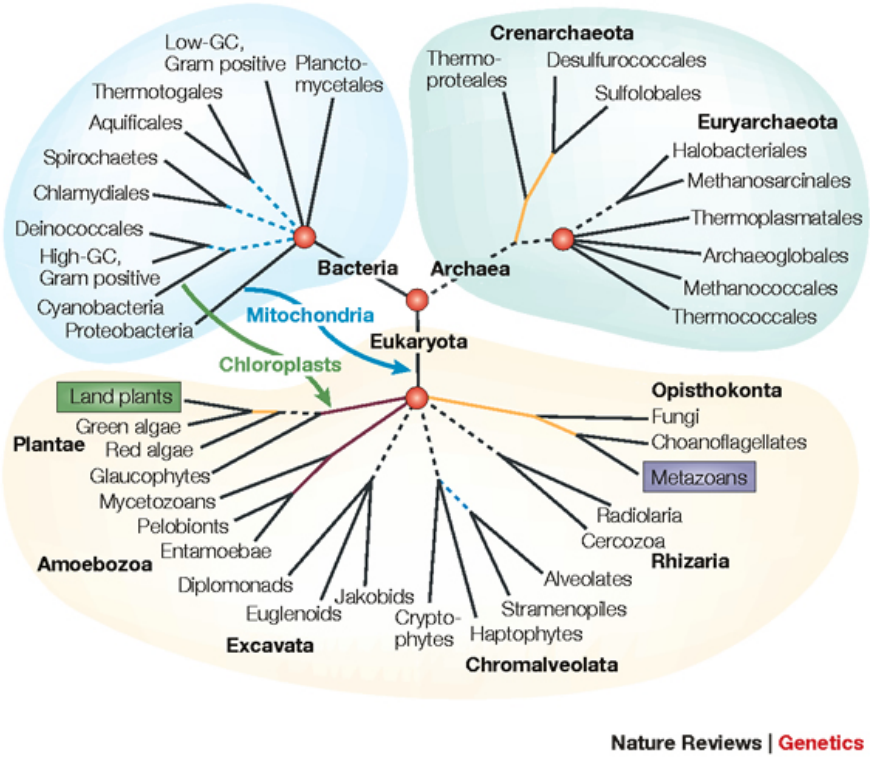
N. Nguyen,
UT-Austin

Plus many many other people…

- First study (Wickett, Mirarab, et al., PNAS 2014) had ~100 species and ~800 genes, gene trees and alignments estimated using SATé, and a coalescent-based species tree estimated using ASTRAL
- Second study: Plant Tree of Life based on transcriptomes of ~1200 species, and more than 13,000 gene families (most not single copy)

Challenges:
Species tree estimation from conflicting gene trees
Gene tree estimation of datasets with > 100,000 sequences

Hard Computational Problems



NP-hard problems

Large datasets

- 100,000+ sequences
- thousands of genes

"Big data" complexity:

- heterogeneity
- model misspecification
- fragmentary sequences
- errors in input data
- streaming data

Two dimensions

- Number of genes (or total number of sites)
  - Thousands of genes for multi-gene analyses
  - Thousands of sites for single genes, millions for multi-gene analyses
  - Some types of analyses can be parallelized
- Number of species
  - Many datasets have thousands of species
  - The Tree of Life will have millions
  - Number of trees on n leaves is (2n-5)!!
  - Parallelism is much more complicated

Two dimensions

- Number of genes (or total number of sites)
  - Thousands of genes for multi-gene analyses
  - Thousands of sites for single genes, millions for multi-gene analyses
  - Some types of analyses can be parallelized
- Number of species
  - Many datasets have thousands of species
  - The Tree of Life will have millions
  - Number of trees on n leaves is (2n-5)!!
  - Parallelism is much more complicated

Divide-and-conquer

- Million-sequence multiple sequence alignments
- Genome-scale phylogeny estimation with up to 1,000 species and 1,000 genes
- DACTAL (almost alignment-free tree estimation)
- DCM-NJ (boosting distance-based methods)

Divide-and-conquer key to the improvements in scalability and accuracy, and produces embarrassingly parallel algorithms.

Divide-and-conquer

- Million-sequence multiple sequence alignments
- Genome-scale phylogeny estimation with up to 1,000 species and 1,000 genes
- DACTAL (almost alignment-free tree estimation)
- DCM-NJ (boosting distance-based methods)

Divide-and-conquer key to the improvements in scalability and accuracy, and produces embarrassingly parallel algorithms.

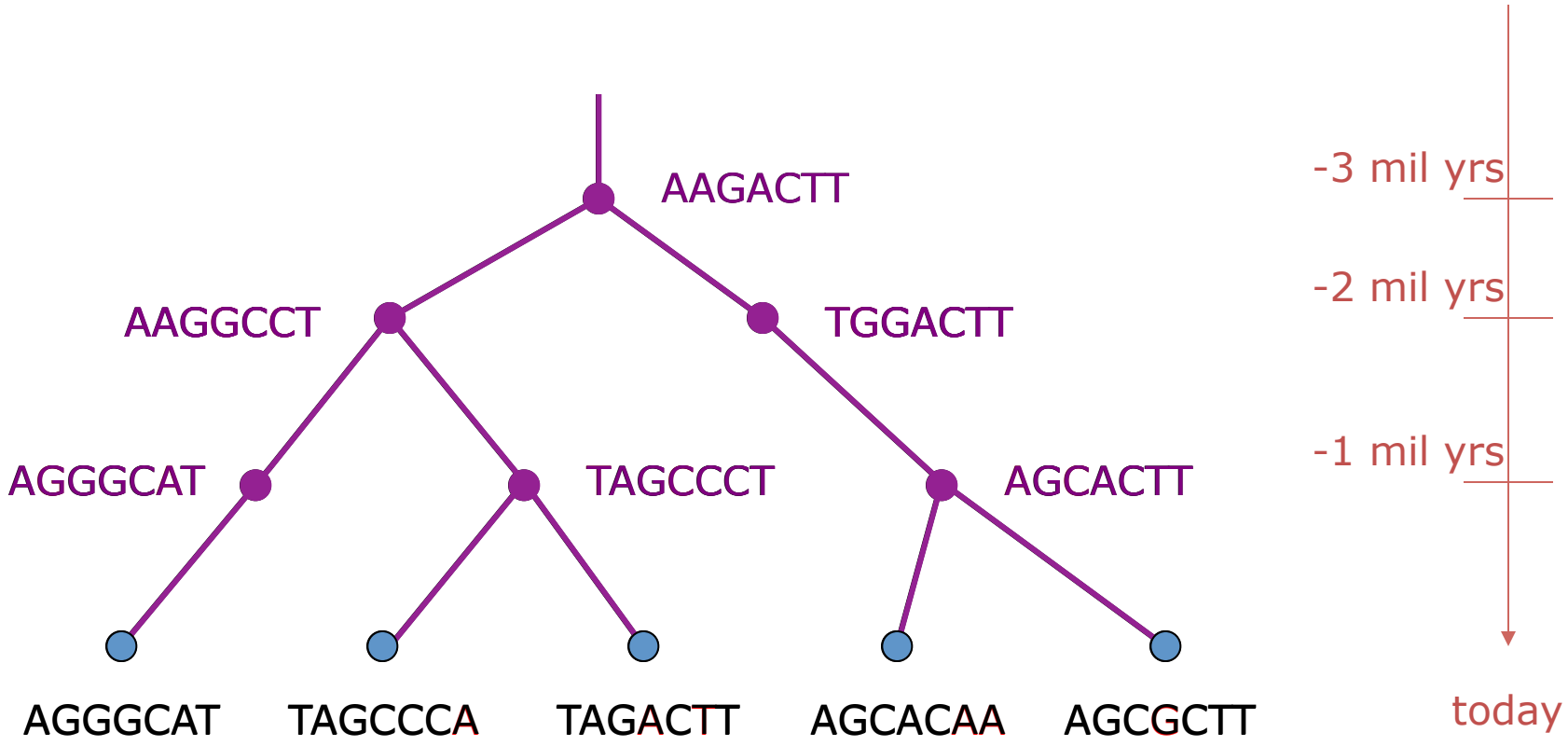
Today's Talk

- Brief overview of phylogenomic pipeline.
- Parallel algorithms in phylogenetics:
  - Ultra-large Multiple Sequence Alignment
- Themes
- Outstanding problems

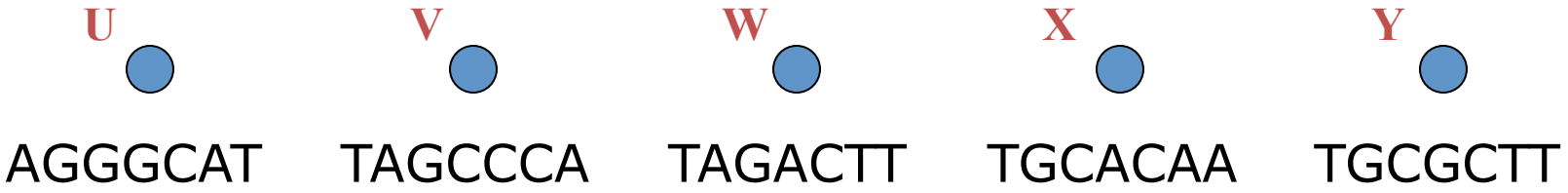
Phylogenomic Pipeline

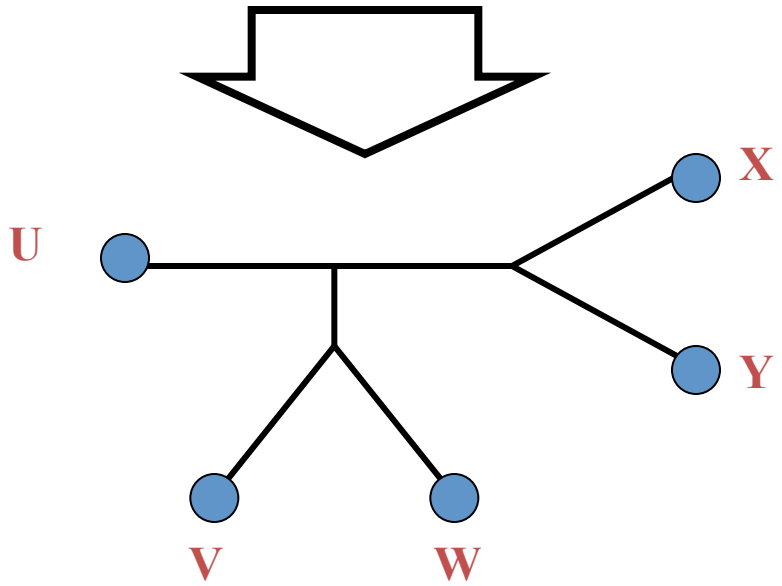
- Select taxon set and markers
- Gather and screen sequence data, possibly identify orthologs
- For each gene:
  - Compute multiple sequence alignment
  - Construct phylogenetic tree
- Compute species tree or network:
  - Combine the estimated gene trees, OR
  - Estimate a tree from a concatenation of the multiple sequence alignments
- Use species tree with branch support and dates to understand biology

DNA Sequence Evolution

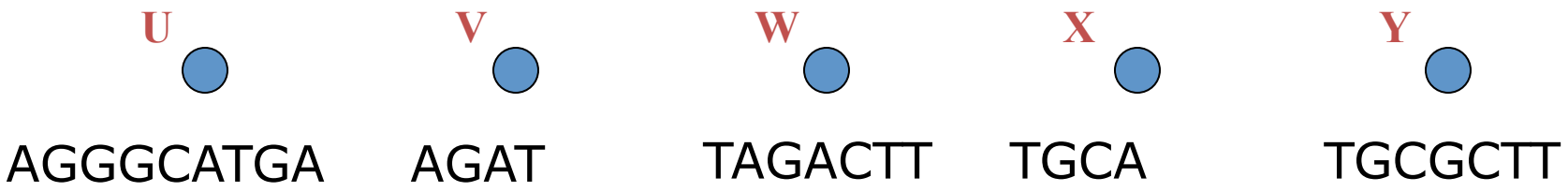


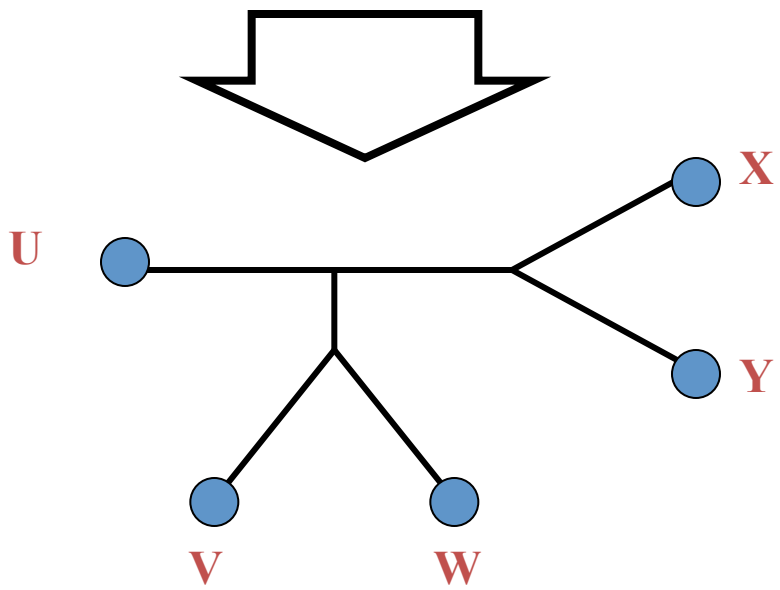
The Classical Phylogeny Problem



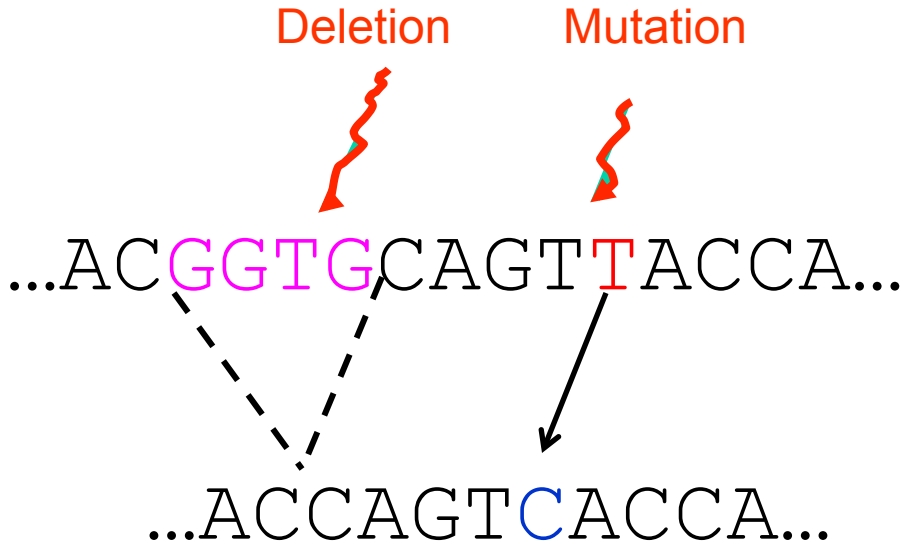


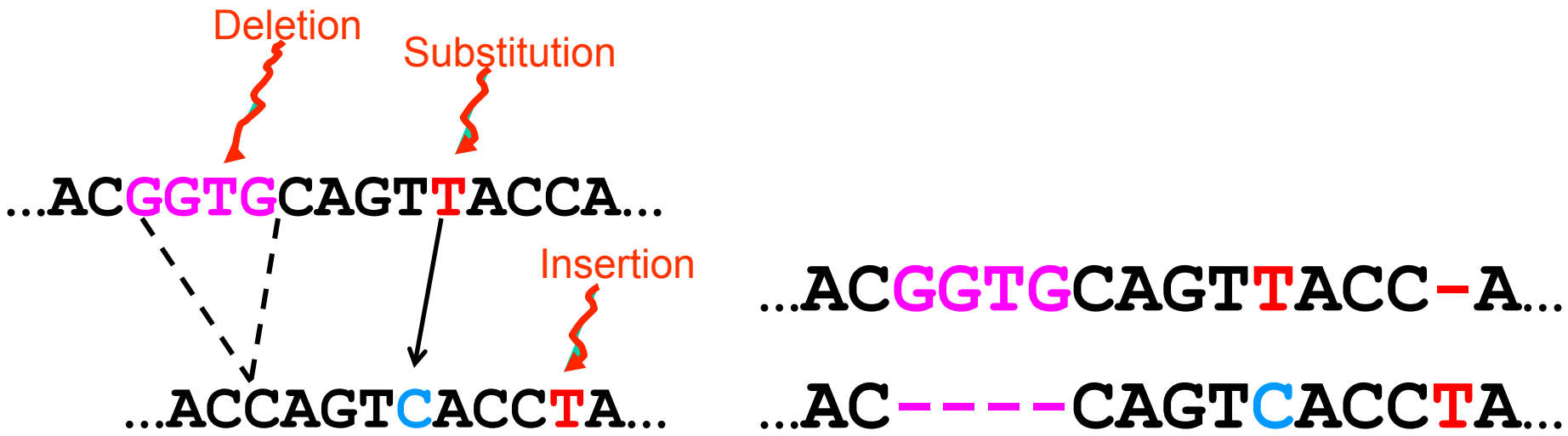
However…





Indels (insertions and deletions)





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
```



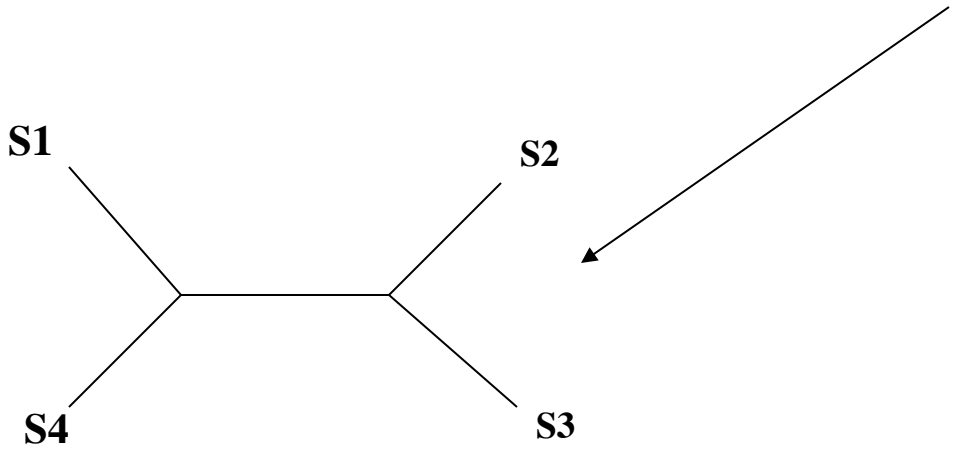
```
S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAC--GACCGC--
S3 = TAG-CT-------GACCGC--
S4 = -------TCAC--GACCGACA
```

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
```



Phylogenomic pipeline

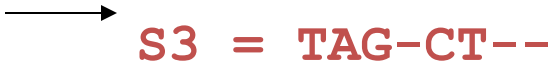
- Select taxon set and markers
- Gather and screen sequence data, possibly identify orthologs
- For each gene:
  - Compute multiple sequence alignment
  - Construct phylogenetic tree
- Compute species tree or network:
  - Combine the estimated gene trees, OR
  - Estimate a tree from a concatenation of the multiple sequence alignments
- Use species tree with branch support and dates to understand biology

Phylogenomic pipeline

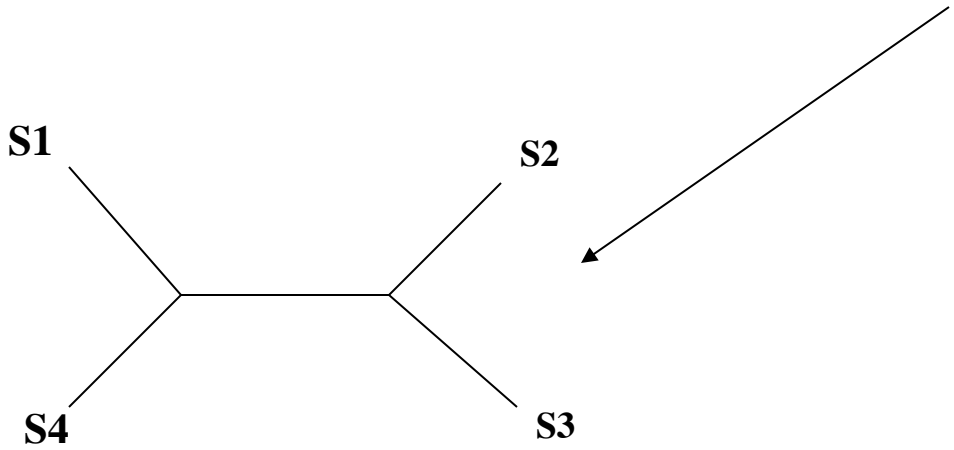
- Select taxon set and markers
- Gather and screen sequence data, possibly identify orthologs
- For each gene:
  - Compute multiple sequence alignment
  - Construct phylogenetic tree
- Compute species tree or network:
  - Combine the estimated gene trees, OR
  - Estimate a tree from a concatenation of the multiple sequence alignments
- Use species tree with branch support and dates to understand biology

First Align, then Compute the Tree

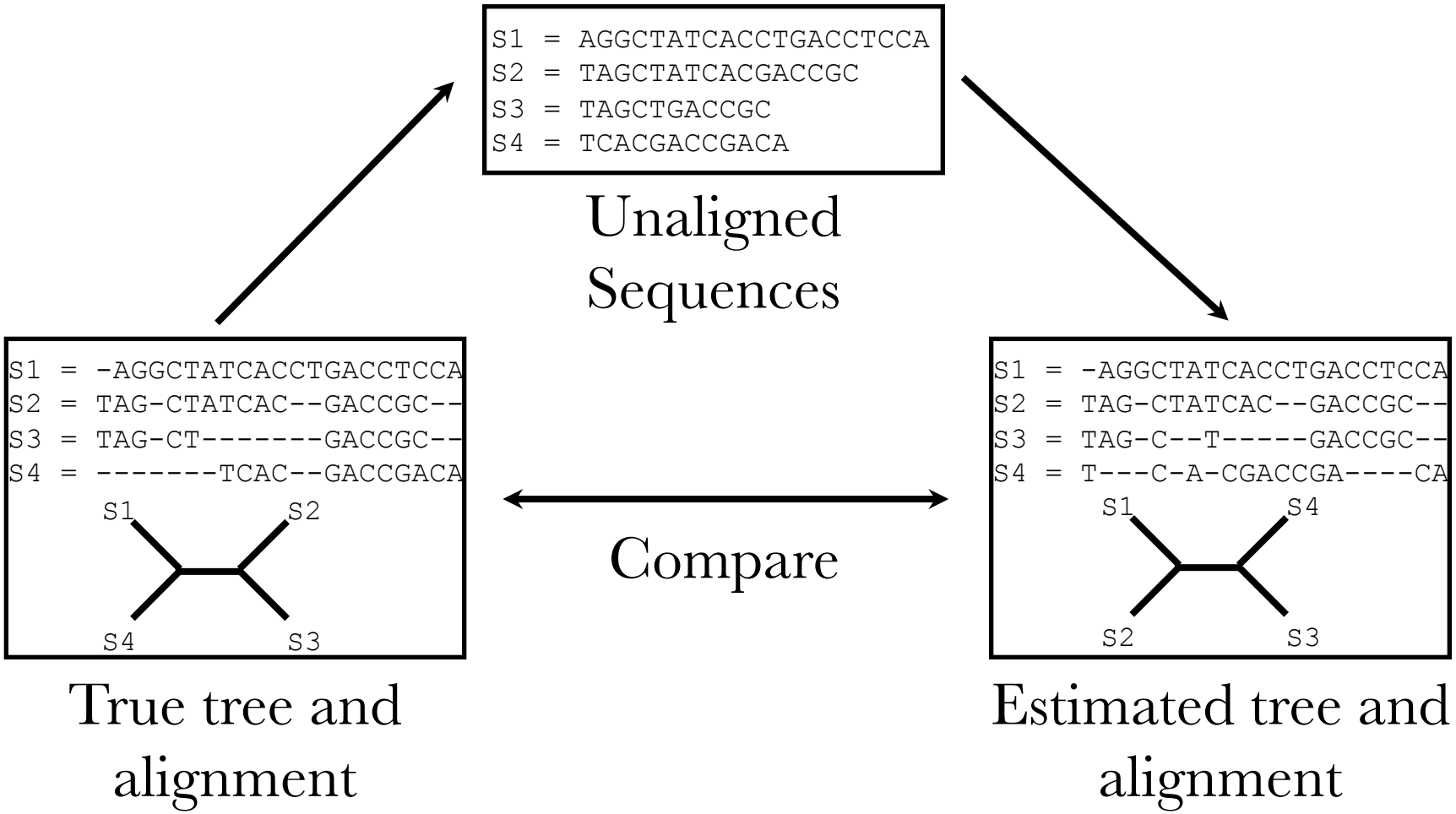
```
S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC
S3 = TAGCTGACCGC
S4 = TCACGACCGACA
```



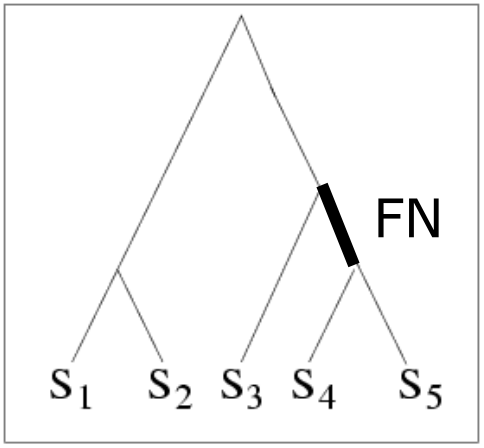
```
S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAC--GACCGC--
S3 = TAG-CT-------GACCGC--
S4 = -------TCAC--GACCGACA
```



Simulation Studies



Quantifying Error



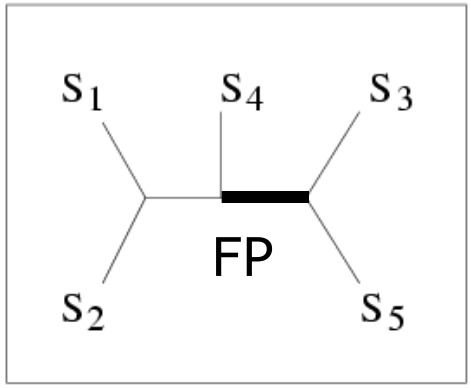
TRUE TREE

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

Two-phase estimation

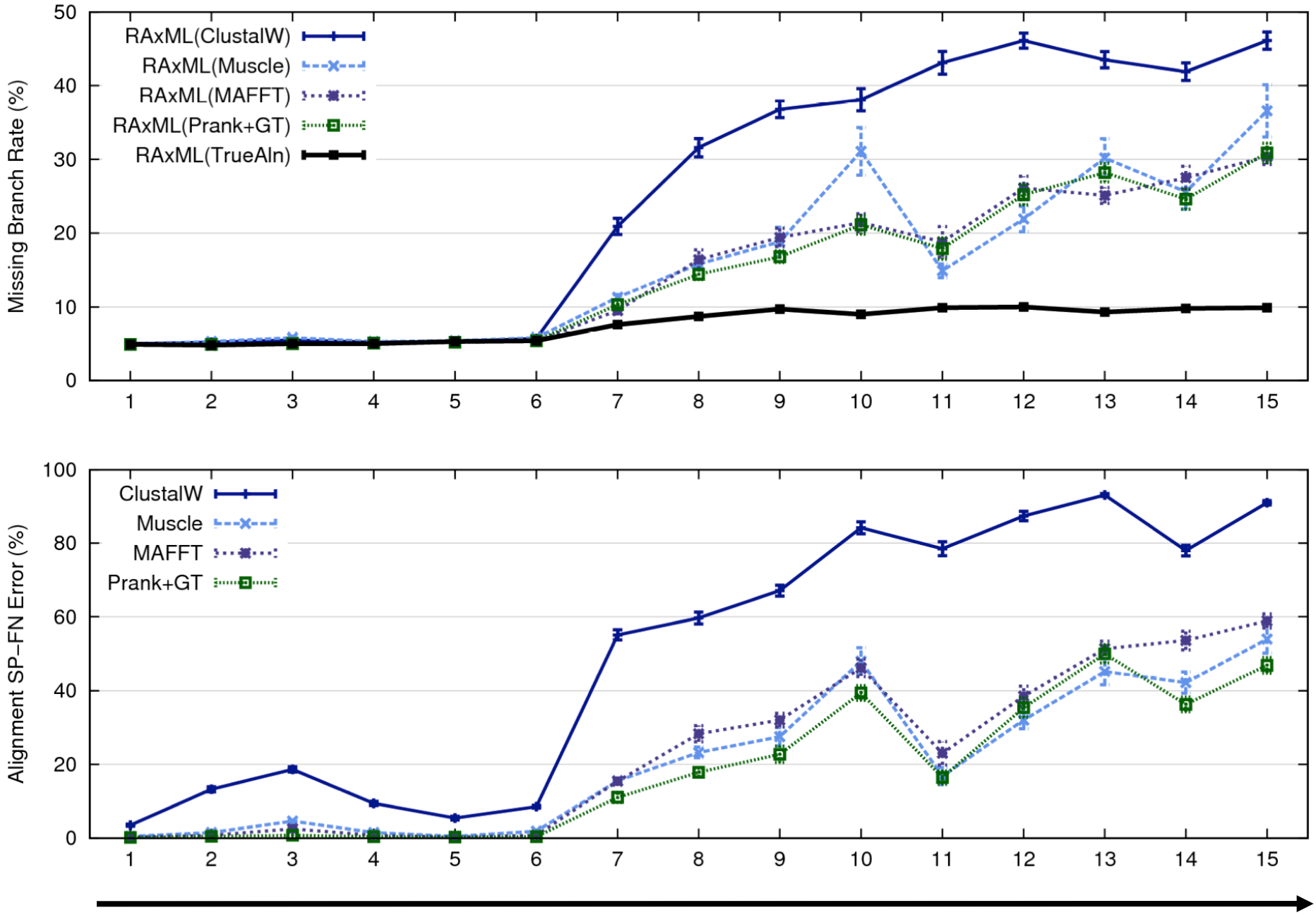
Alignment methods

- 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.

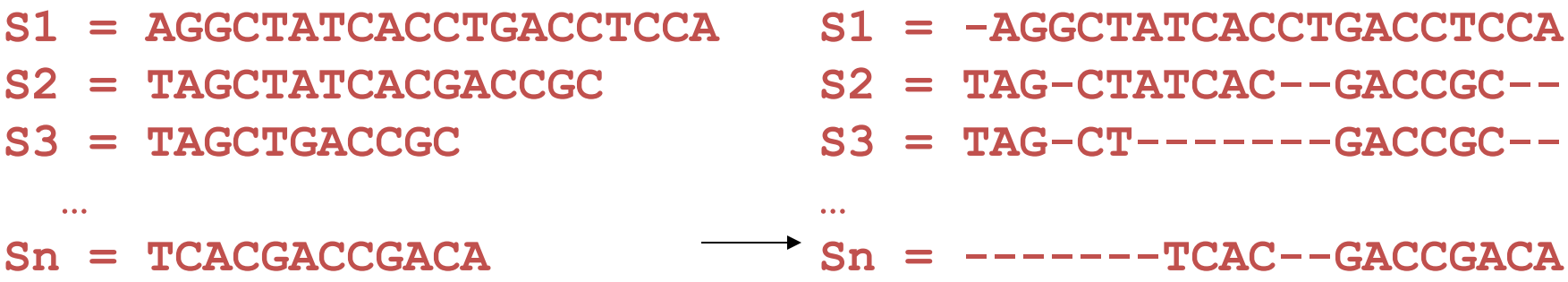
***RAxML**: heuristic for large-scale ML optimization*


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)

Large-scale Alignment Estimation

- Many genes are considered unalignable due to high rates of evolution
- Only a few methods can analyze large datasets
- Alignment error increases with rate of evolution, and results in tree estimation error

Multiple Sequence Alignment (MSA): *a scientific grand challenge*[1]



Novel techniques needed for scalability and accuracy

NP-hard problems and large datasets

Current methods do not provide good accuracy

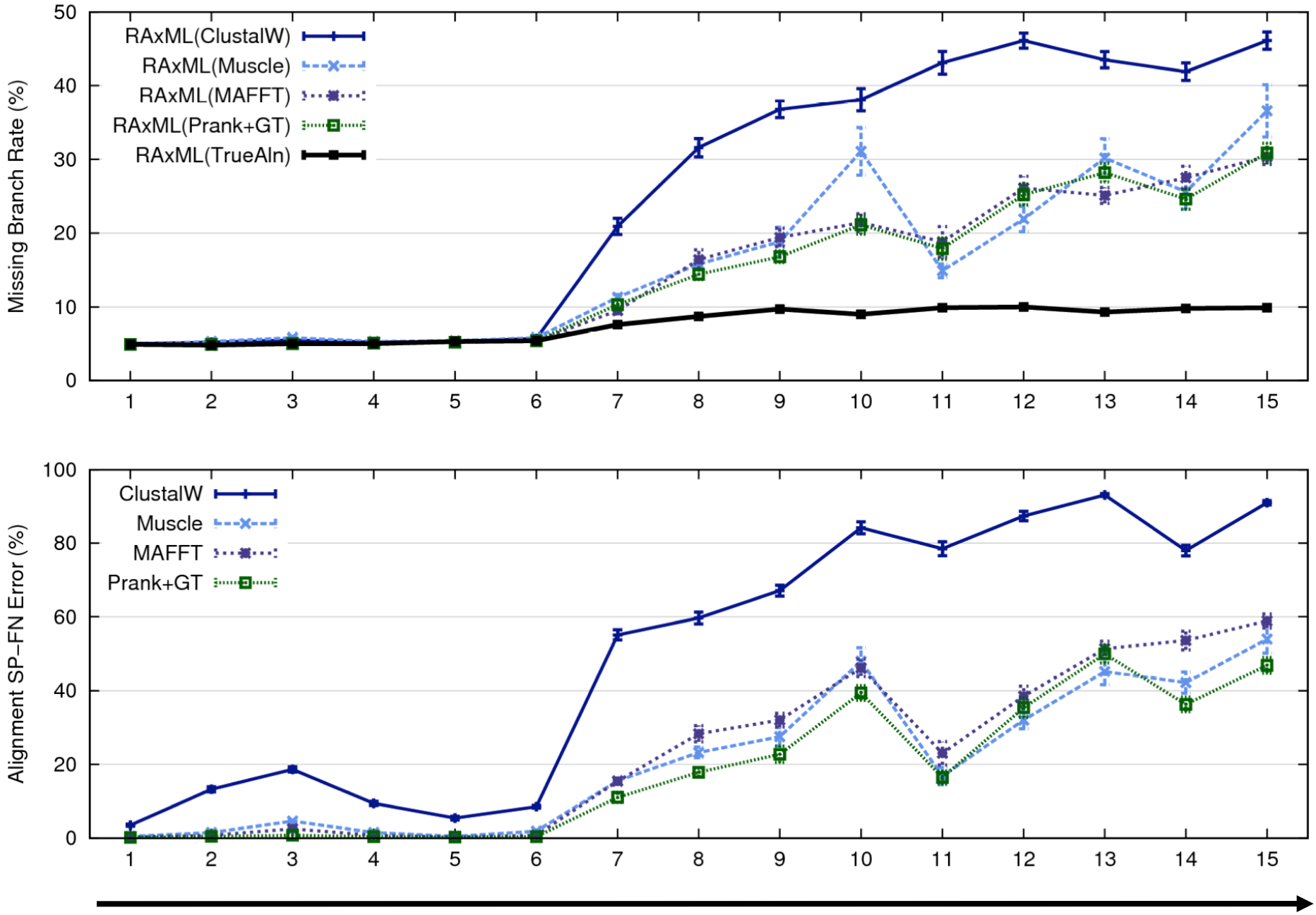
Few methods can analyze even moderately large datasets

Many important applications besides phylogenetic estimation

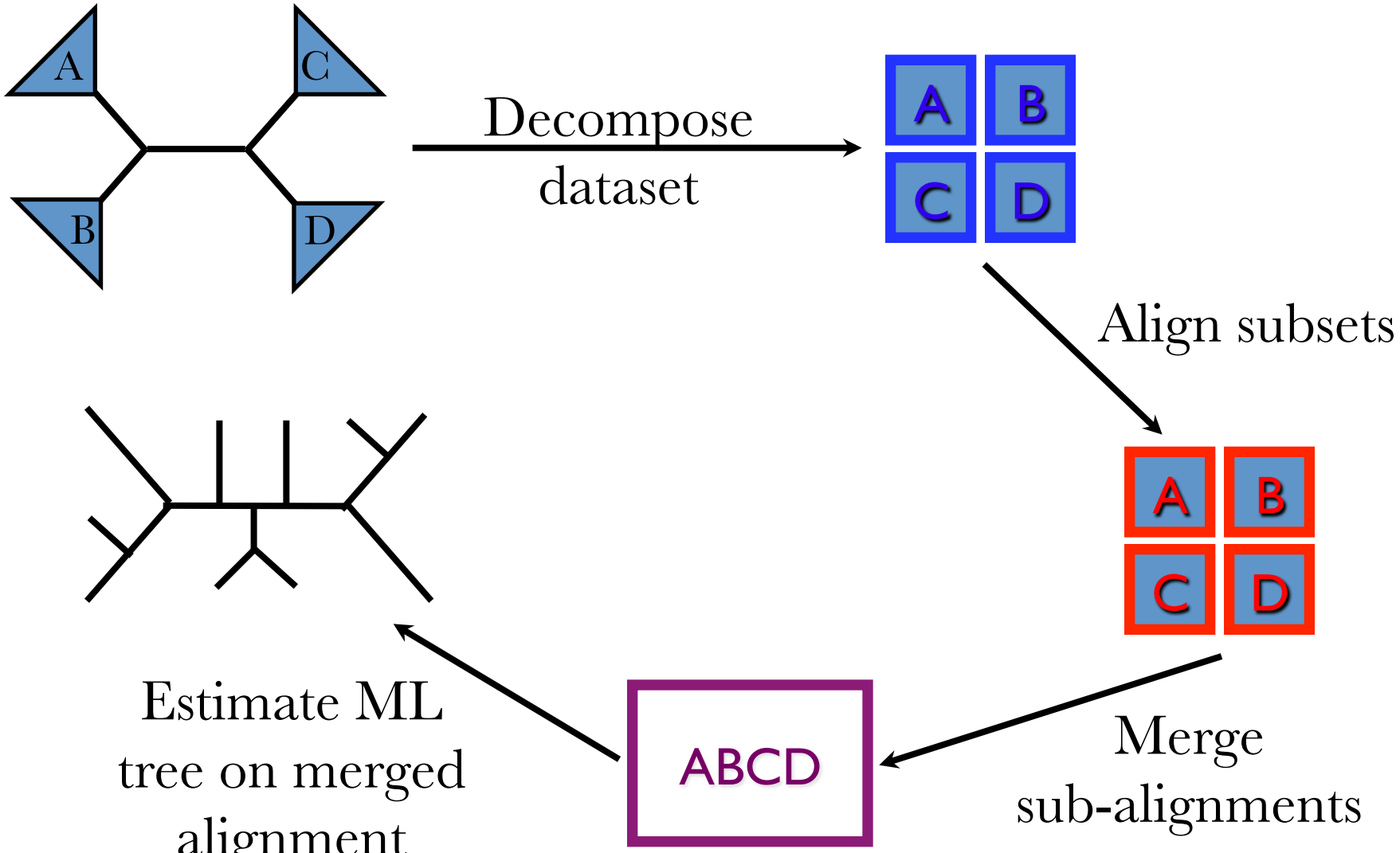
[1] Frontiers in Massive Data Analysis, National Academies Press, 2013

Large-scale Alignment Estimation

- Many genes are considered unalignable due to high rates of evolution
- Only a few methods can analyze large datasets
- Alignment error increases with rate of evolution, and results in tree estimation error

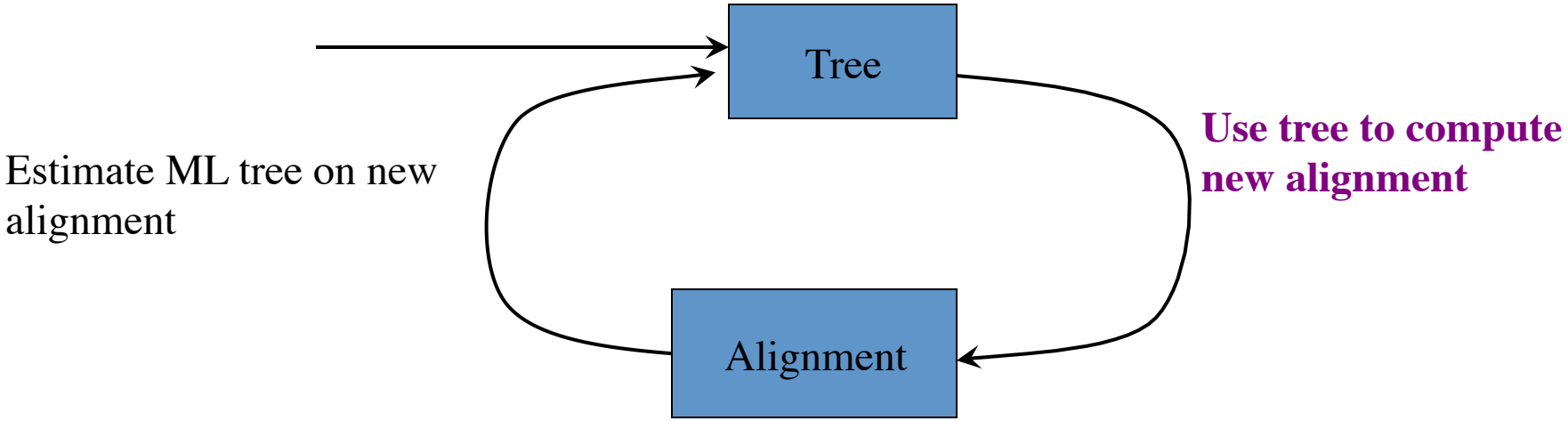

Missing Branch Rate (%)
RAxML(ClustalW)
RAxML(Muscle)
RAxML(MAFFT)
RAxML(Prank+GT)
RAxML(TrueAln)
Alignment SP–FN Error (%)
ClustalW
Muscle
MAFFT
Prank+GT
1000-taxon models, ordered by difficulty (Liu et al., 2009)

Re-aligning on a tree



SATé and PASTA Algorithms

Obtain initial alignment and estimated ML tree

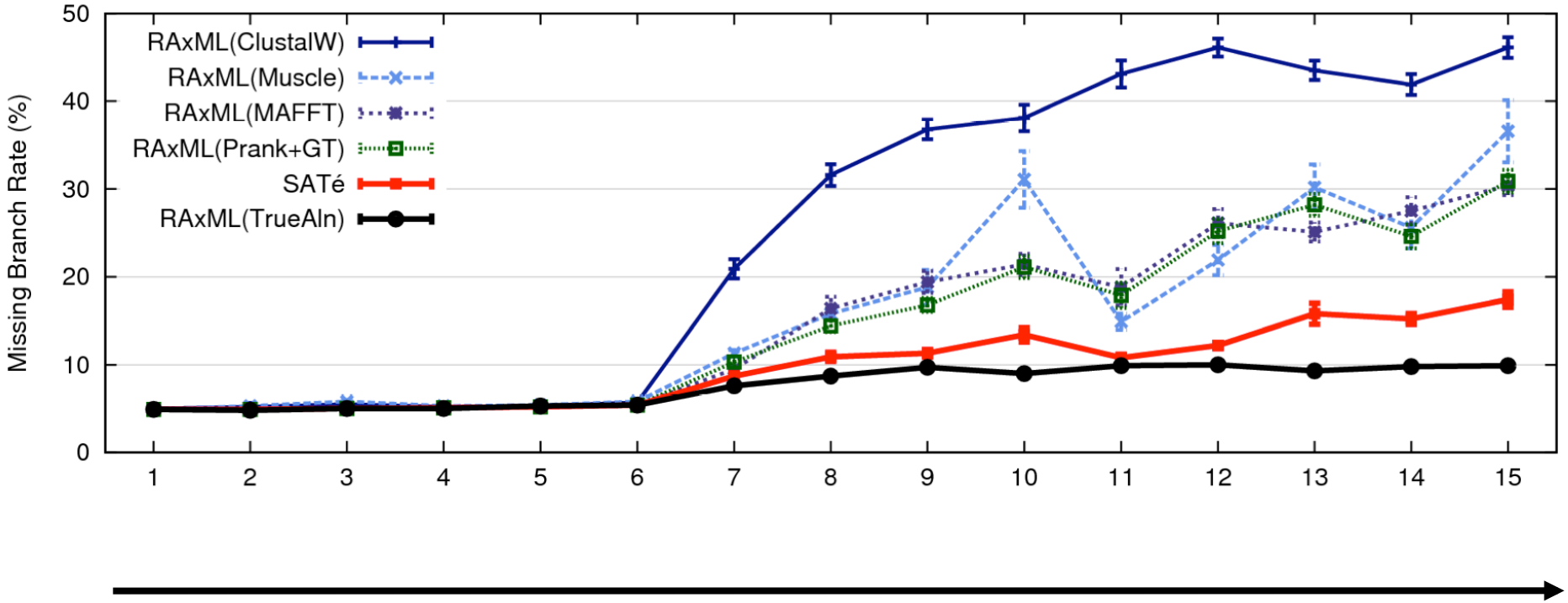


Repeat until termination condition, and

return the alignment/tree pair with the best ML score

SATé-1: Liu et al., Science 2009; SATé-2, Liu et al., Systematic Biology 2012;
PASTA: Mirarab et al., RECOMB 2014 and J. Computational Biology 2014

SATé-1 (Science 2009) performance



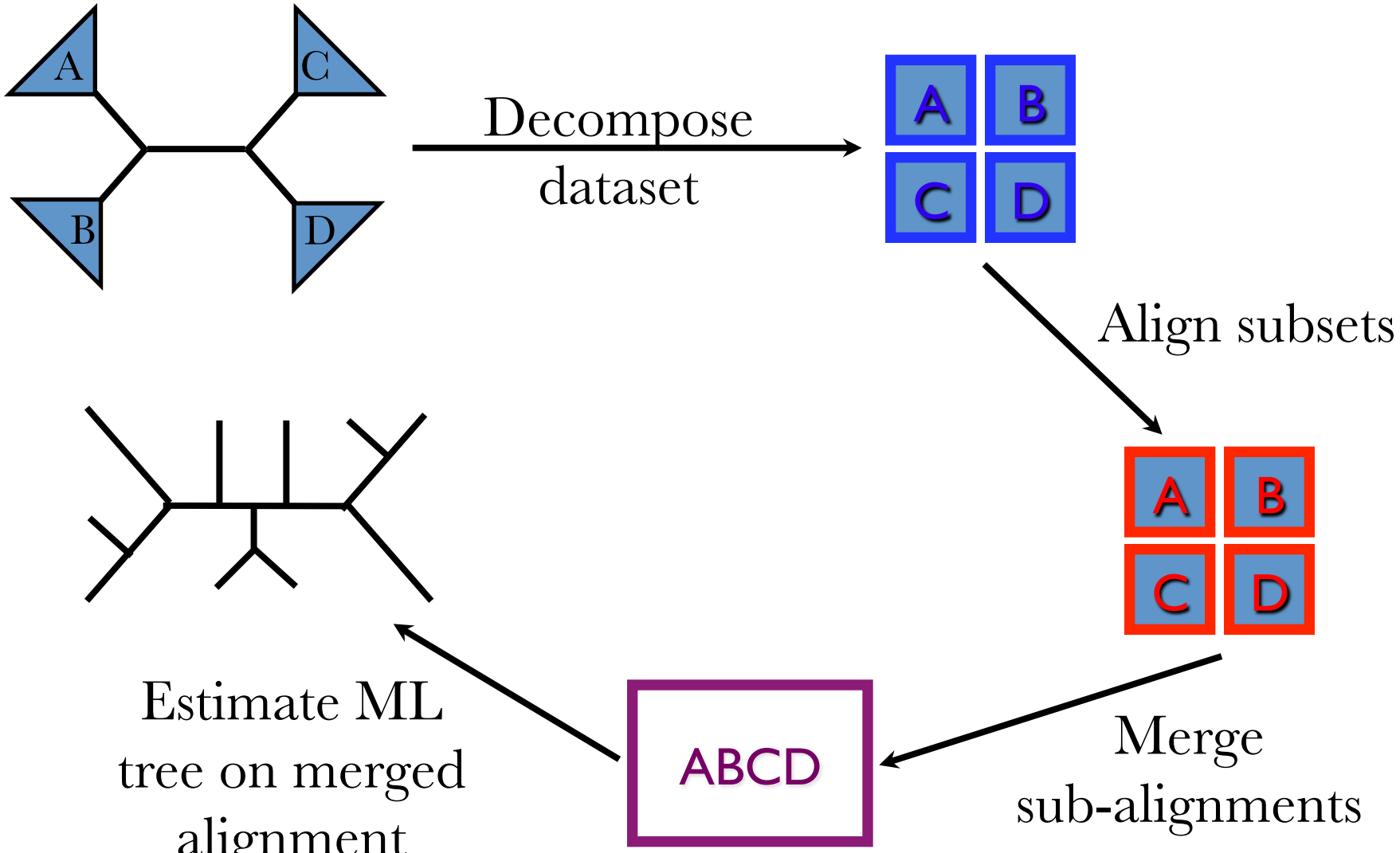
1000-taxon models, rate of evolution generally increases from left to right

SATé-1 is based on MAFFT to align subsets. Results shown are for a 24-hour analysis, on desktop machines.

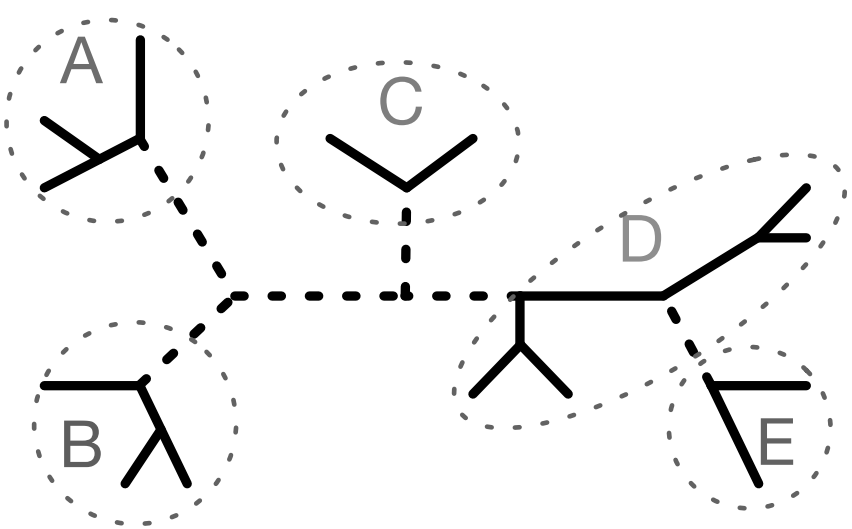
Similar improvements seen for biological datasets.

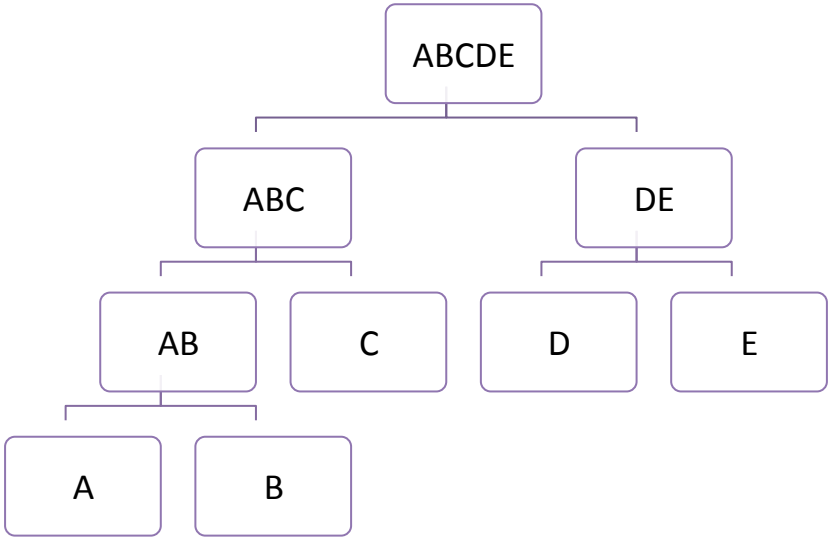
SATé-1 can analyze up to about 8,000 sequences.

SATé-I decomposition (clades)



SATé-II: centroid edge decomposition

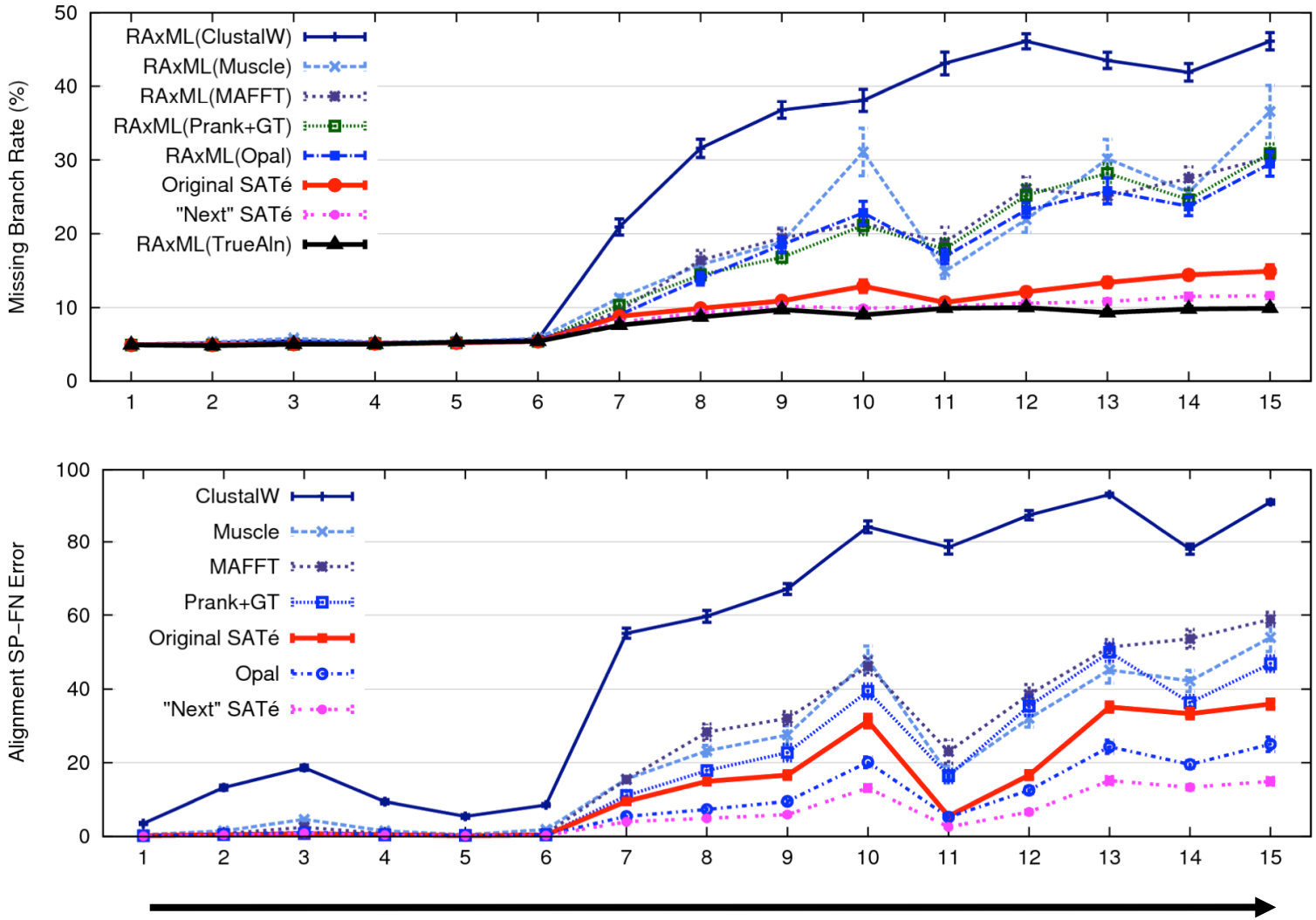




SATé-II makes all subsets small (user parameter), and can analyze 50K sequences.

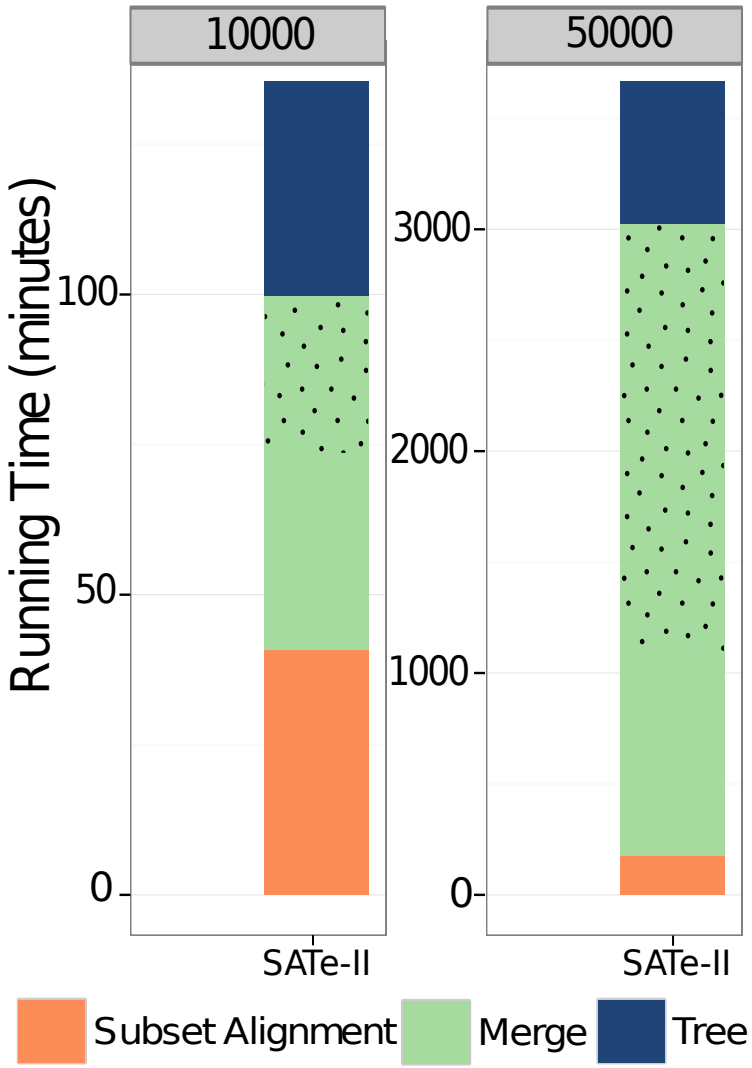
(Recall that the SATé-I decomposition produced clades and had bigger subsets; limited to 8K sequences.)

SATé-1 and SATé-2 (Systematic Biology, 2012)

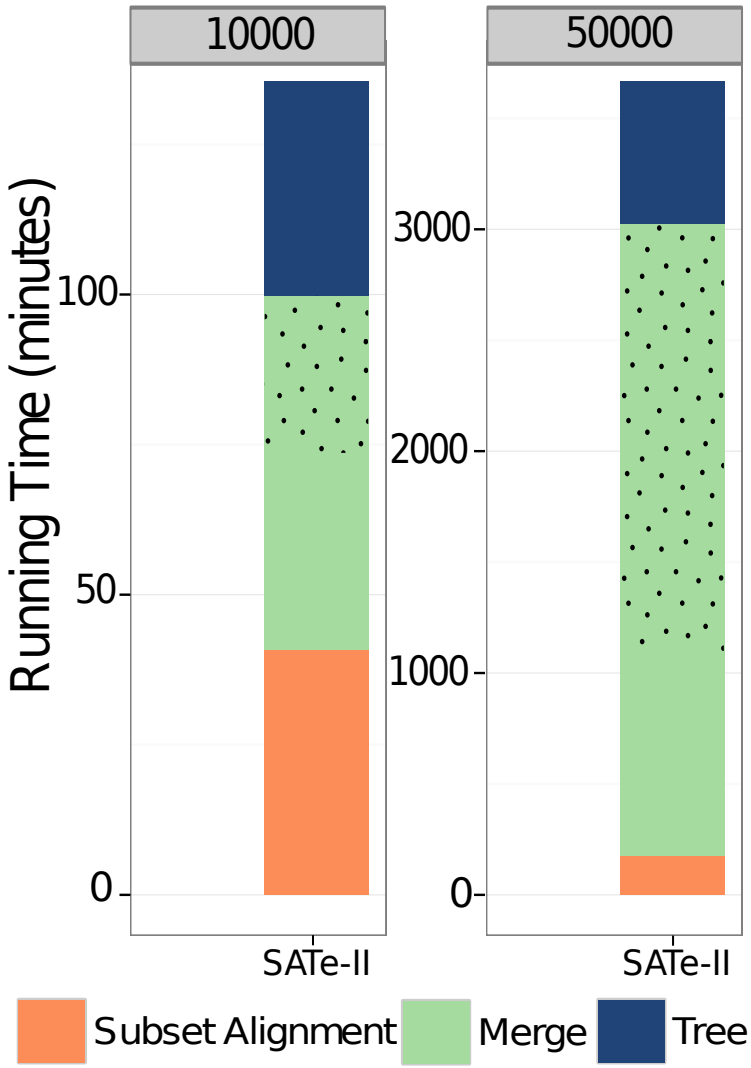


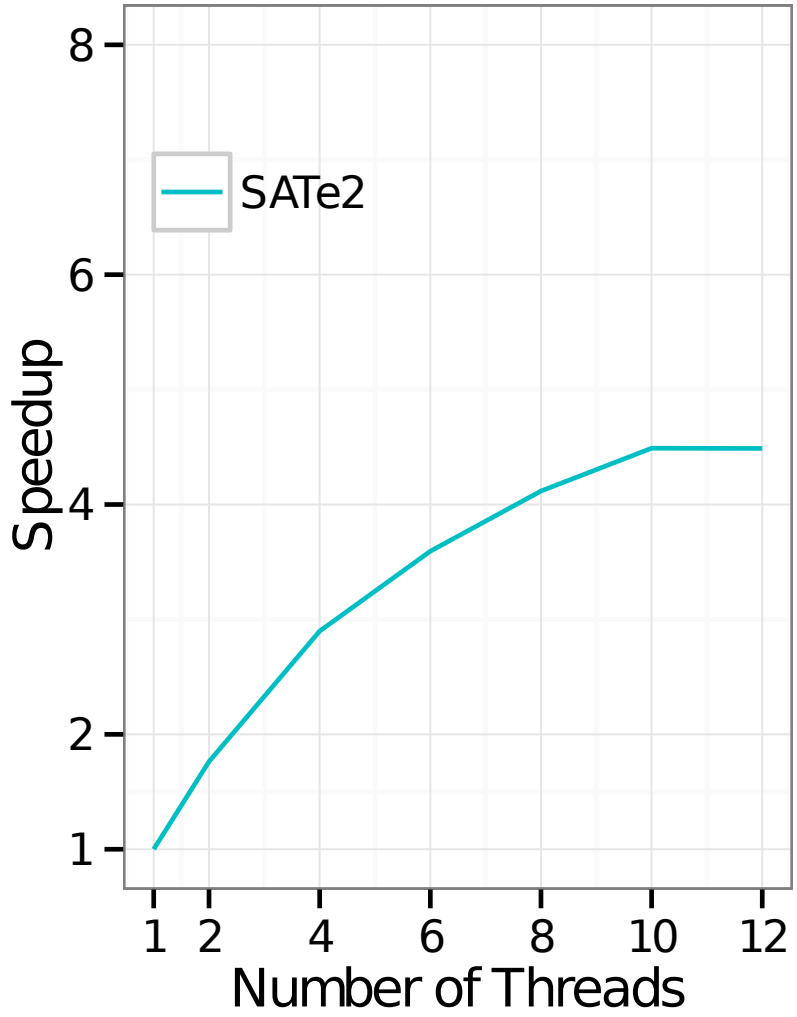
SATé-1: up to 8K
SATé-2: up to ~50K

SATé-II running time profiling

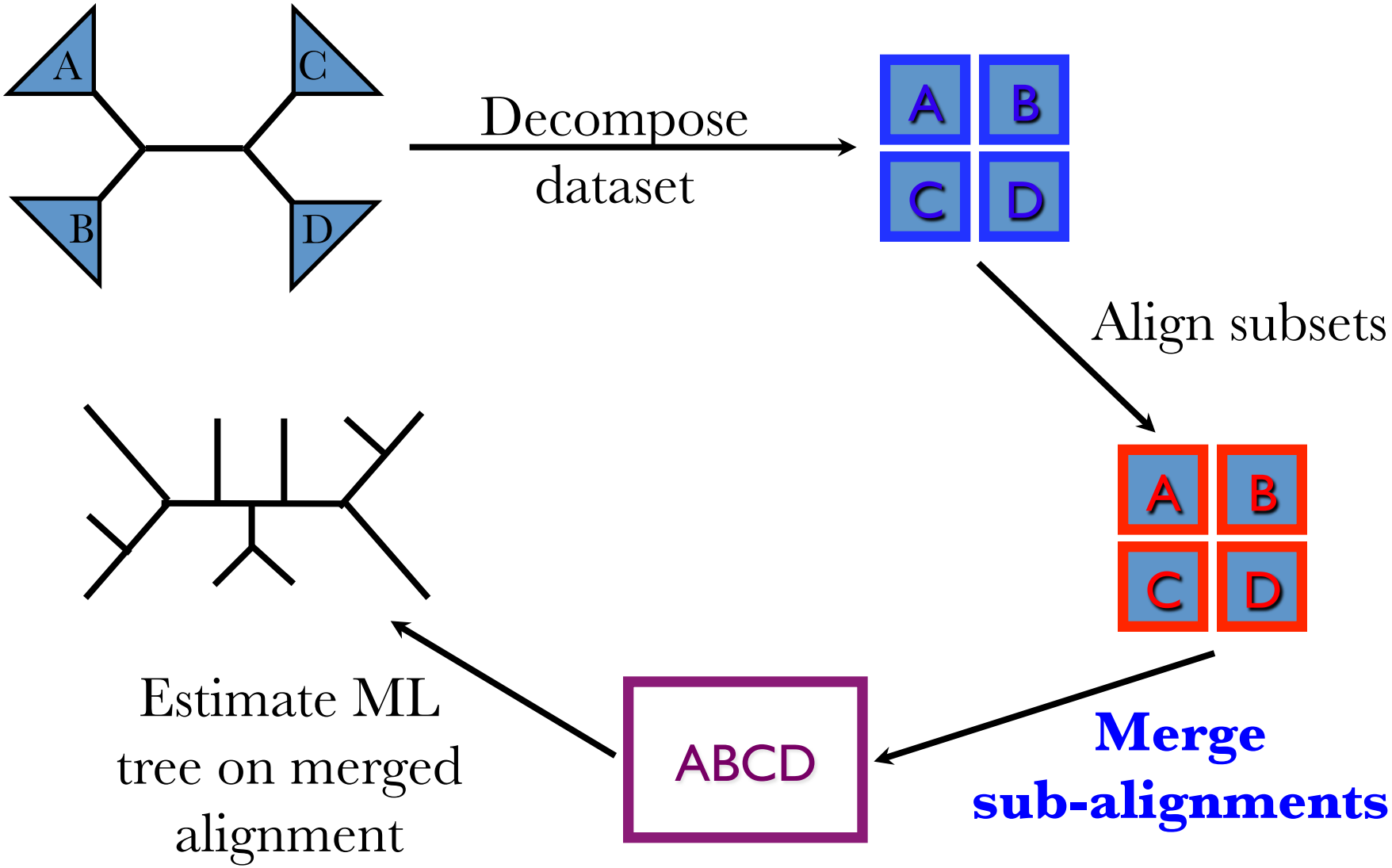


SATé-II running time profiling

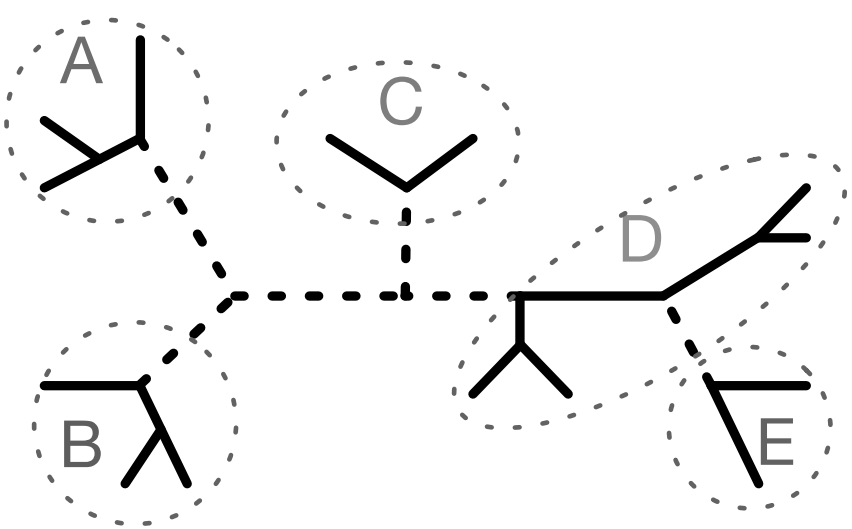


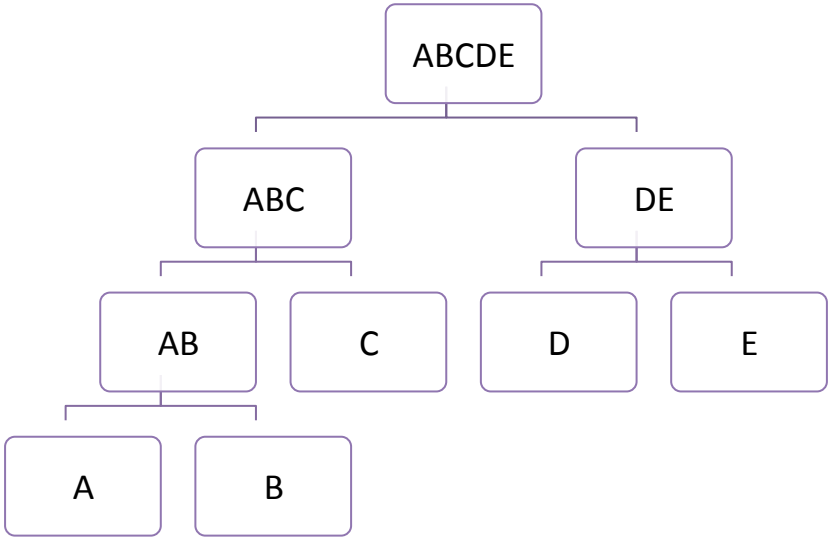


PASTA: SATé-II with a new merger technique



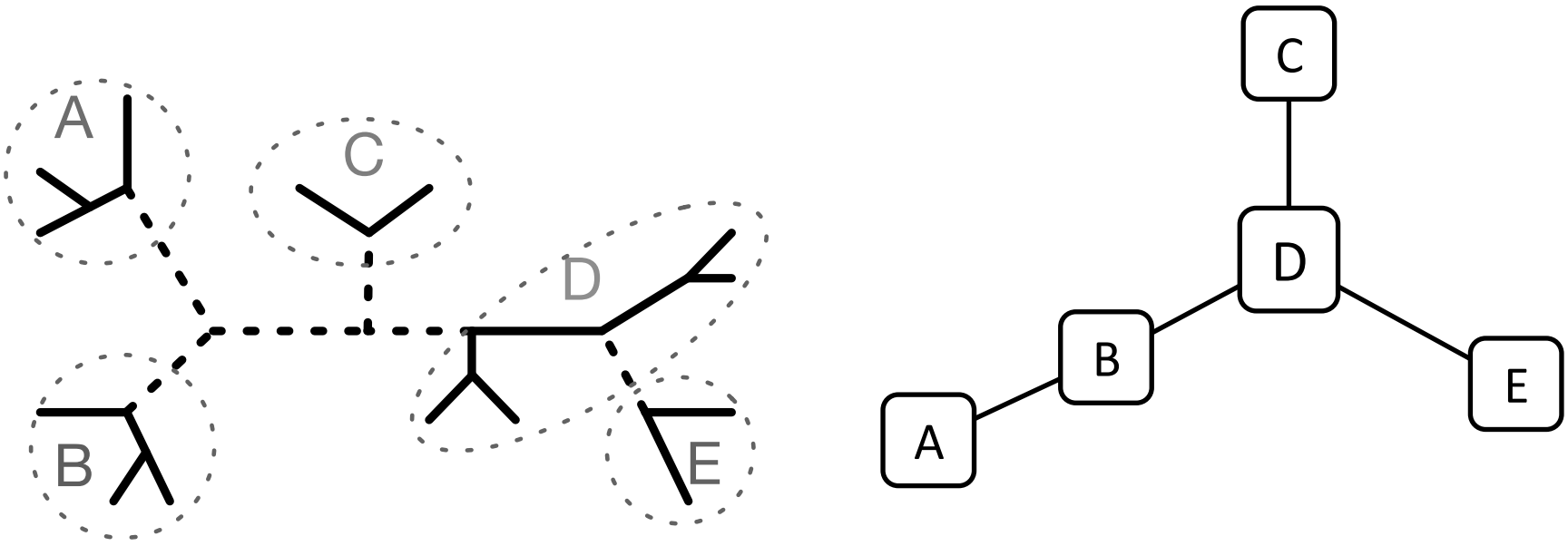
SATé: merger strategy





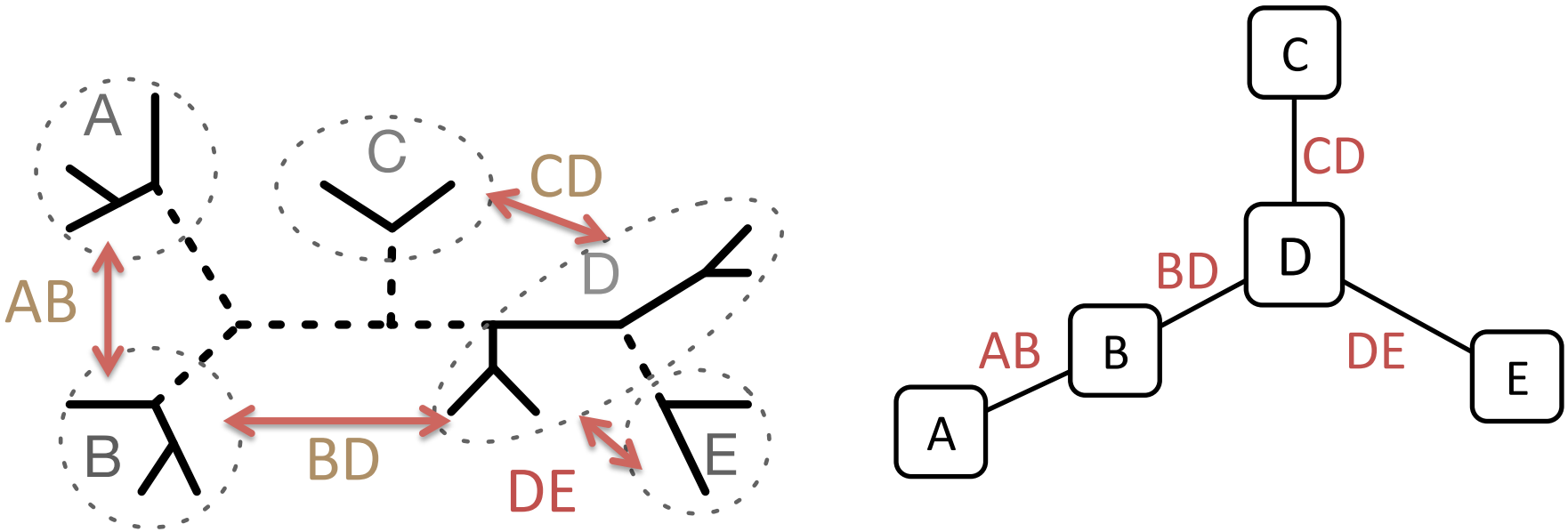
Both SATé's use the same hierarchical merger strategy.
On large (50K) datasets, the last pairwise merger can use more than 70% of the running time

PASTA merging: Step 1



Compute a spanning tree connecting alignment subsets

PASTA merging: Step 2



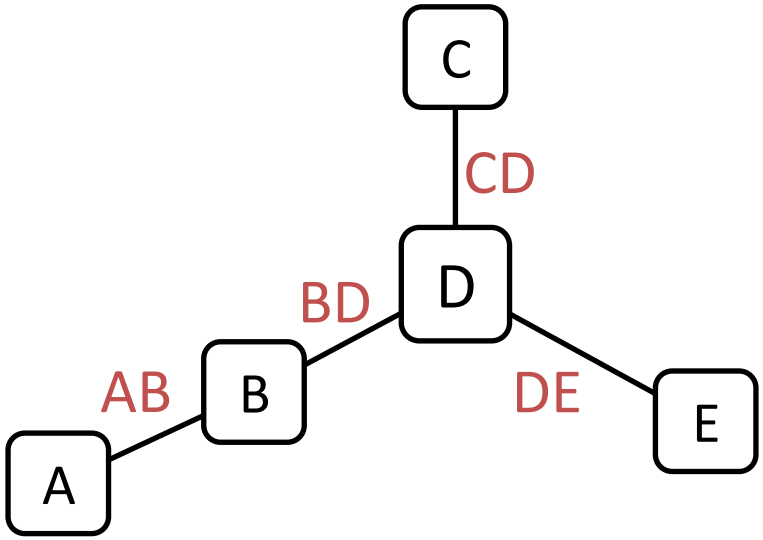
Use Opal (or Muscle) to merge adjacent subset alignments in the spanning tree

PASTA merging: Step 3

AB + BD = ABD

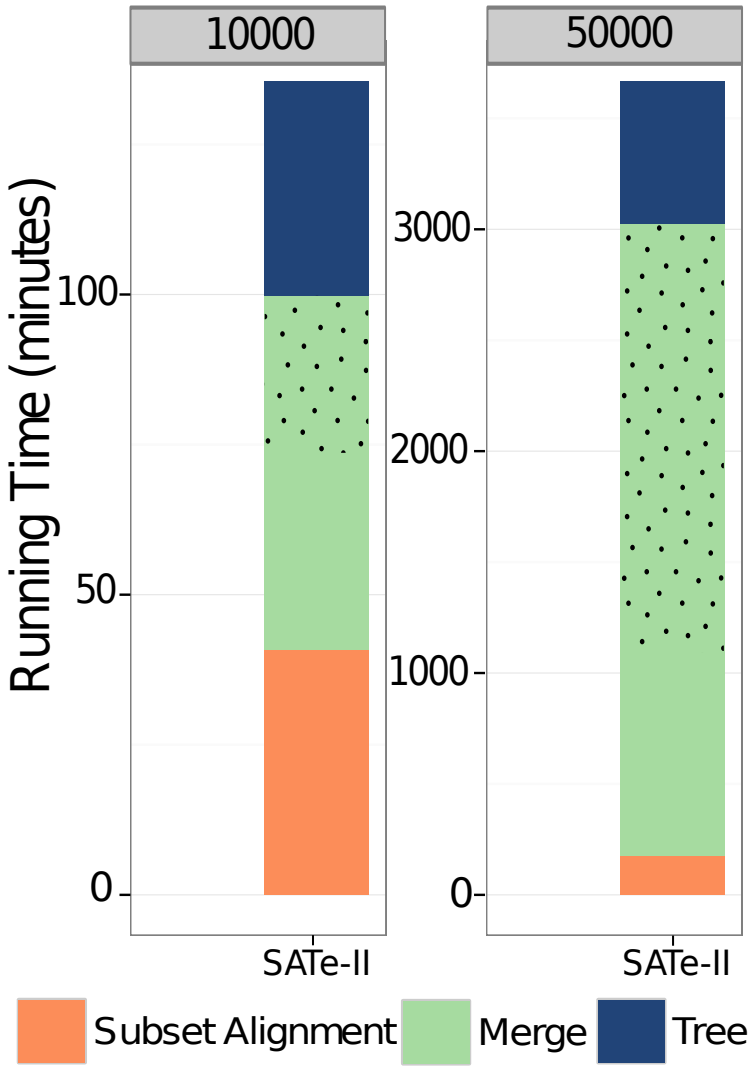
ABD + CD = ABCD

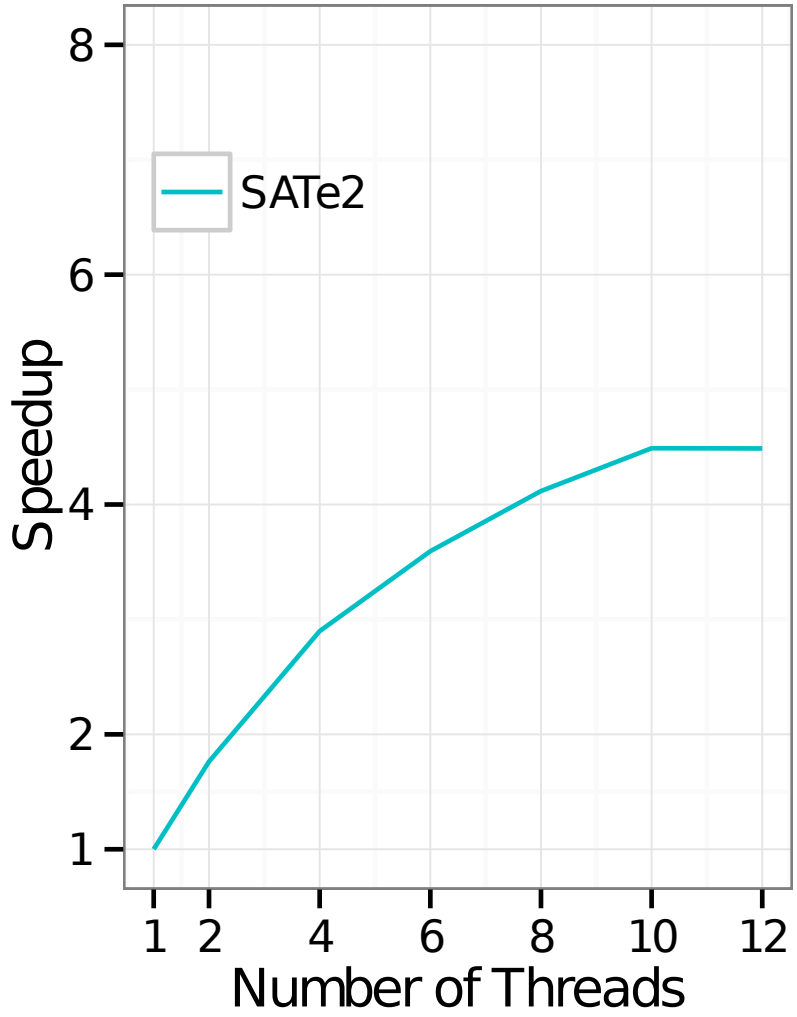
ABCD + DE = ABCDE



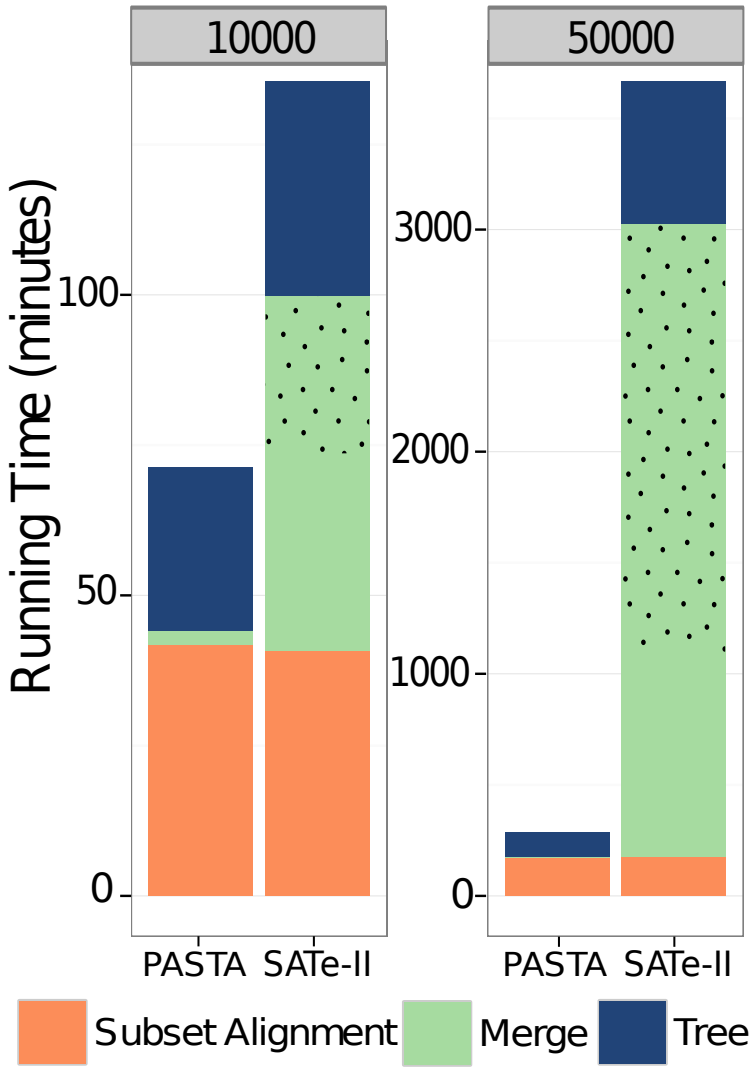
Use transitivity to merge all pairwise-merged alignments from Step 2 into final an alignment on entire dataset

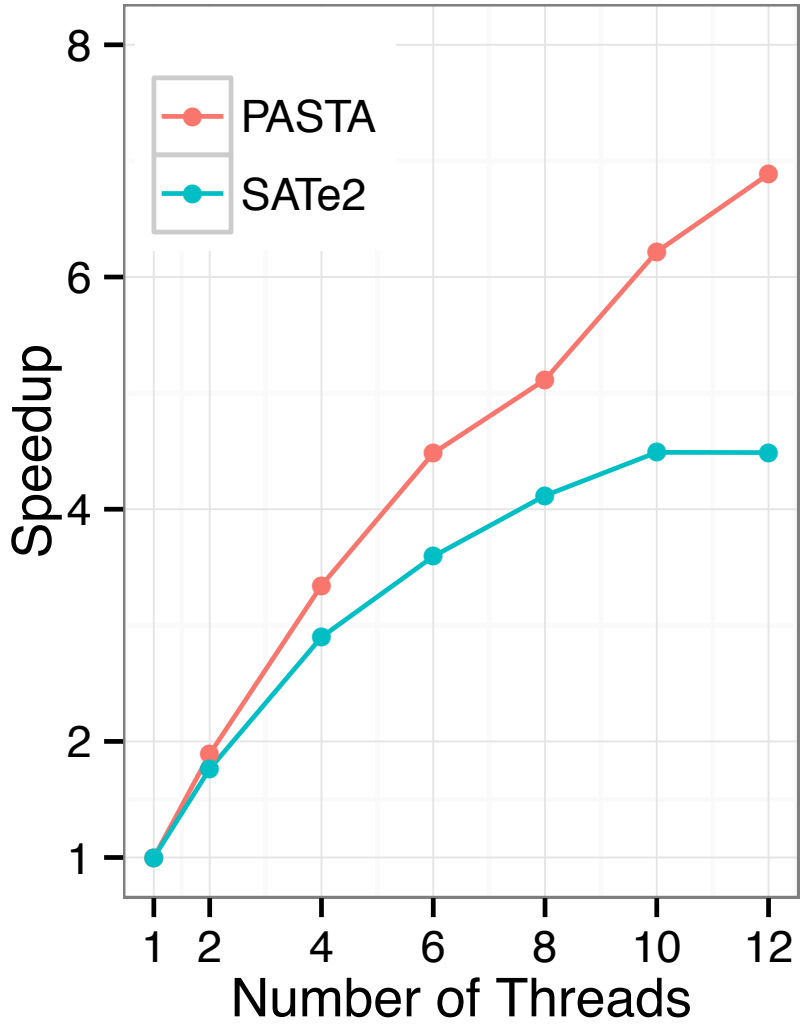
SATé-II running time profiling



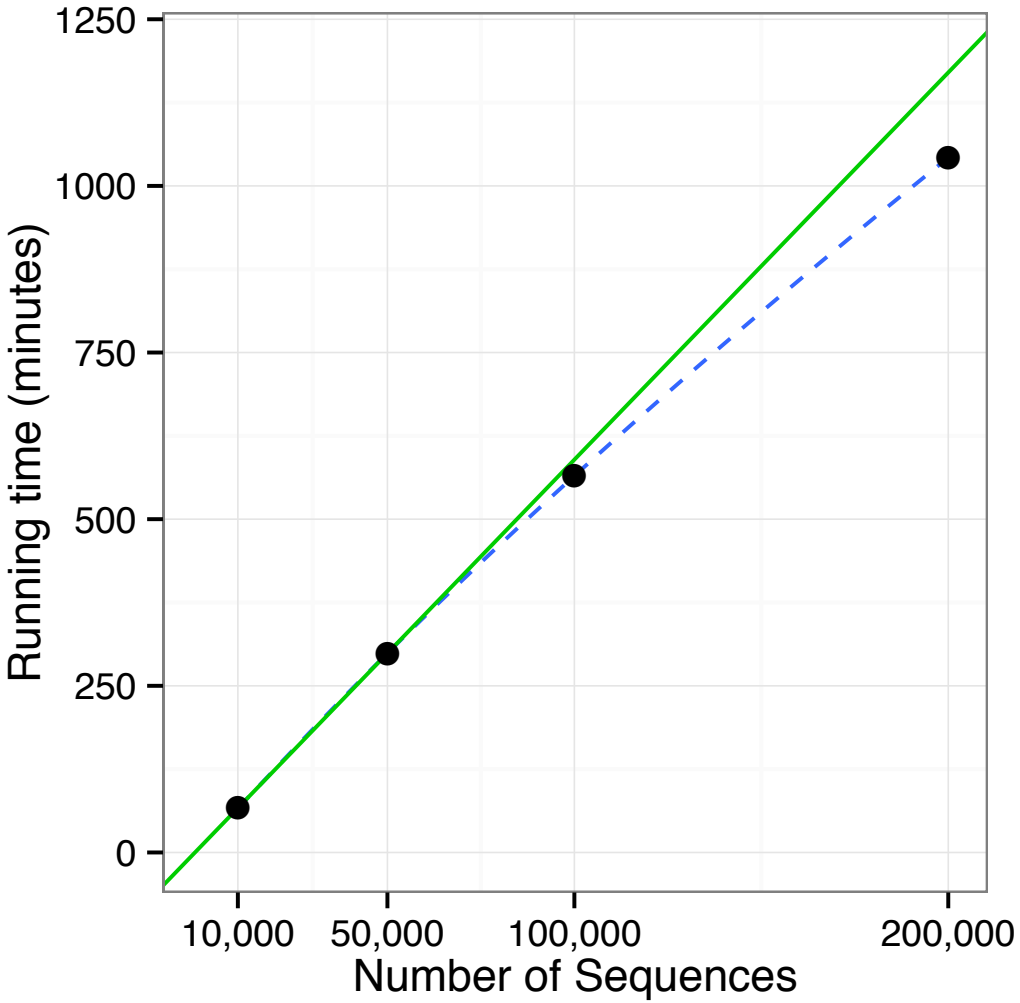


PASTA vs. SATé-II profiling and scaling





PASTA Running Time and Scalability



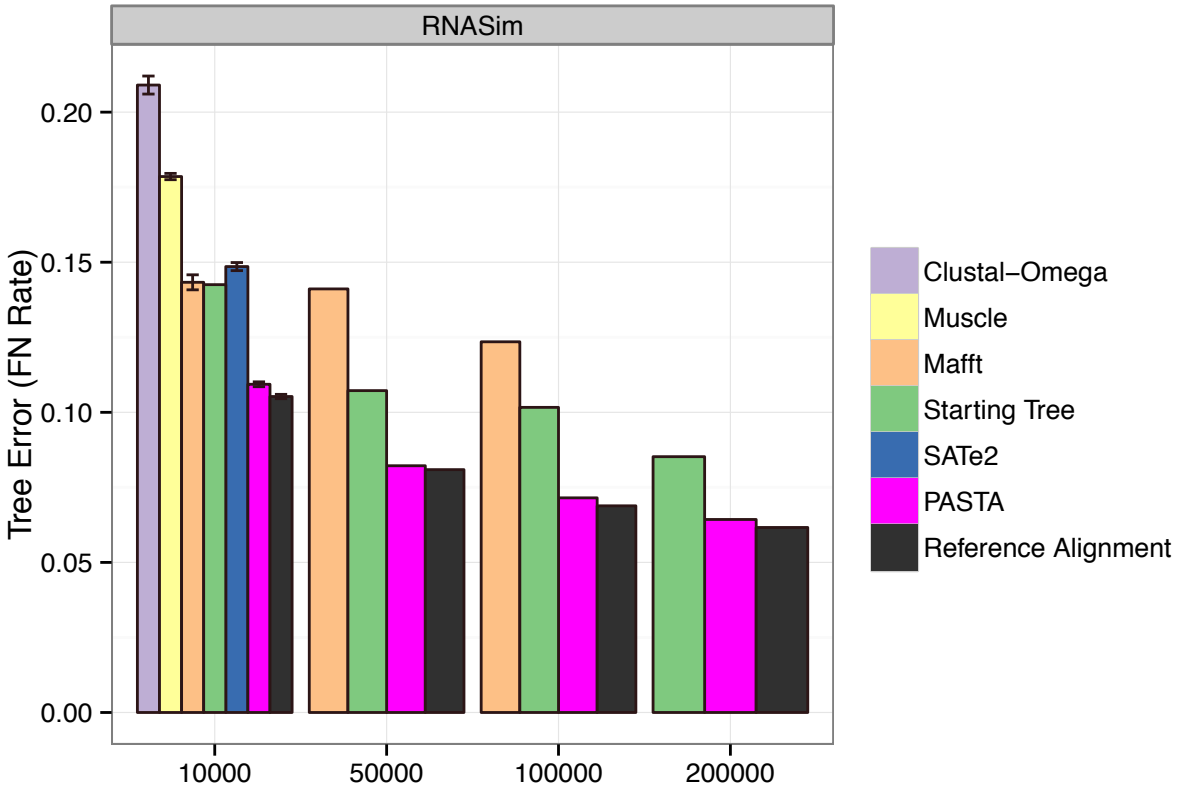
- One iteration
- Using
  - 12 cpus
  - 1 node on Lonestar TACC
  - Maximum 24 GB memory
- Showing wall clock running time
  - ~ 1 hour for 10k taxa
  - ~ 17 hours for 200k taxa

Massive Parallelism in PASTA

- Division step: very fast (not worse than $O(n^2)$)
- So 1,000,000-sequence dataset becomes:
  - ~5000 subsets of 200 sequences each
  - Each analyzed independently
  - Can tailor subset analysis to features of the data
- Merging step: very fast and also massively parallel (independent pairwise mergers, then transitivity)

The only part of PASTA that isn't well parallelized is the tree estimation step in each iteration!

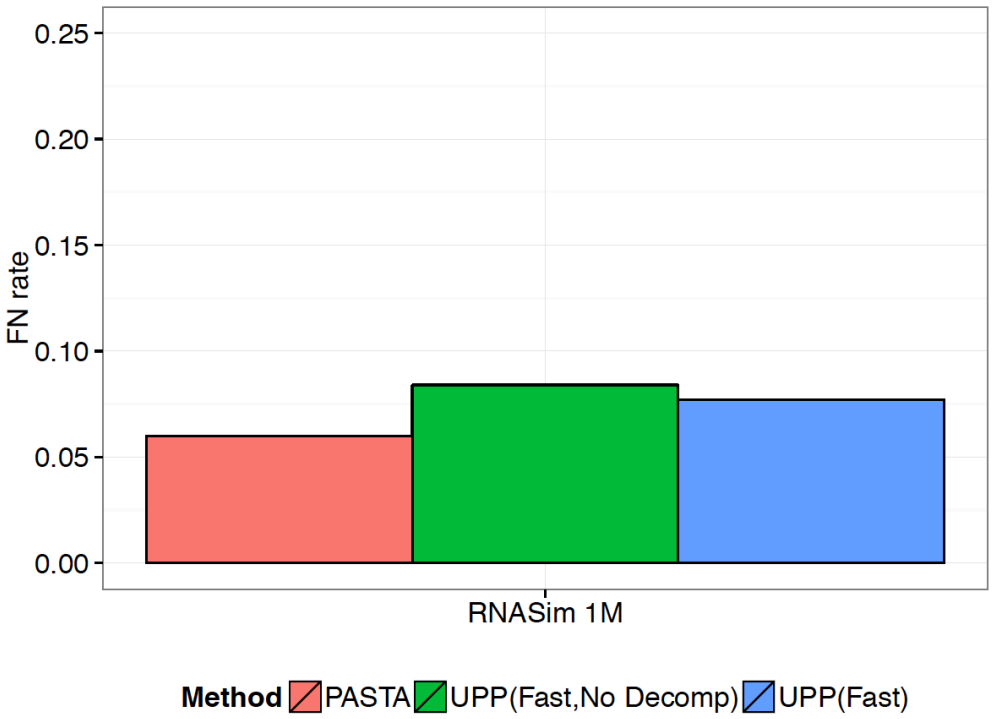
PASTA: even better than SATé-2



PASTA: Mirarab, Nguyen, and Warnow, J Comp. Biol. 2015

- Simulated RNASim datasets from 10K to 200K taxa
- Limited to 24 hours using 12 CPUs
- Not all methods could run (missing bars could not finish)

RNASim Million Sequences: tree error



Using 12 TACC processors:

- UPP(Fast,NoDecomp) took 2.2 days,
- UPP(Fast) took 11.9 days, and
- PASTA took 10.3 days

UPP: Nguyen et al. RECOMB 2015 and Genome Biology 2015 (also uses divide-and-conquer and is highly parallelizable.)

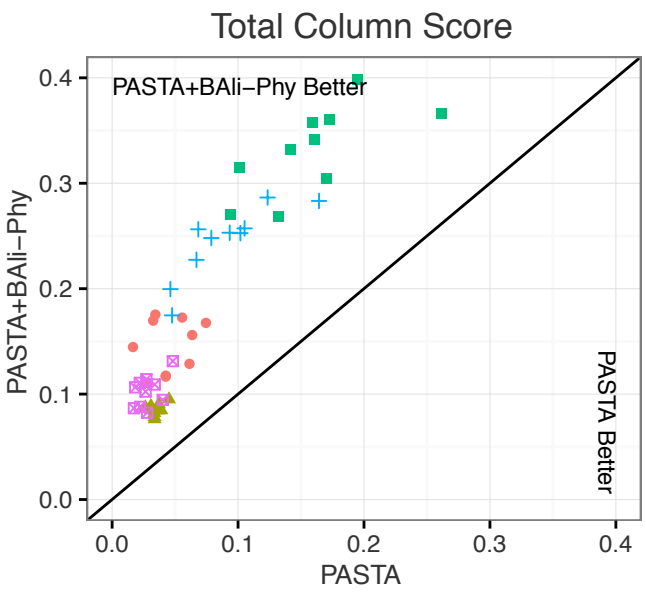
PASTA and SATé-II: MSA “boosters”

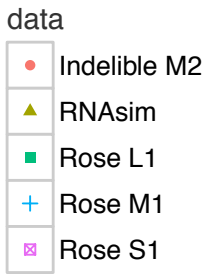
- PASTA and SATé-II are techniques for improving the scalability of MSA methods to large datasets.
- We showed results here using MAFFT-l-ins-i to align small subsets with 200 sequences.

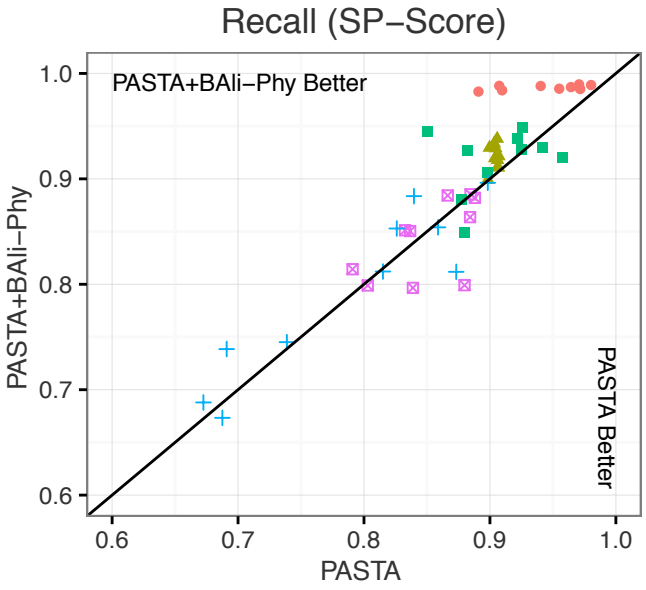
Gold Standard: Statistical co-estimation

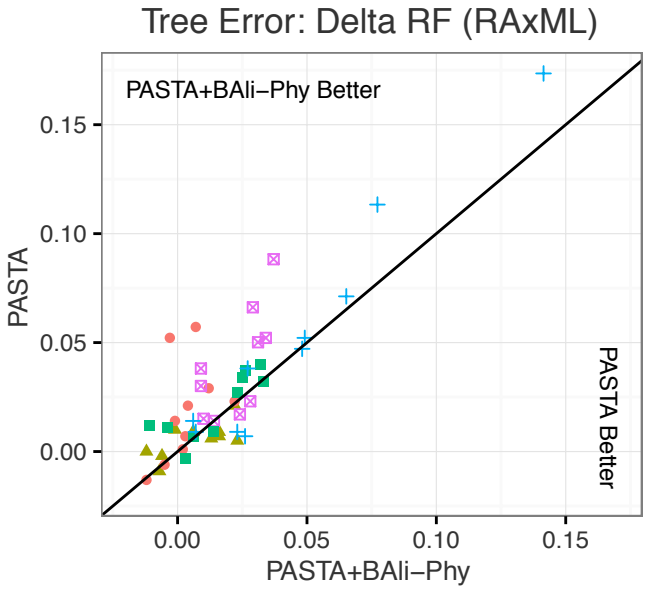
- Improved accuracy can be obtained through co-estimation of alignments and trees.
- BAli-Phy (Redelings and Suchard, 2005), a Bayesian method, is the leading co-estimation method.
- However, BAli-Phy is limited to small datasets (at most 100 sequences), and even these analyses can take weeks (due to convergence issues).
- We integrated BAli-Phy into PASTA (replacing MAFFT), with decompositions to at most 100 sequences.

Comparing default PASTA to PASTA+BAli-Phy on simulated datasets with 1000 sequences





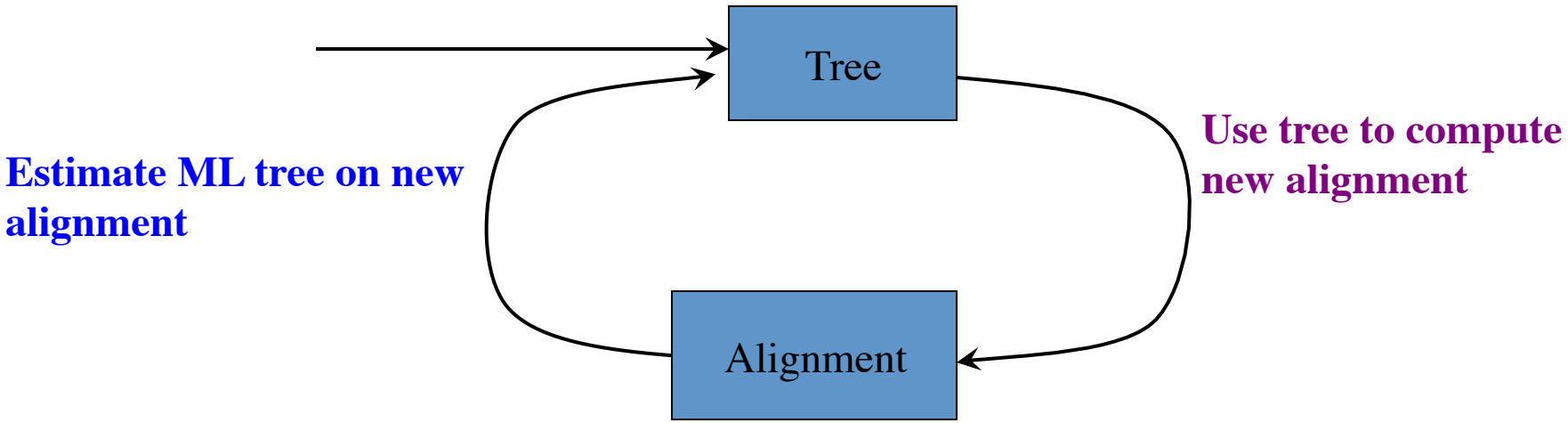




Decomposition to 100-sequence subsets, one iteration of PASTA+BAli-Phy

SATé and PASTA Algorithms

Obtain initial alignment and estimated ML tree



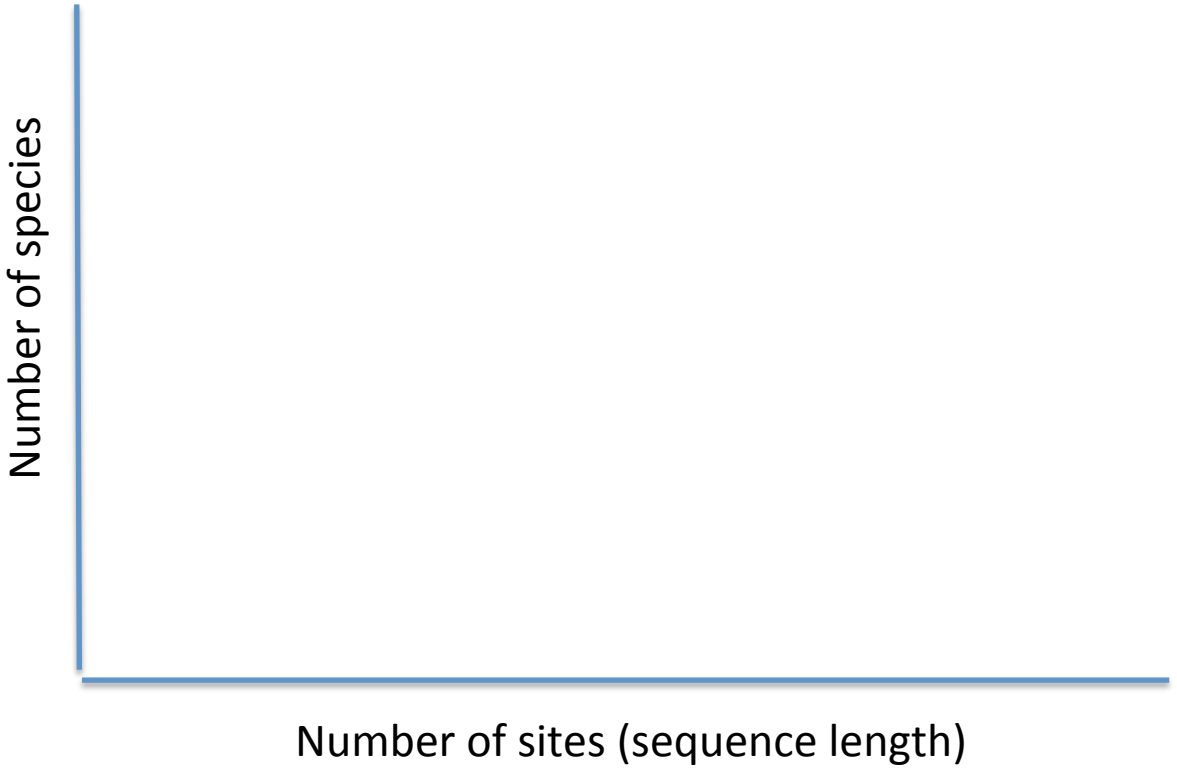
Repeat until termination condition, and

return the alignment/tree pair with the best ML score

Major Open Problem

- Scalable maximum likelihood tree estimation
  - Input: Multiple sequence alignment (and specified model)
  - Output: Tree and numeric parameters to maximize probability of the sequences under the model

Maximum Likelihood Heuristics

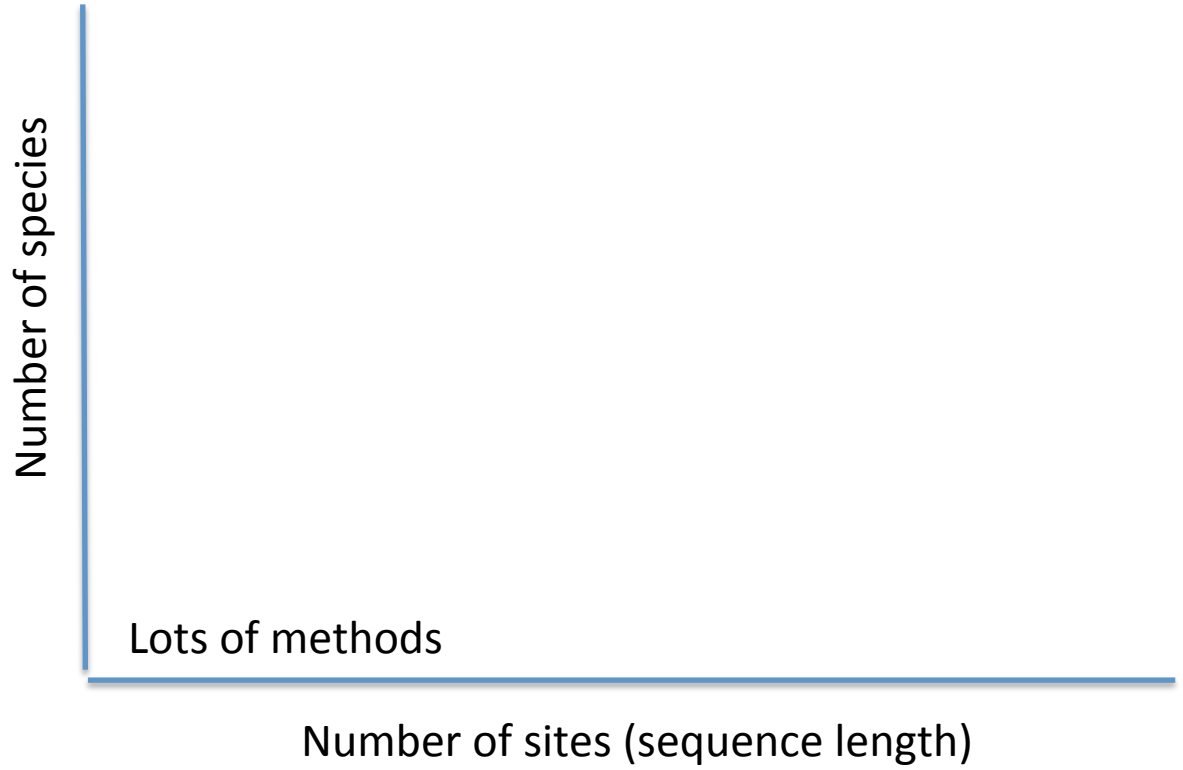


Maximum Likelihood (NP-hard):

Input is a multiple sequence alignment,

Output is a tree maximizing the probability of generating the input sequences.

Maximum Likelihood Heuristics

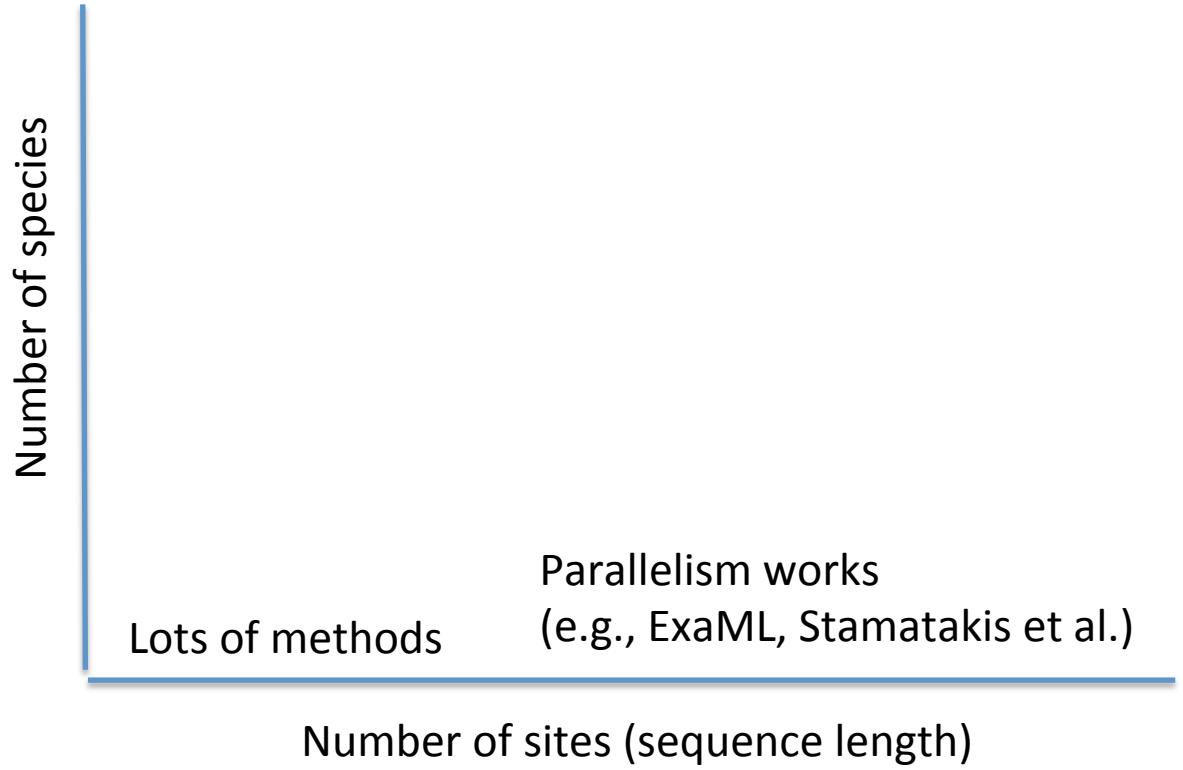


Maximum Likelihood (NP-hard):

Input is a multiple sequence alignment,

Output is a tree maximizing the probability of generating the input sequences.

Maximum Likelihood Heuristics

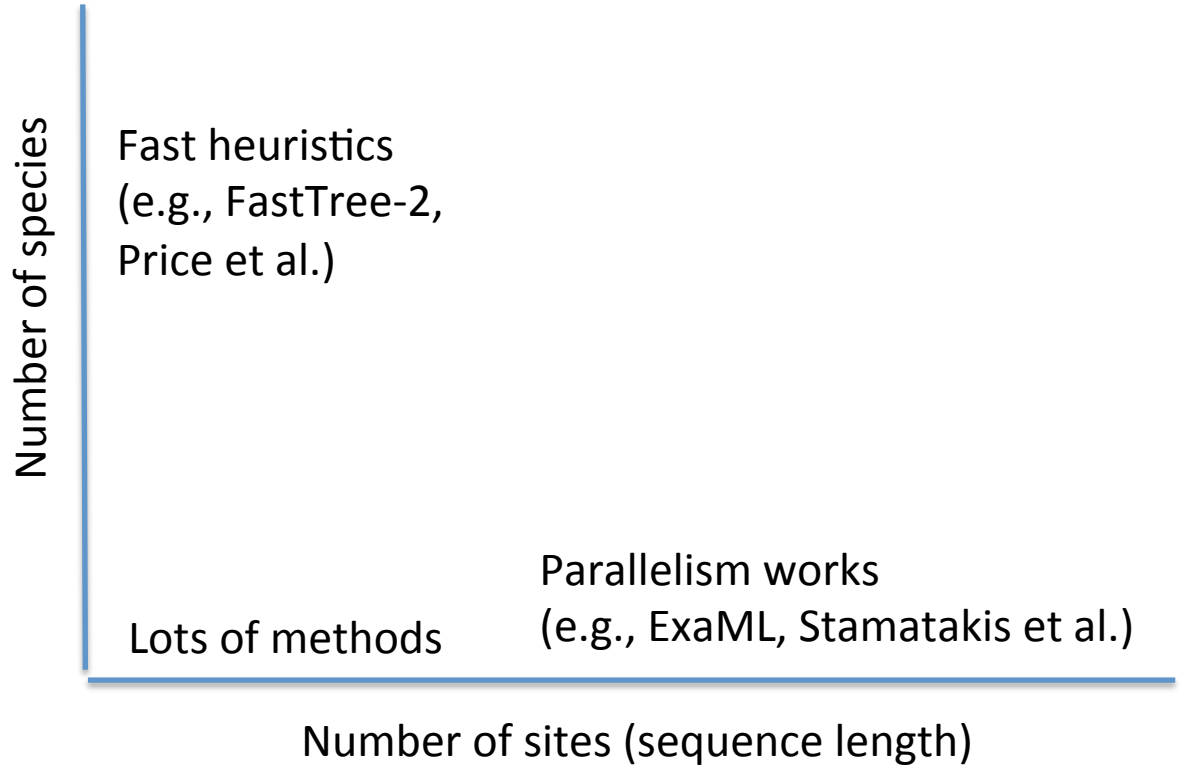


Maximum Likelihood (NP-hard):

Input is a multiple sequence alignment,

Output is a tree maximizing the probability of generating the input sequences.

Maximum Likelihood Heuristics

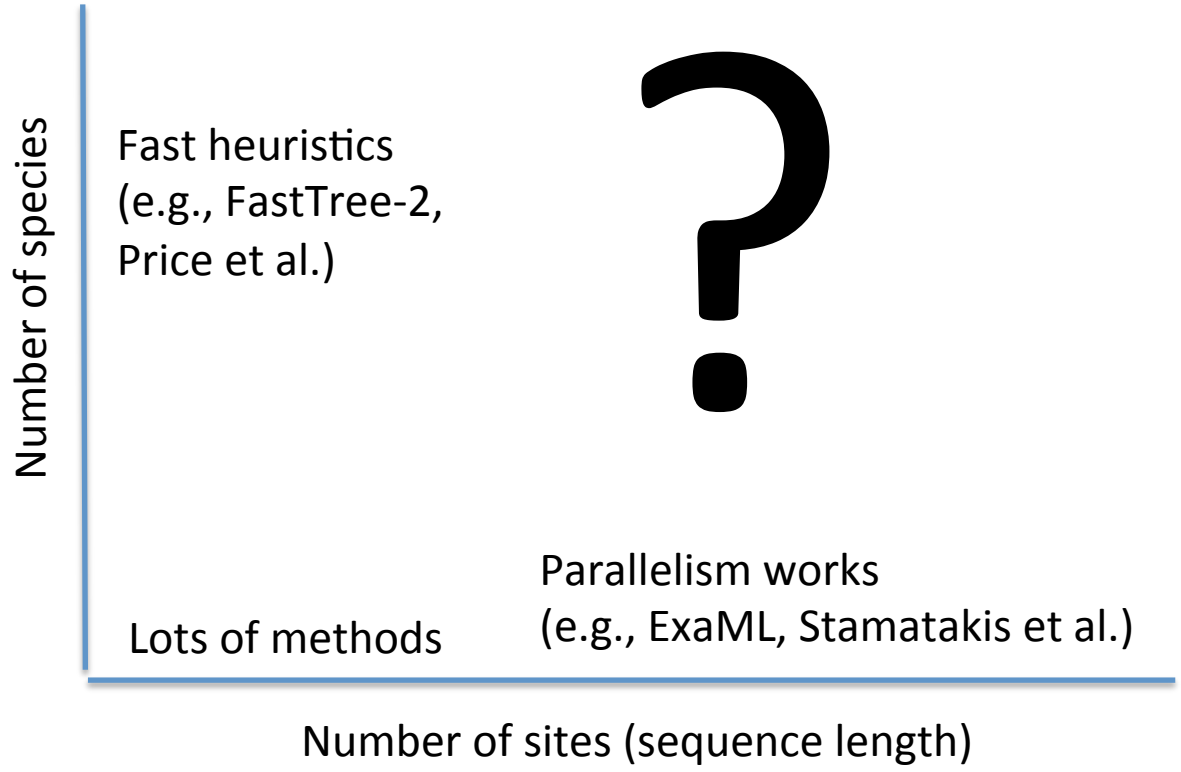


Maximum Likelihood (NP-hard):

Input is a multiple sequence alignment,

Output is a tree maximizing the probability of generating the input sequences.

Maximum Likelihood Heuristics



Maximum Likelihood (NP-hard):

Input is a multiple sequence alignment,

Output is a tree maximizing the probability of generating the input sequences.

Major Open Problems

- Scalable maximum likelihood tree estimation
  - Input: Multiple sequence alignment (and specified model)
  - Output: Tree and numeric parameters to maximize probability of the sequences under the model
  - Comments: RAxML cannot analyze large numbers of sequences efficiently; FastTree cannot analyze long alignments (and has poor parallelism)

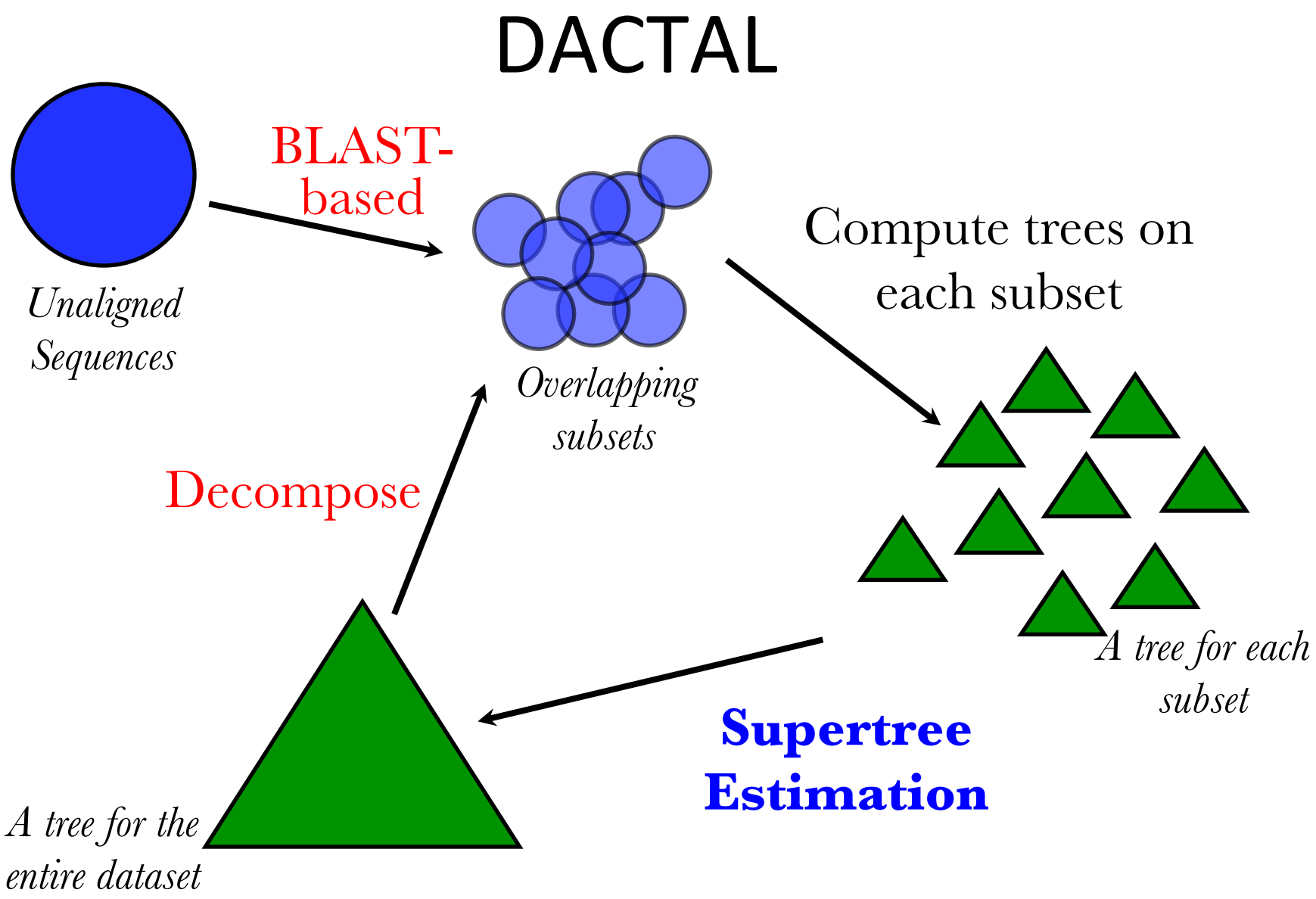
Second Major Open Problem

Supertree estimation

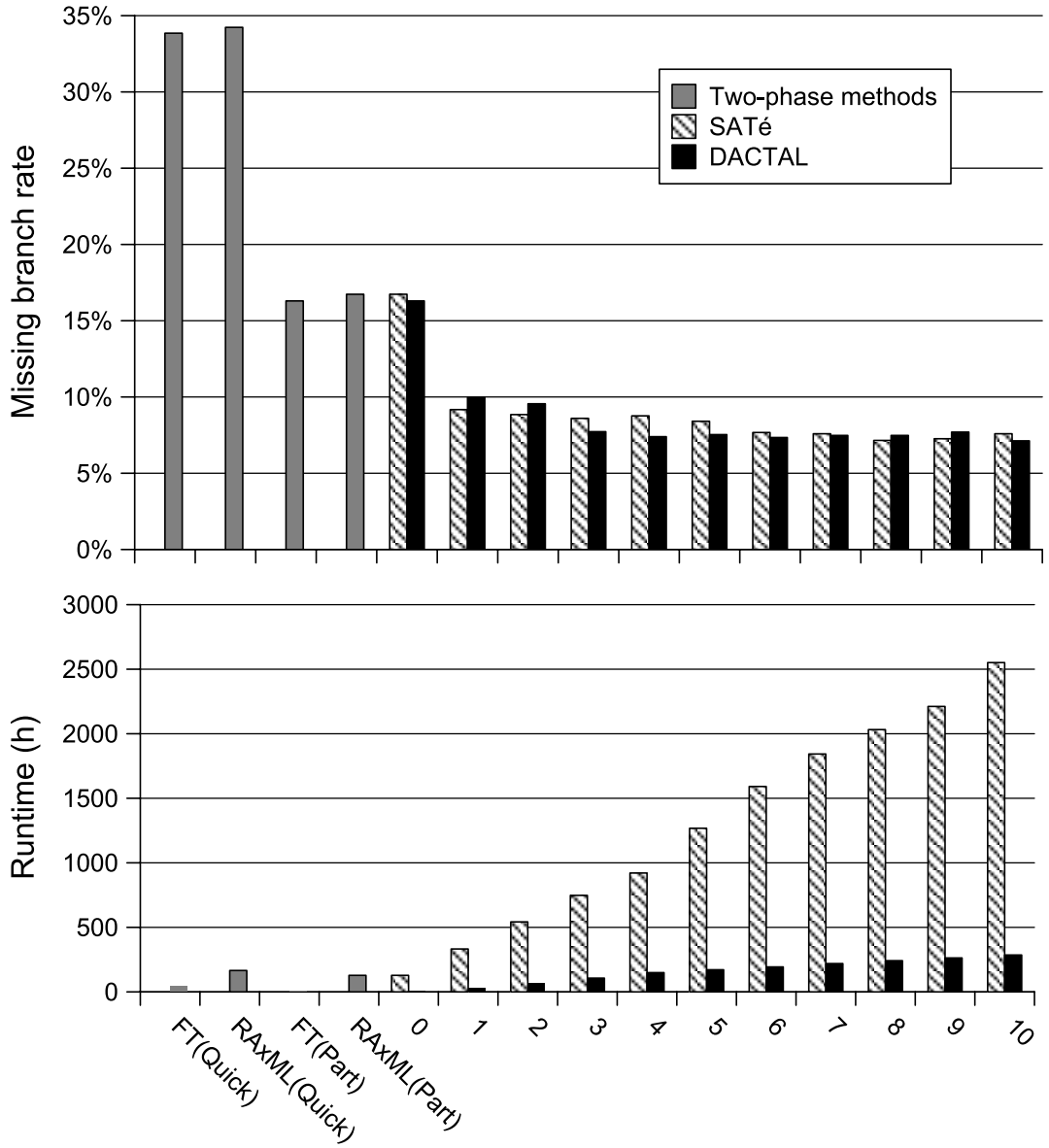
- Input: Set $\mathcal{T}$ of unrooted trees on subsets of S
- Output: Tree T that minimizes total distance to trees in $\mathcal{T}$

Comments:

- Basic problem in phylogenetics.
- Key algorithmic step in divide-and-conquer tree estimation methods.
- The best current methods rely on heuristics for NP-hard optimization problems, and so cannot scale to large datasets (for either dimension of large!).


DACTAL
BLAST-based
Unaligned Sequences
Overlapping subsets
Compute trees on each subset
A tree for each subset
Supertree Estimation
Decompose
A tree for the entire dataset

DACTAL compared to SATé and standard methods



16S.T dataset
7350 sequences
from the CRW
(Comparative Ribosomal
Database)

Observations

- Highly accurate statistical methods (especially maximum likelihood and Bayesian methods) have been developed for many problems.

Observations

- Highly accurate statistical methods (especially maximum likelihood and Bayesian methods) have been developed for many problems.
- However, these methods were typically designed for small datasets, and either do not run on large datasets, take too long, or have poor accuracy.

Observations

- Highly accurate statistical methods (especially maximum likelihood and Bayesian methods) have been developed for many problems.
- However, these methods were typically designed for small datasets, and either do not run on large datasets, take too long, or have poor accuracy.
- Relative performance of methods can change with dataset size and heterogeneity!

Observations

- Highly accurate statistical methods (especially maximum likelihood and Bayesian methods) have been developed for many problems.
- However, these methods were typically designed for small datasets, and either do not run on large datasets, take too long, or have poor accuracy.
- Relative performance of methods can change with dataset size and heterogeneity!
- But appropriate divide-and-conquer strategies can make them scale to large datasets, and be massively parallel.

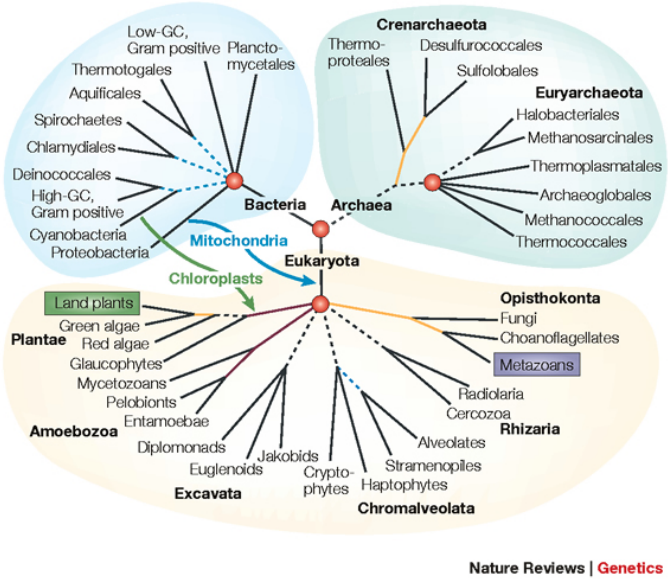
Summary

- Divide-and-conquer in phylogenetics and multiple sequence alignment is very powerful, and can lead to improved accuracy and scalability.
- The ingredients of these strategies are:
  - Existing tree-based approaches for the decomposition step
  - Existing statistical methods for analyzing subsets (often computationally intensive)
  - Combining solutions from subsets is where the research is needed!
- Distributed computing is necessary for some datasets

The Tree of Life: *Multiple Challenges*

Scientific challenges:

- Ultra-large multiple-sequence alignment
- Gene tree estimation
- Metagenomic classification
- Alignment-free phylogeny estimation
- Supertree estimation
- Estimating species trees from many gene trees
- Genome rearrangement phylogeny
- Reticulate evolution
- Visualization of large trees and alignments
- Data mining techniques to explore multiple optima
- Theoretical guarantees under Markov models of evolution



Techniques: applied probability theory, graph theory, supercomputing, and heuristics

Testing: simulations and real data

Acknowledgments





PASTA: Nam Nguyen (now postdoc at UIUC) and Siavash Mirarab (now faculty at UCSD), undergrad: Keerthana Kumar (at UT-Austin)

PASTA+BAli-Phy: Mike Nute (PhD student at UIUC)

DACTAL: Serita Nelesen (now professor at Calvin College)

Current NSF grants: ABI-1458652 (multiple sequence alignment)

Grainger Foundation (at UIUC), and UIUC

TACC, UTCS, Blue Waters, and UIUC campus cluster

PASTA is available on github at https://github.com/smirarab/pasta; see also PASTA+BAli-Phy at http://github.com/MGNute/pasta