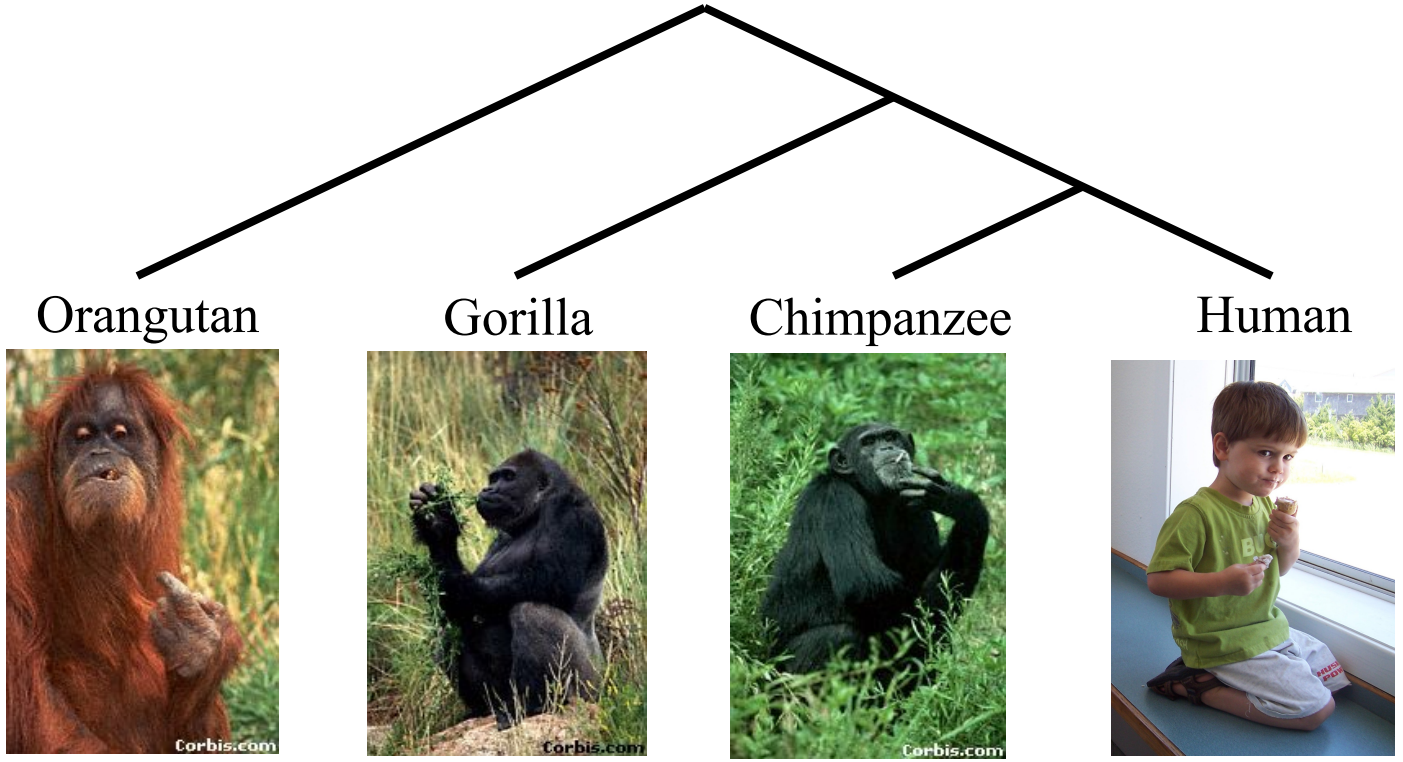
Advancing Genome-Scale Phylogenomic Analysis

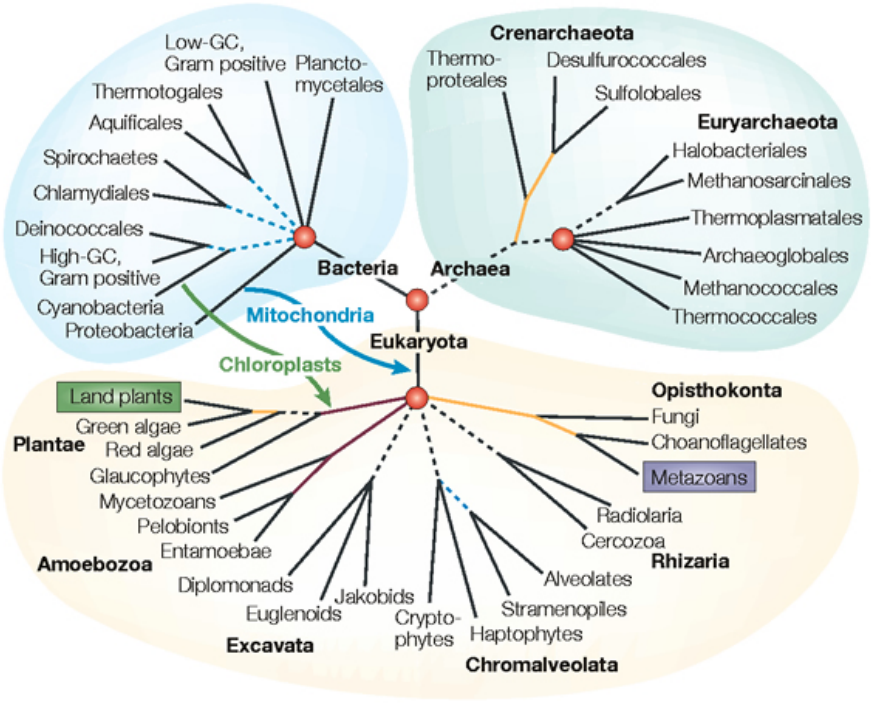
Tandy Warnow
Departments of Computer Science and Bioengineering
Carl R. Woese Institute for Genomic Biology
The University of Illinois at Urbana-Champaign
http://tandy.cs.illinois.edu

Species Tree



From the Tree of the Life Website,
University of Arizona

Phylogenies and Applications



Nature Reviews | Genetics

Basic Biology:

How did life evolve?

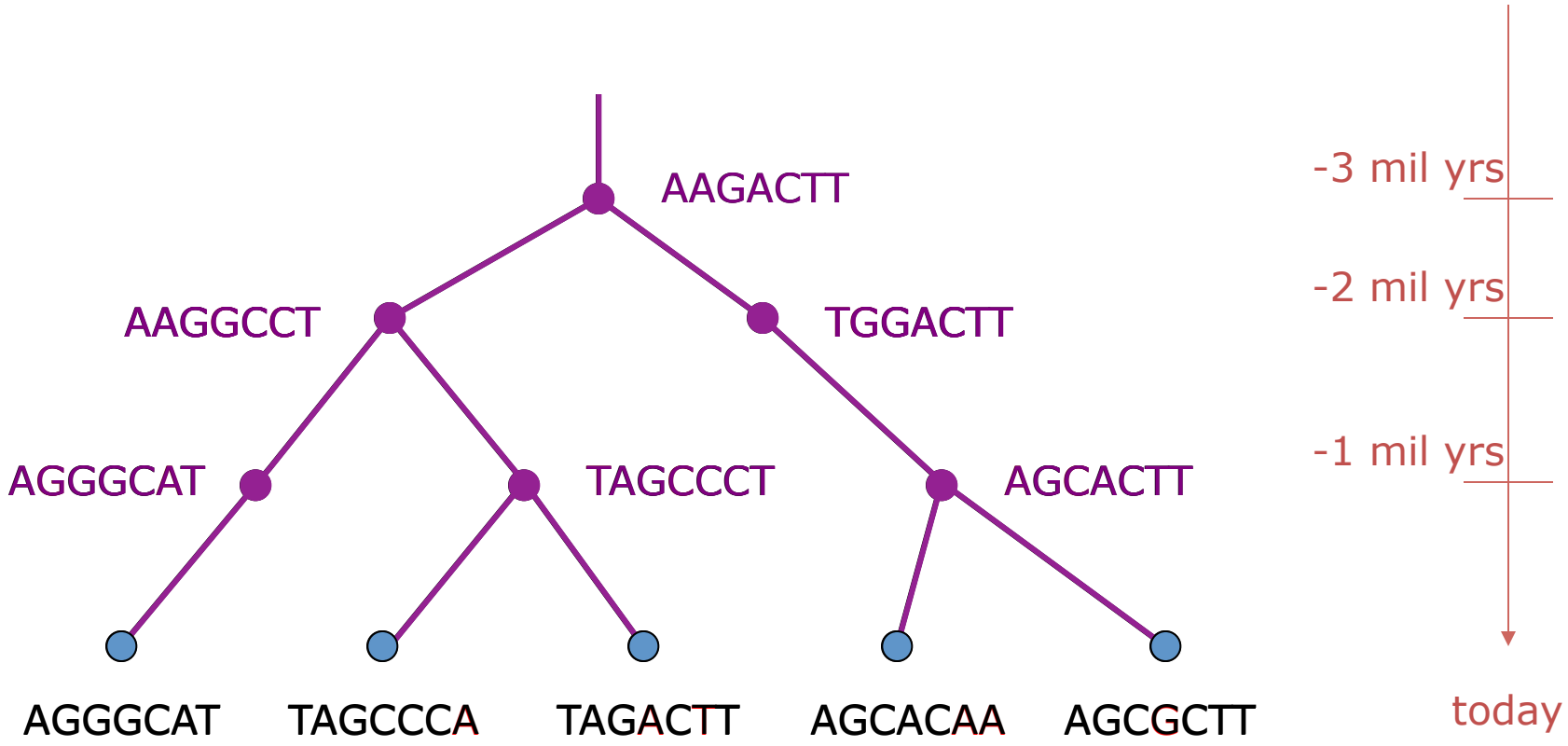
Applications of phylogenies to:

protein structure and function
population genetics
human migrations
metagenomics

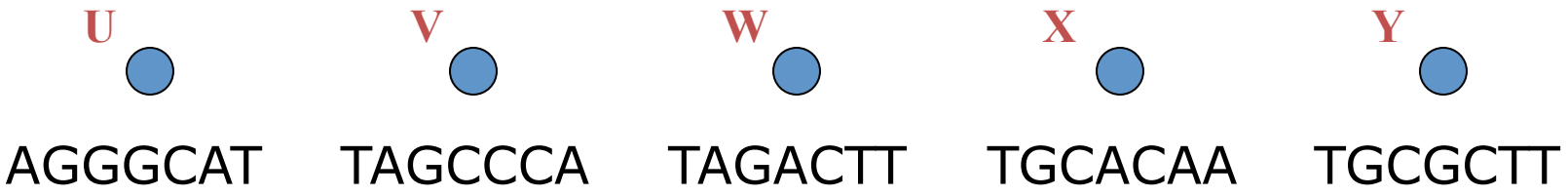
- "Nothing in biology makes sense except in the light of evolution"
  - Theodosius Dobzhansky, 1973 essay in the American Biology Teacher, vol. 35, pp 125-129

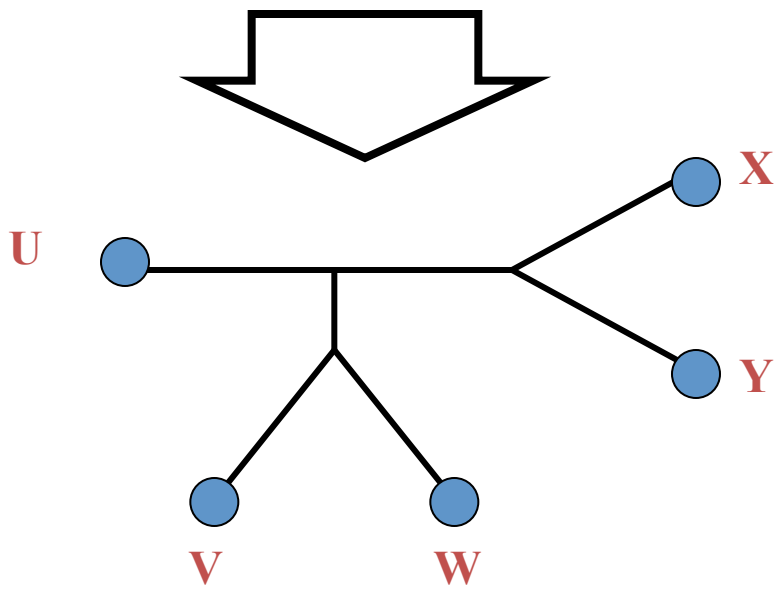
- *"…… nothing in evolution makes sense except in the light of phylogeny ..."*
  - *Society of Systematic Biologists, http://systbio.org/teachevolution.html*

DNA Sequence Evolution

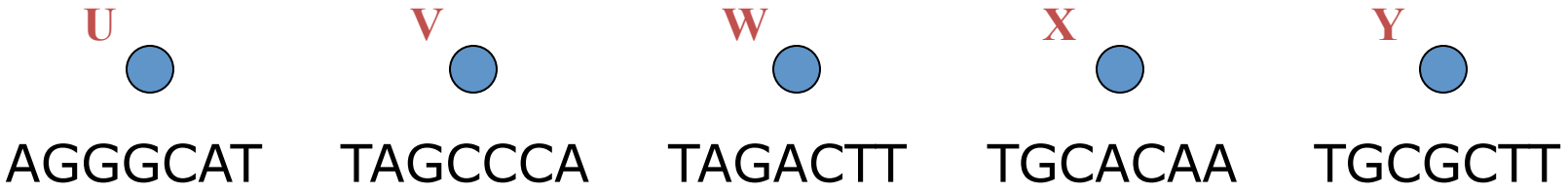


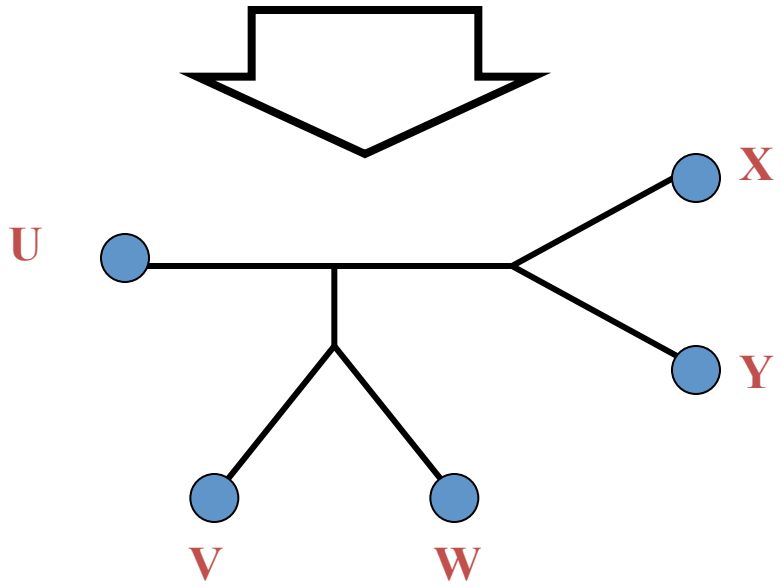
Phylogeny Estimation



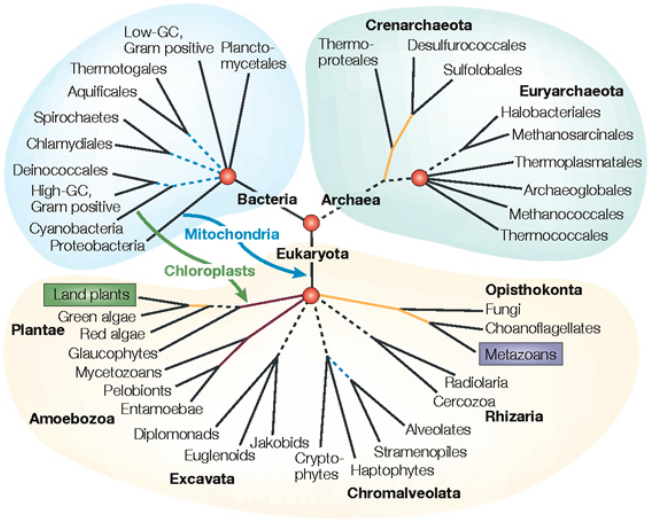


Maximum Likelihood Phylogeny Estimation: A Grand Challenge





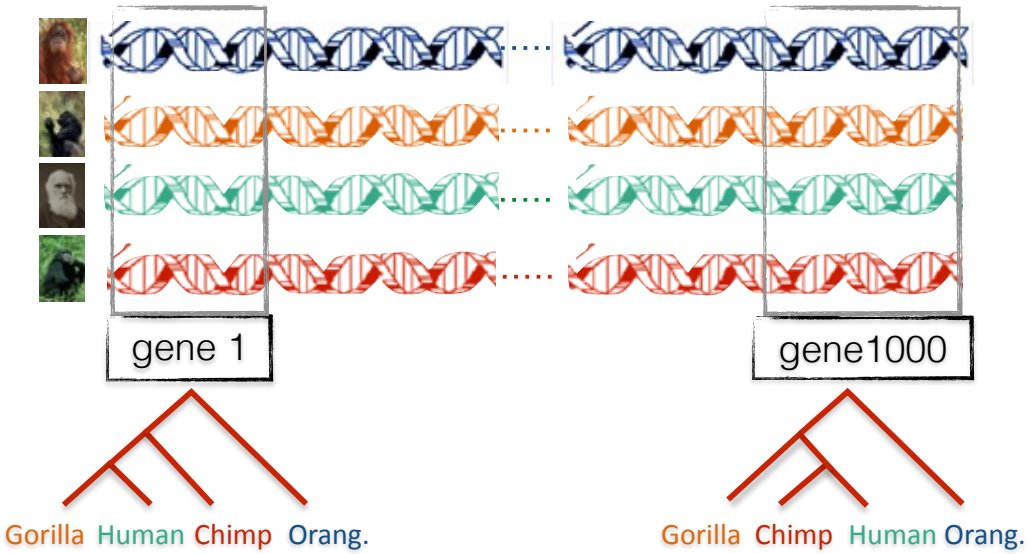
Phylogenomics



Nature Reviews | Genetics

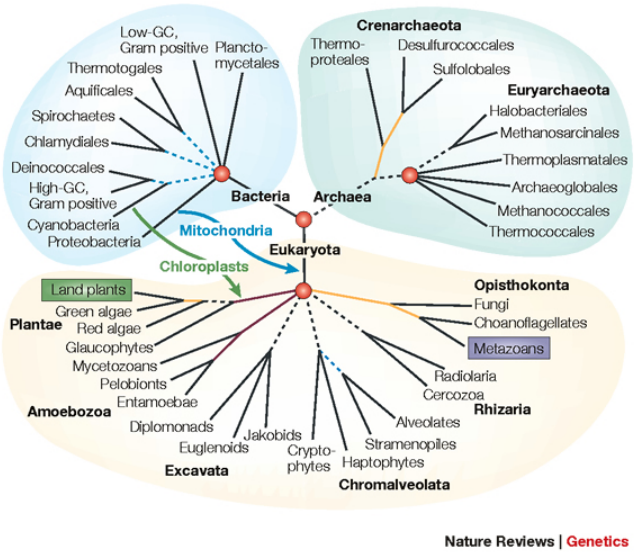
Phylogeny + genomics = genome-scale phylogeny estimation

Gene tree discordance



Incomplete Lineage Sorting (ILS) is a dominant cause of gene tree heterogeneity

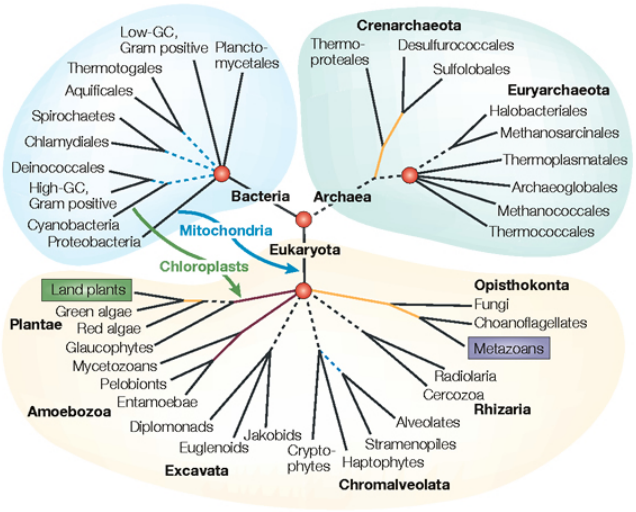
Phylogenomic Pipelines



Nature Reviews | Genetics

- Finding related genomic regions (homology detection)
- Multiple sequence alignment
- Maximum Likelihood phylogeny estimation for single genes
- Species tree estimation from multiple conflicting genes
- Answer biological questions

Computational Grand Challenges



Nature Reviews | Genetics

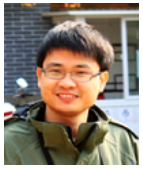
- NP-hard problems
- Exact solutions infeasible – heuristics necessary
- Parallelism helpful but does not address scalability with number of species
- Large datasets:
  - 100,000+ sequences
  - 10,000+ genes
- “BigData” complexity

Avian Phylogenomics Project

E Jarvis, HHMI

MTP Gilbert, Copenhagen

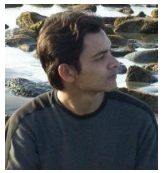
G Zhang, BGI



T. Warnow UT-Austin

S. Mirarab UT-Austin

Md. S. Bayzid, UT-Austin



Plus many many other people…

- Approx. 50 species, whole genomes, 14,000 loci
- Jarvis, Mirarab, et al., *Science* 2014

- Challenge #1: Maximum likelihood analysis of concatenated multiple sequence alignments **took 250 CPU years** using supercomputers around the globe.
- Challenge #2: Massive gene tree heterogeneity

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…

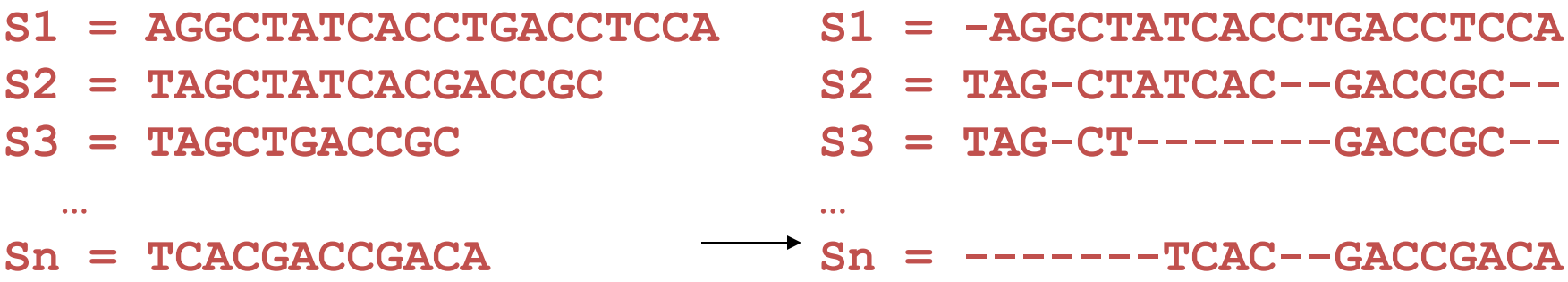
- PNAS 2014 (about 100 species and 800 genes)
- Second analysis underway, much larger dataset (~1200 species and ~1000 loci)

- Challenge #1: Construct multiple sequence alignment and tree for more than 100,000 sequences
- Challenge #2: Massive gene tree heterogeneity

Three of my favorite Grand Challenges

- Multiple Sequence Alignment: Methods for large-scale MSA (up to 1,000,000 sequences, including fragments)
- Phylogenomics: Methods for multi-locus species tree estimation that are robust to gene tree incongruence due to incomplete lineage sorting (ILS) and horizontal gene transfer (HGT)
- Supertree estimation: Methods that combine smaller trees into larger trees (useful for divide-and-conquer)

Multiple Sequence Alignment (MSA): *an important grand challenge*[1]



Novel techniques needed for scalability and accuracy

NP-hard problems and large datasets

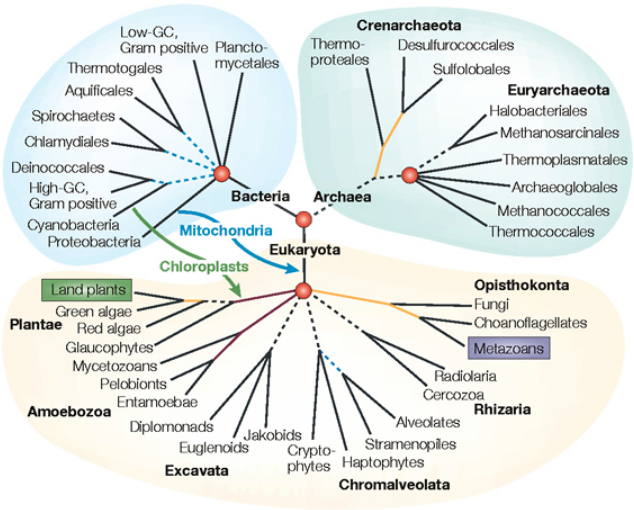
Current methods do not provide good accuracy

Few methods can analyze even moderately large datasets

Many important applications besides phylogenetic estimation

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

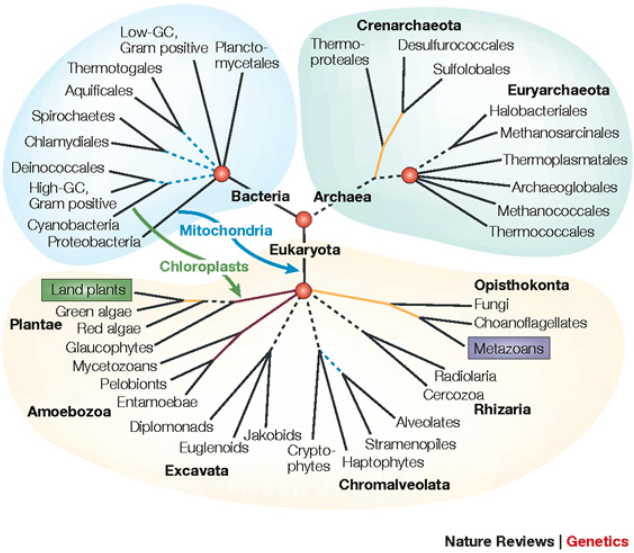
Consequences of MSA difficulties



Nature Reviews | Genetics

- Biologists restrict dataset size or use inadequate methods to be able to analyze their data.
- Alignment accuracy is reduced, which impacts downstream analyses:
  - Protein structure and function prediction
  - Metagenomic taxon identification
  - Phylogeny estimation
  - Detection of positive selection
- Scientific discoveries are jeapordized.

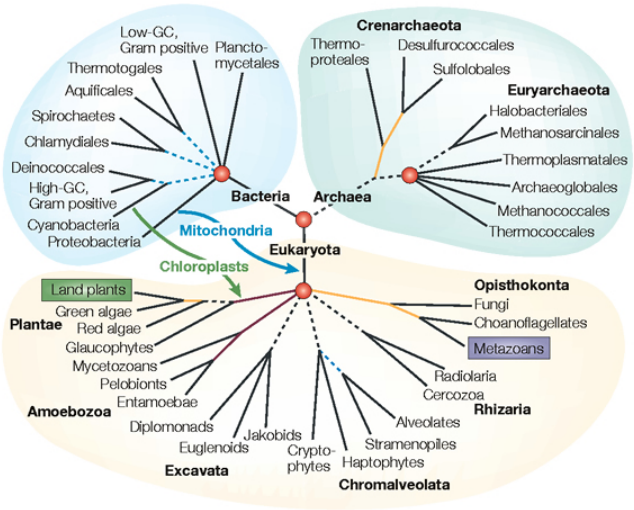
Today's talk: MSA estimation



Nature Reviews | Genetics

- PASTA: divide-and-conquer method to "boost" a MSA method
- Default PASTA (using MAFFT): can align 1,000,000 sequences with high accuracy and speed

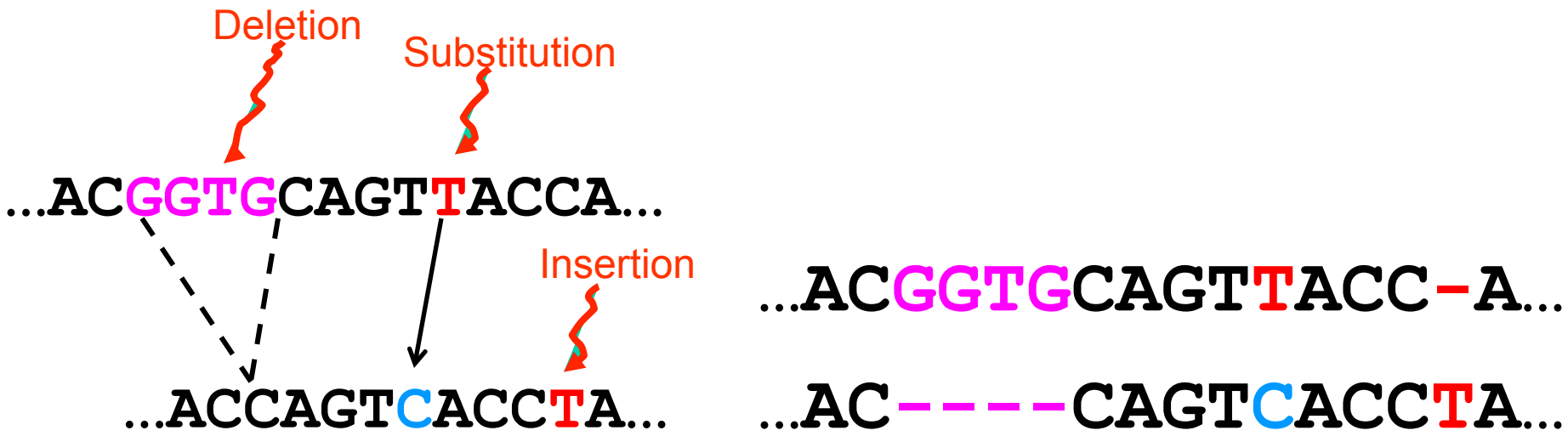
Today's talk: MSA estimation



Nature Reviews | Genetics

- PASTA: divide-and-conquer method to "boost" a MSA method
- Default PASTA (using MAFFT): can align 1,000,000 sequences with high accuracy and speed
- PASTA+BAli-Phy: integrating Bayesian statistical alignment method into PASTA, scaling to 10,000 sequences

BLUE WATERS
SUSTAINED PETASCALE COMPUTING



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
```



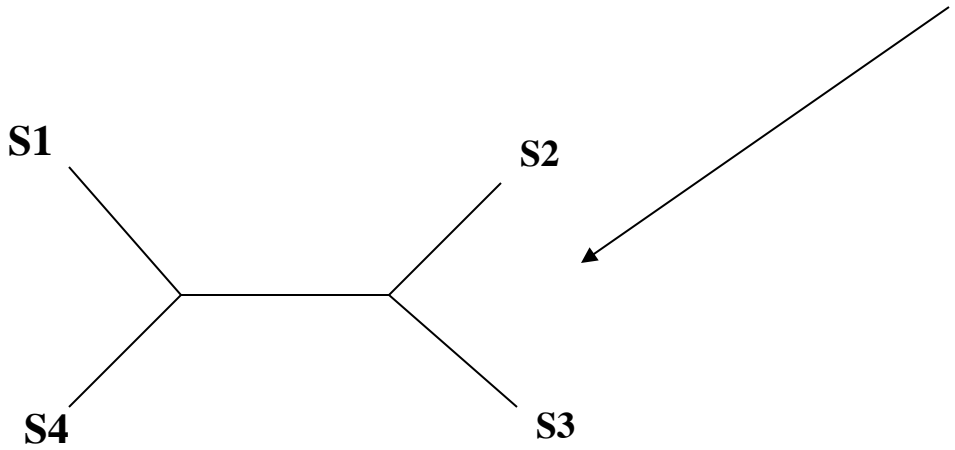
```
S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAC--GACCGC--
S3 = TAG-CT-------GACCGC--
S4 = -------TCAC--GACCGACA
```

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
```

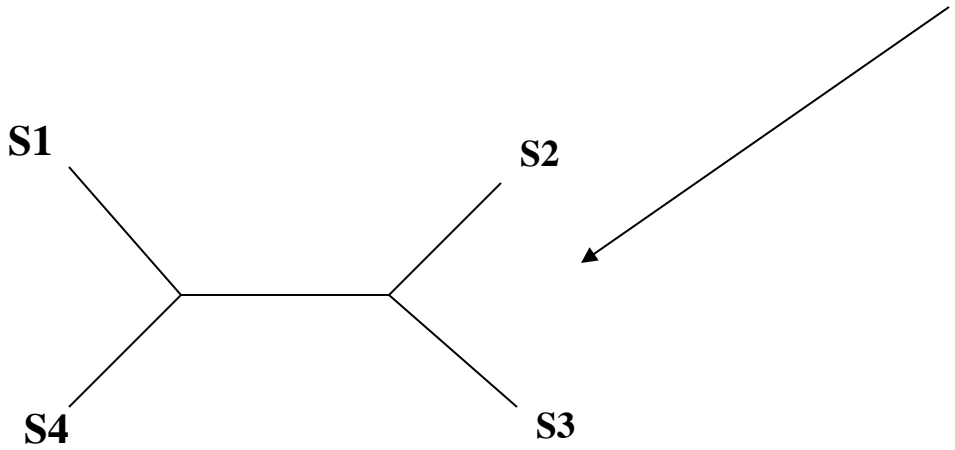


Statistical co-estimation would be much better!!!

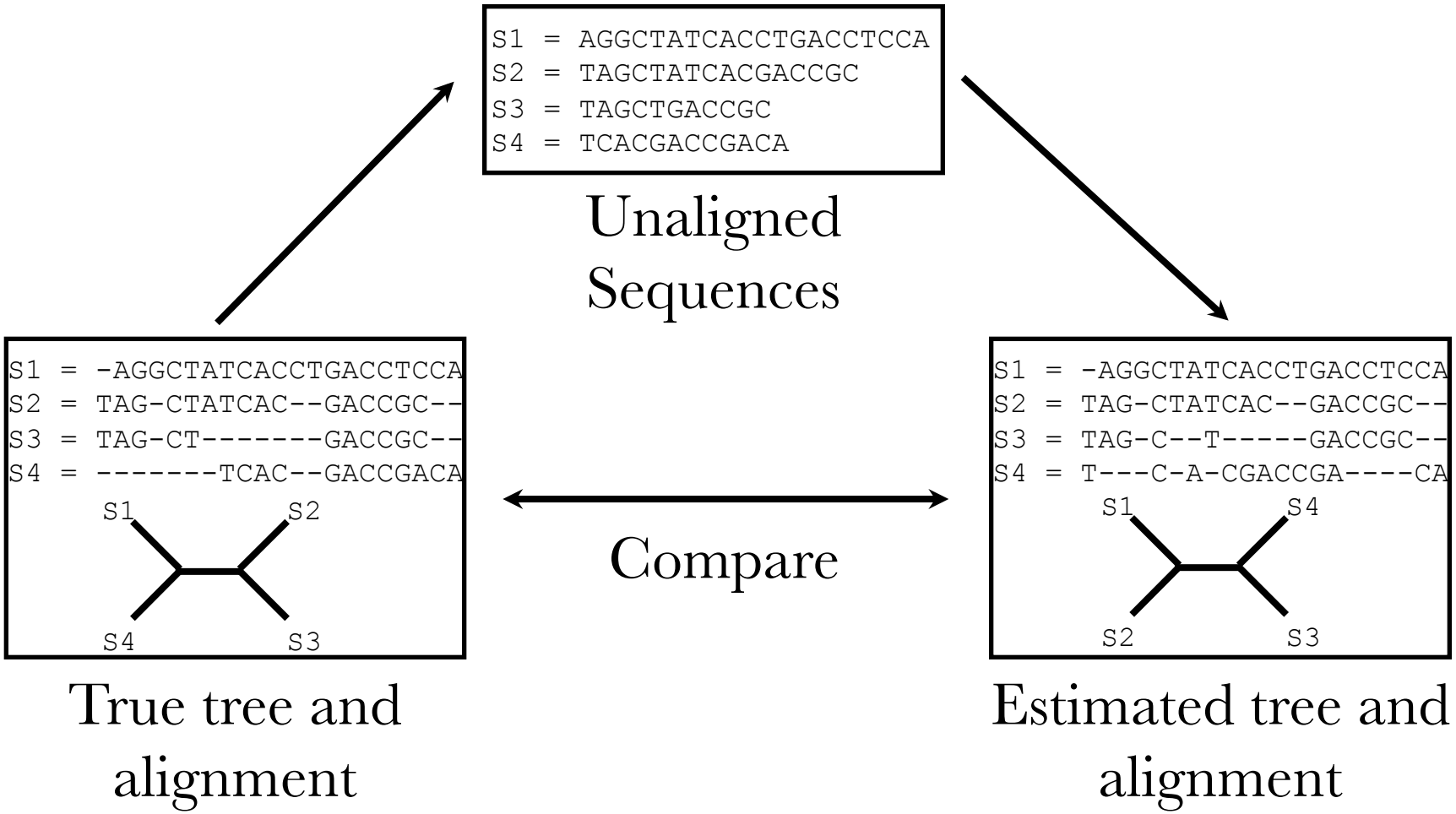
```
S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC
S3 = TAGCTGACCGC
S4 = TCACGACCGACA
```

→

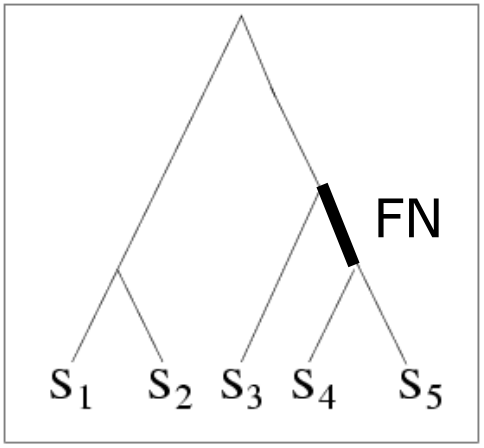
```
S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAC--GACCGC--
S3 = TAG-CT-------GACCGC--
S4 = -------TCAC--GACCGACA
```



Simulation Studies



Quantifying Error



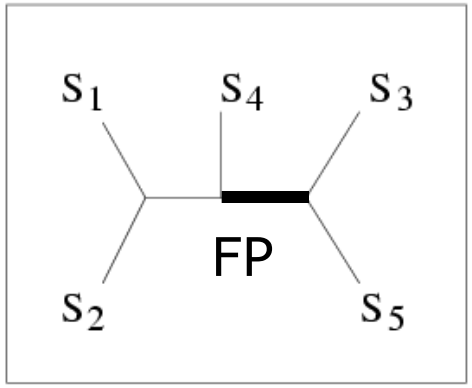
TRUE TREE

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



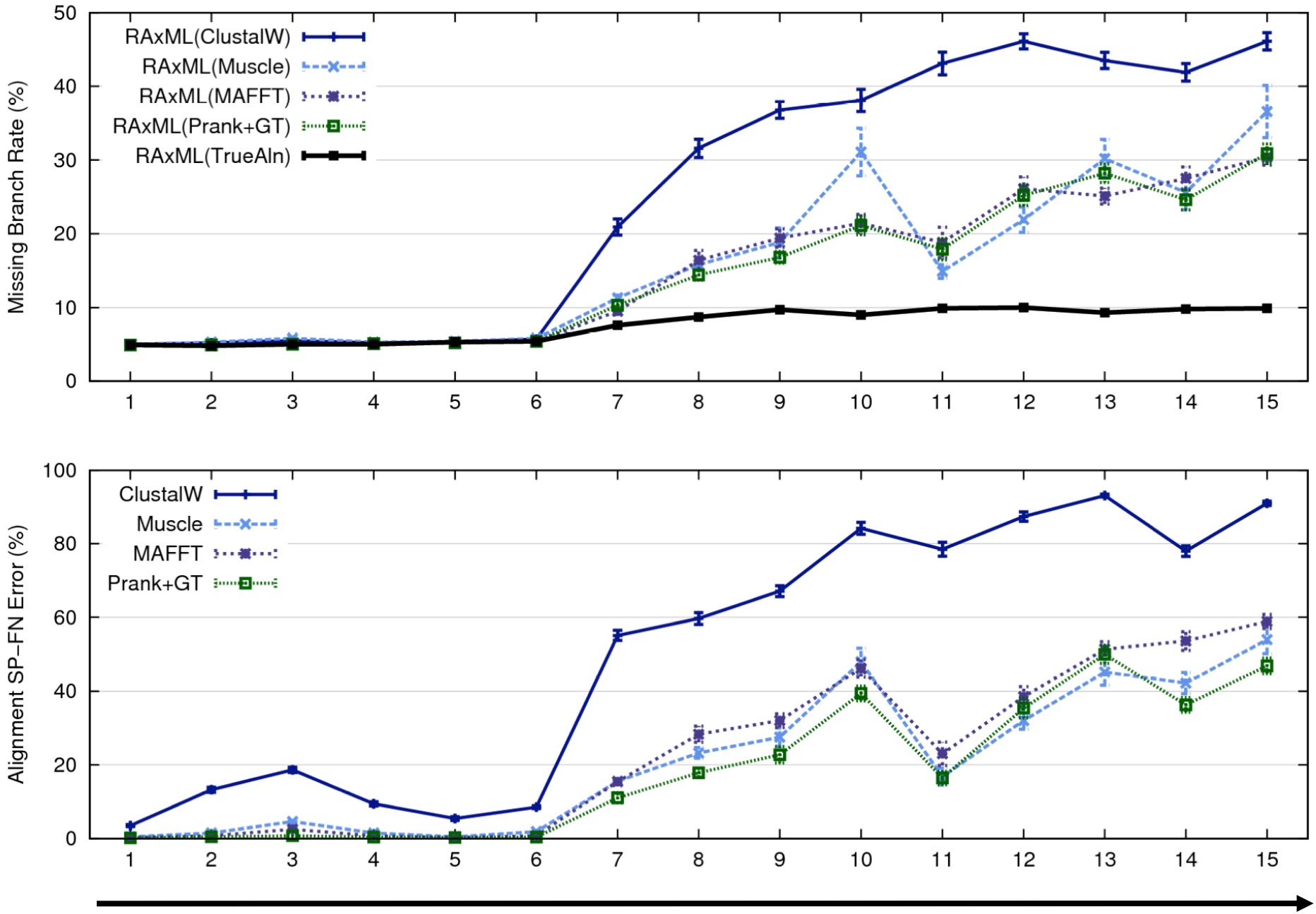
INFERRED TREE

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*

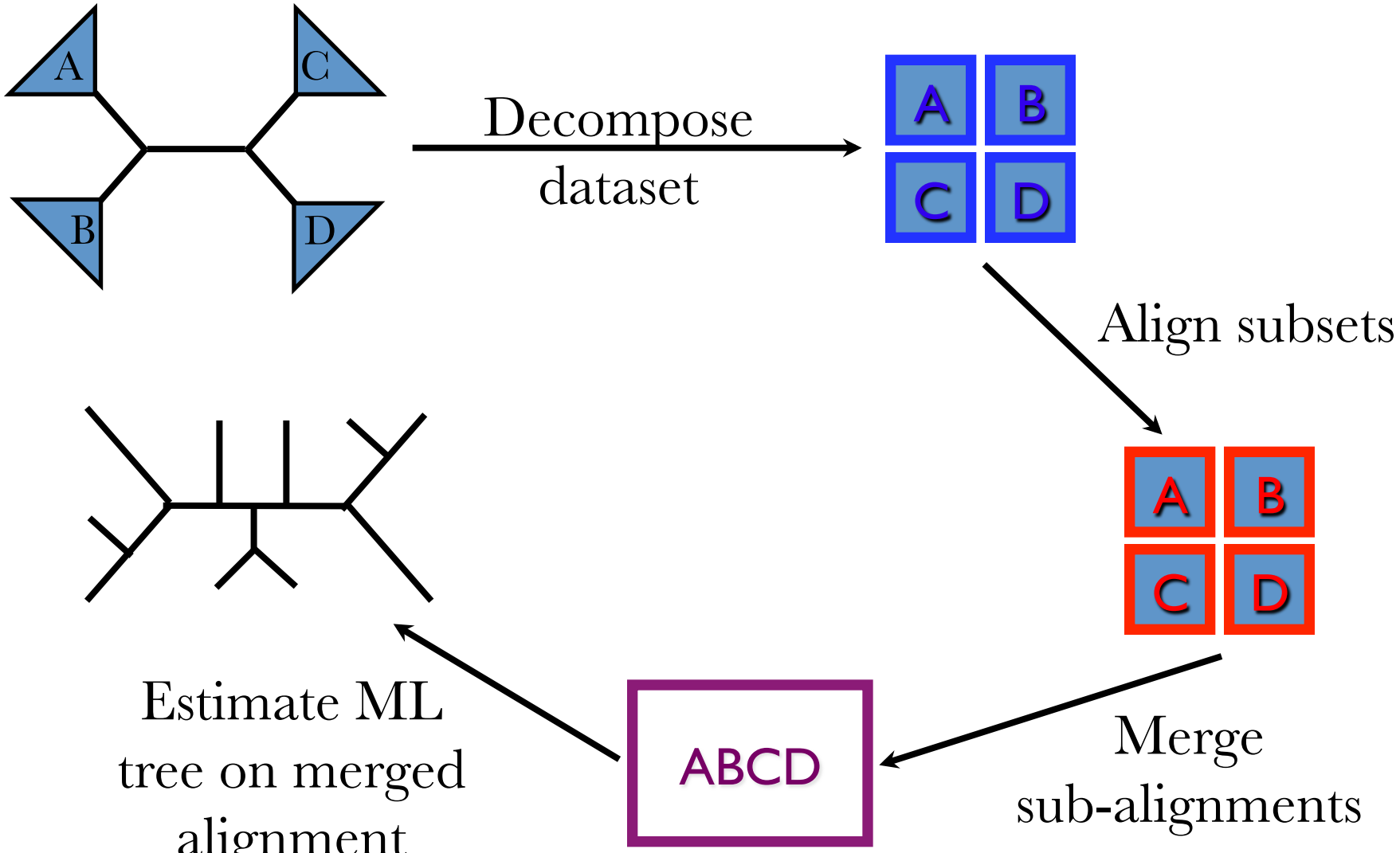

Missing Branch Rate (%)
RAxML(ClustalW)
RAxML(Muscle)
RAxML(MAFFT)
RAxML(Prank+GT)
RAxML(TrueAln)
Alignment SP–FN Error (%)
ClustalW
Muscle
MAFFT
Prank+GT
1000-taxon models, ordered by difficulty (Liu et al., 2009)

Key Observations

- Datasets that are large and have high rates of evolution are difficult to align accurately.
- However, datasets with slow rates of evolution can be aligned with high accuracy.
- Not all MSA methods can run on large datasets (and some cannot even run on moderate-sized datasets).

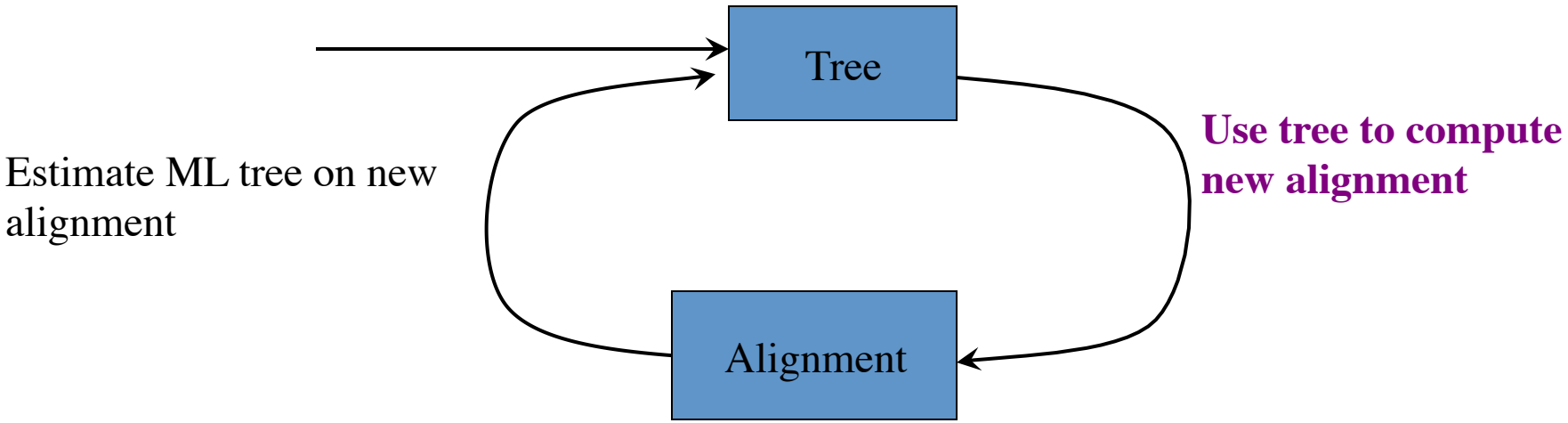
These observations suggest *divide-and-conquer* to boost MSA methods to larger datasets.

Re-aligning on a tree (boosting an MSA method)



SATé and PASTA Algorithms

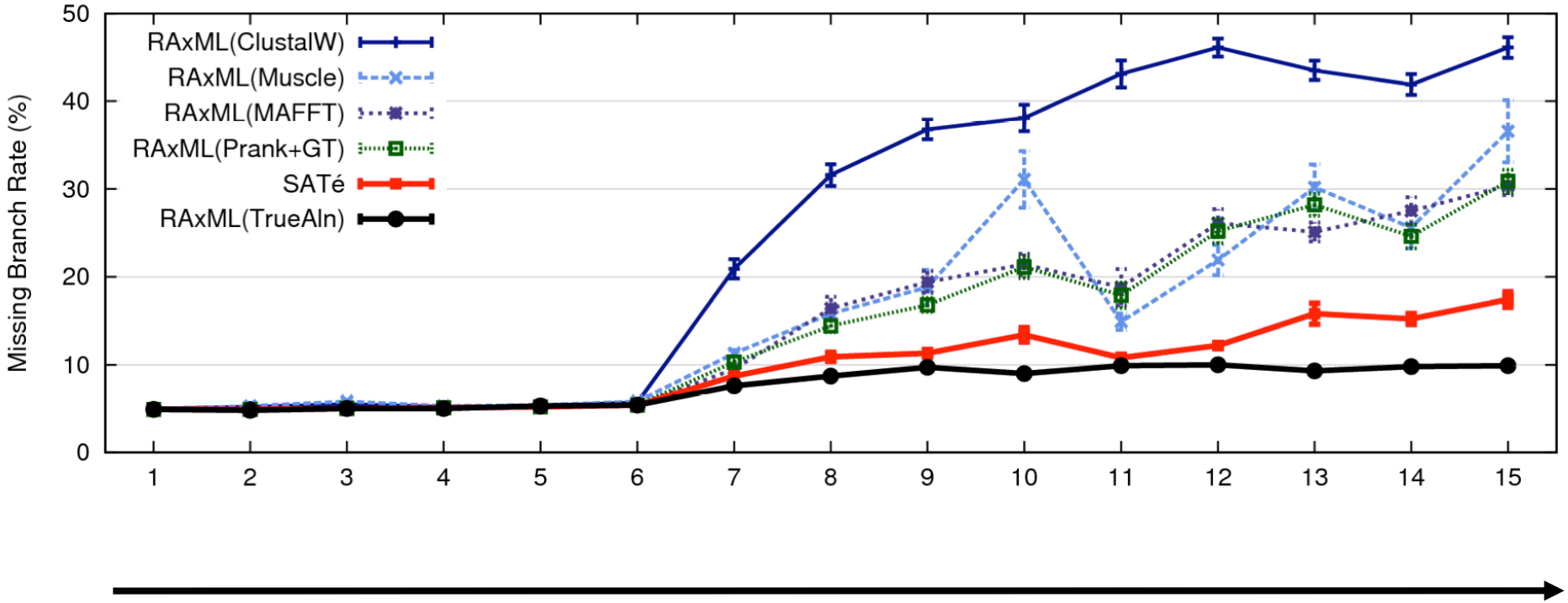
Obtain initial alignment and estimated ML tree



Repeat until termination condition, and

return the alignment/tree pair with the best ML score

SATé-1 (Science 2009) performance



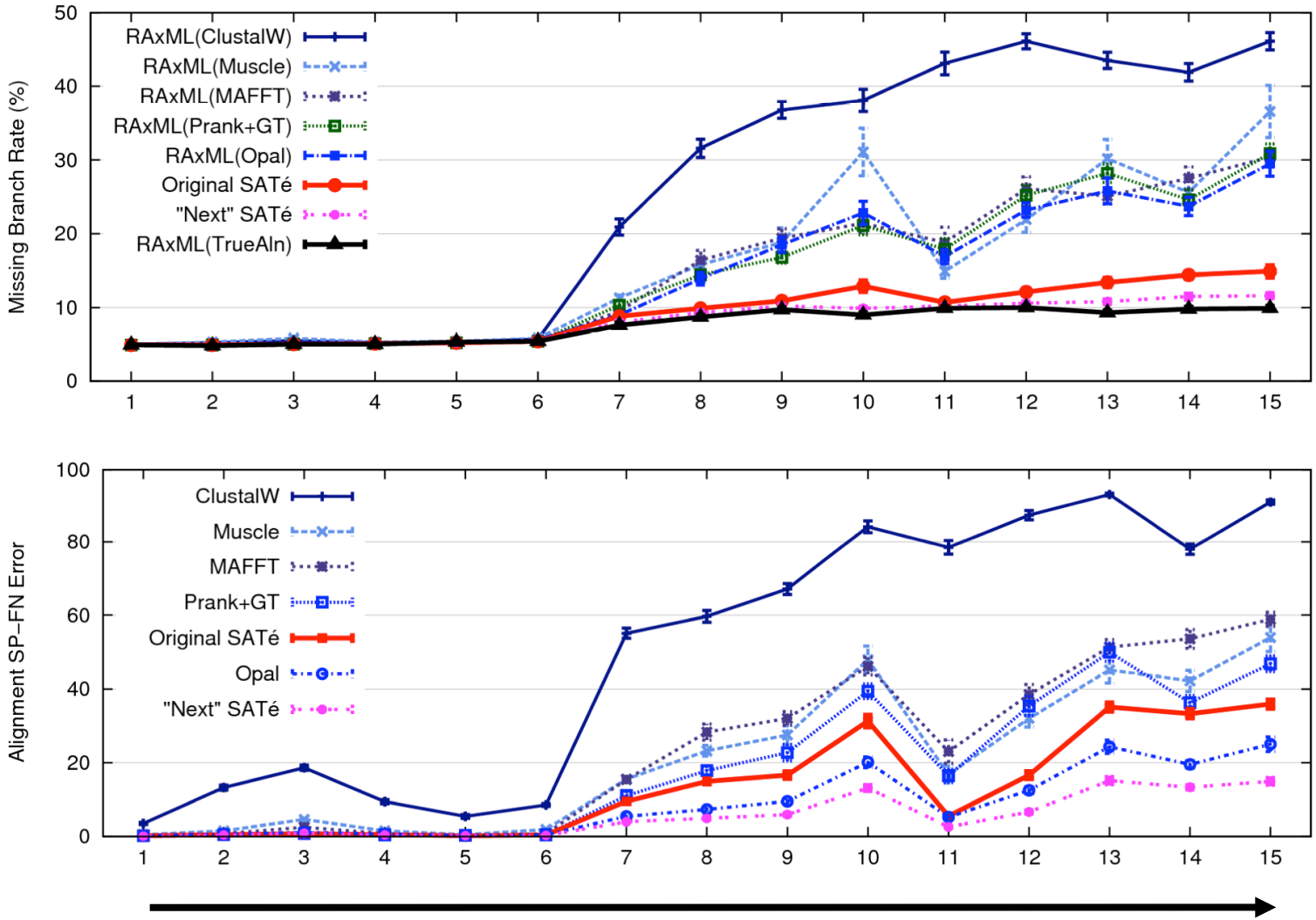
1000-taxon models, ordered by difficulty – rate of evolution generally increases from left to right

SATé-1 24 hour analysis, on desktop machines

(Similar improvements for biological datasets)

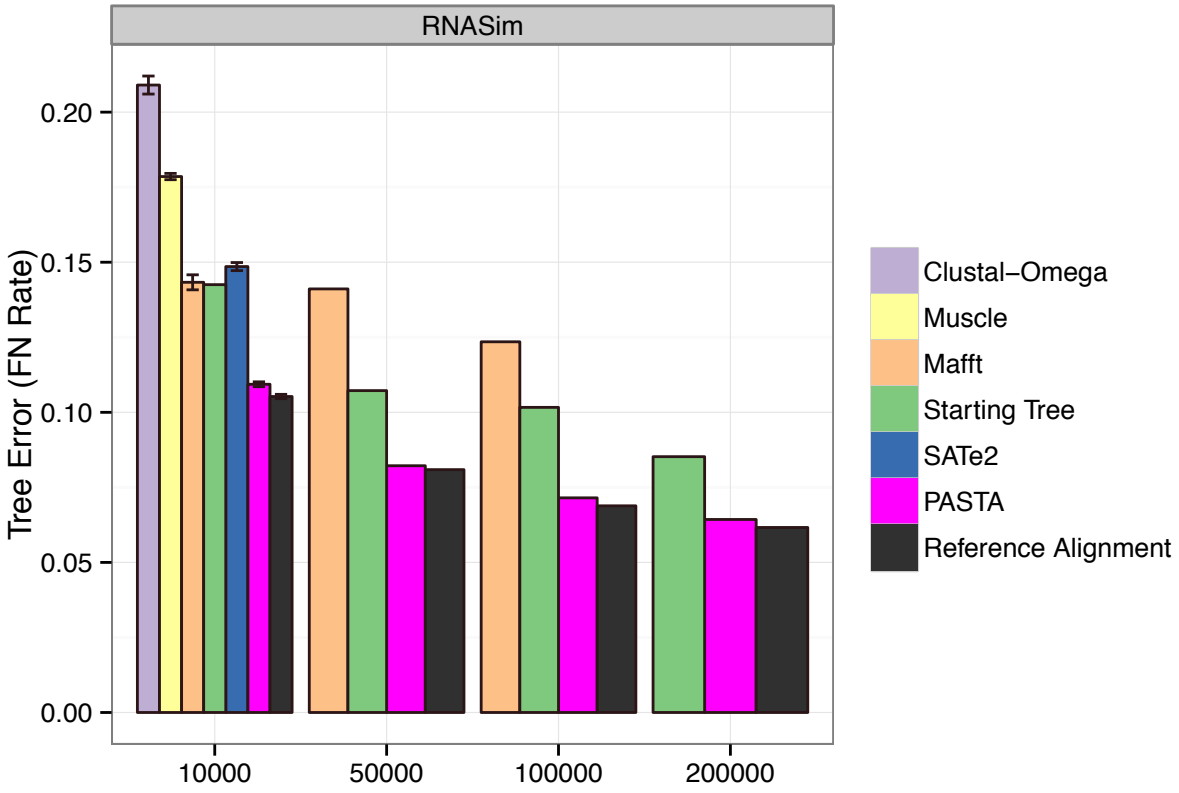
SATé-1 can analyze up to about 8,000 sequences.

SATé-1 vs. SATé-2 (Systematic Biology, 2012)



SATé-1: up to 8K
SATé-2: up to ~50K

PASTA: even better than SATé-2



- 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)

UPP

UPP = “Ultra-large multiple sequence alignment using Phylogeny-aware Profiles”

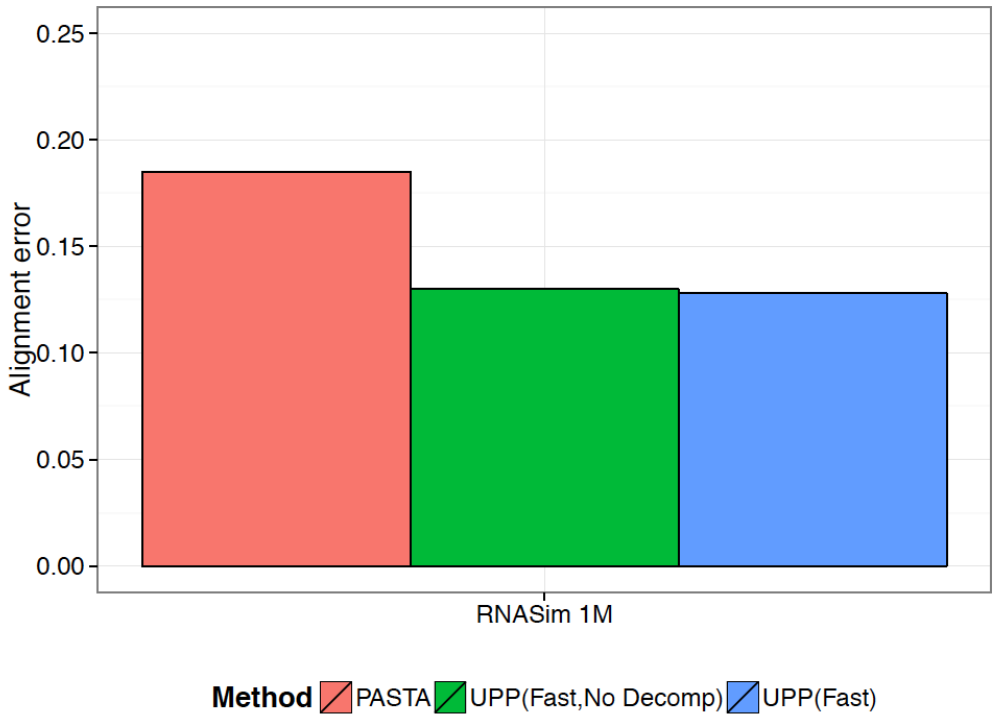
Nguyen, Mirarab, and Warnow. Genome Biology, 2014.

Purpose: highly accurate large-scale multiple sequence alignments, even in the presence of fragmentary sequences.

UPP Algorithmic Approach

1. Select small random subset of full-length sequences, and build “backbone alignment”
2. Construct an “Ensemble of Hidden Markov Models” on the backbone alignment
3. Add all remaining sequences to the backbone alignment using the Ensemble of HMMs

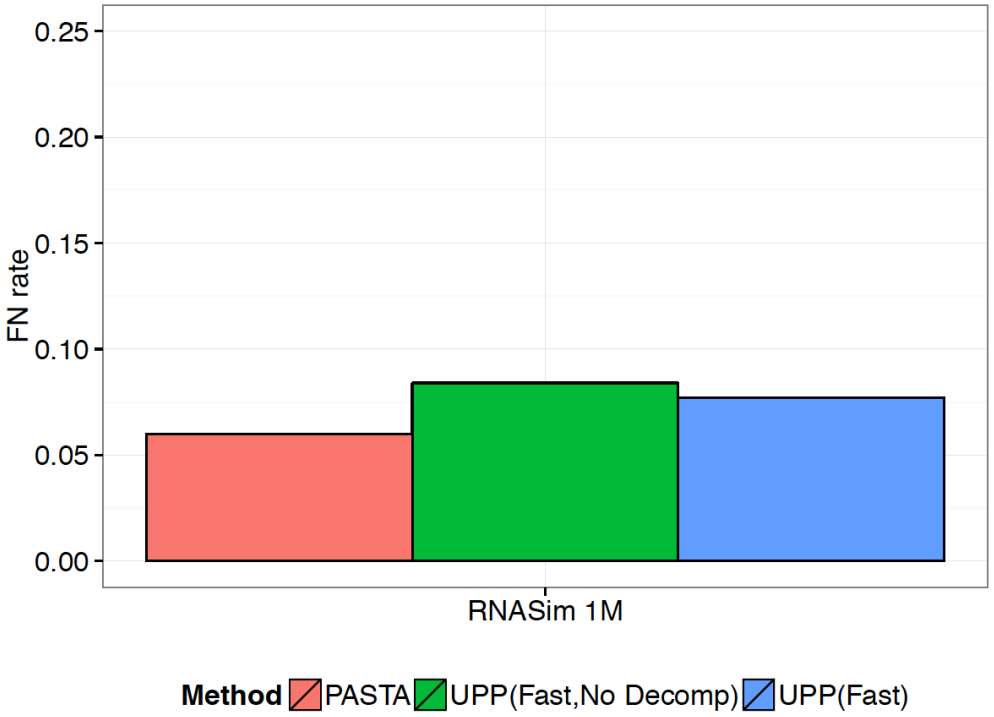
RNASim Million Sequences: alignment error



Notes:

- We show alignment error using average of SP-FN and SP-FP.
- UPP variants have better alignment scores than PASTA.
- (Not shown: Total Column Scores – PASTA more accurate than UPP)
- No other methods tested could complete on these data

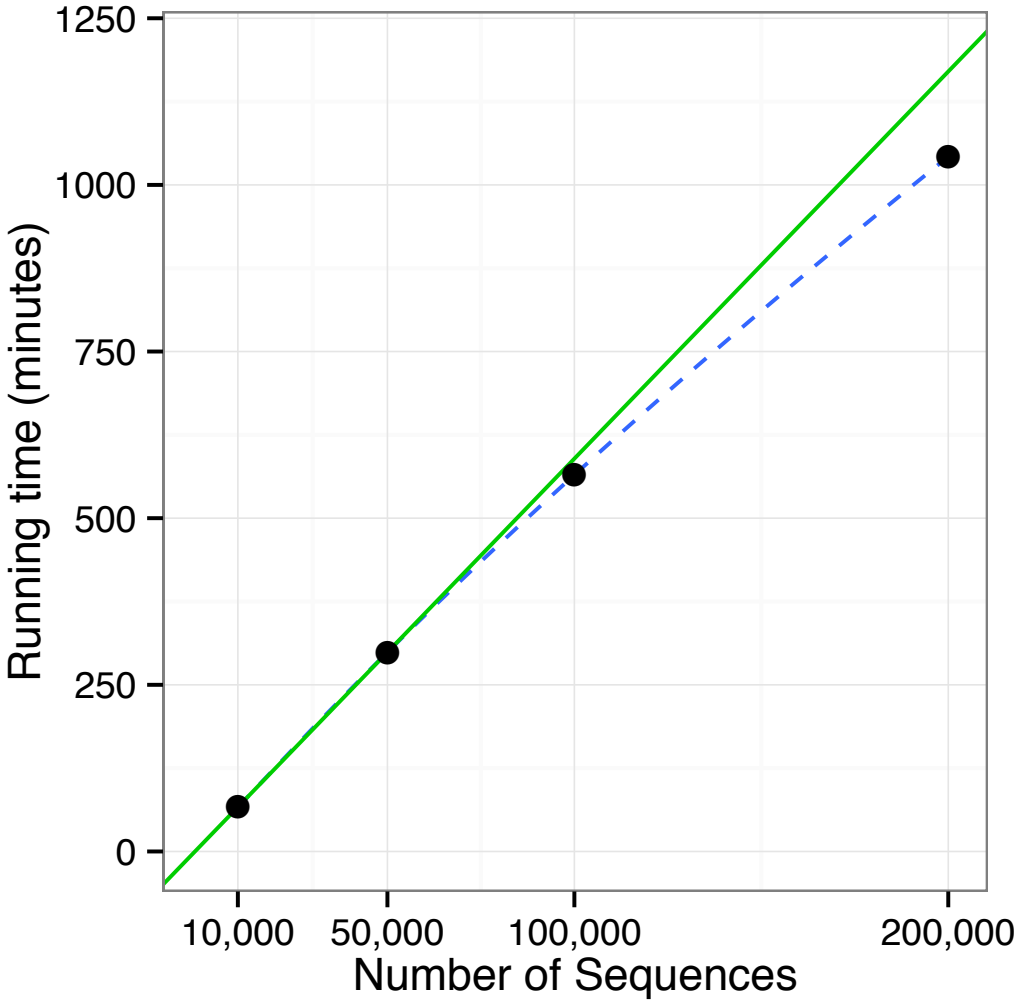
RNASim Million Sequences: tree error



Using 12 TACC processors:

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

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

PASTA and UPP: boosters of MSA methods

- PASTA
  - Combines iteration and divide-and-conquer to "boost" a preferred MSA method to large datasets; we showed results based on MAFFT
- UPP
  - Step 1: Constructs a "backbone" tree and an alignment on a small random subset of the sequences
  - Step 2: Aligns all the remaining sequences to the backbone alignment
  - We showed results where default PASTA computed the backbone alignment and tree.

PASTA and UPP: boosters of MSA methods

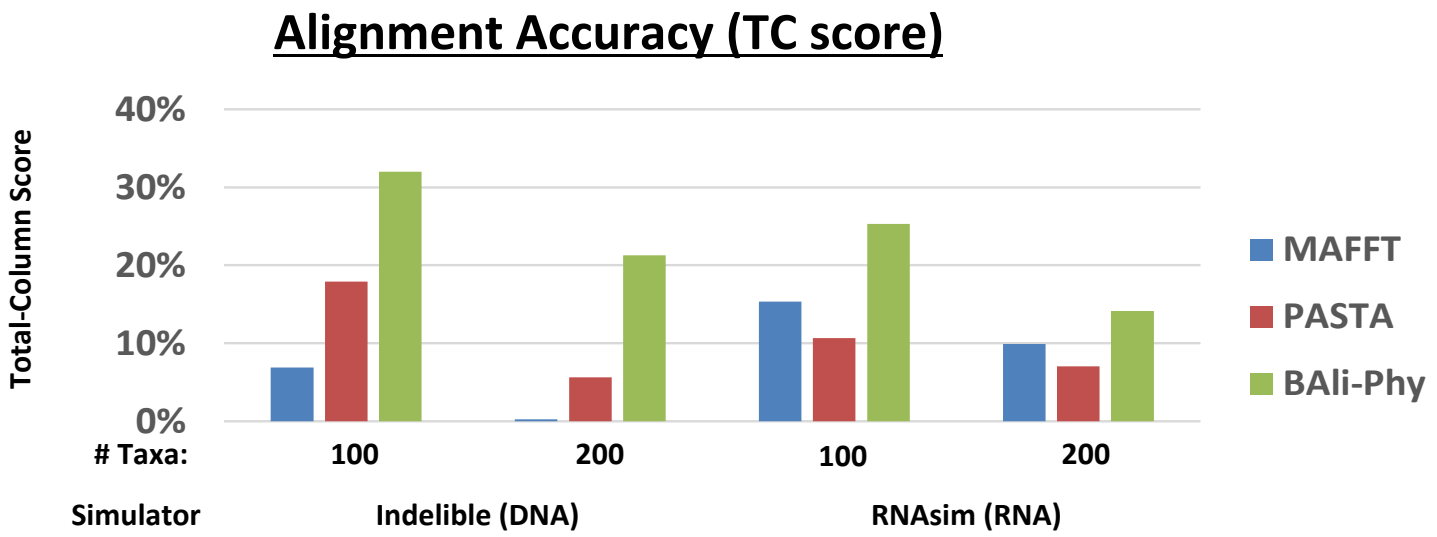
- PASTA
  - Combines iteration and divide-and-conquer to "boost" a preferred MSA method to large datasets; we showed results based on MAFFT
- UPP
  - Step 1: Constructs a "backbone" tree and an alignment on a small random subset of the sequences
  - Step 2: Aligns all the remaining sequences to the backbone alignment
  - We showed results where default PASTA computed the backbone alignment and tree.

Challenge: Can we boost statistical alignment methods?

BAli-Phy: leading statistical co-estimation method

- BAli-Phy (Redelings and Suchard, 2005):
  - Statistical co-estimation of the sequence alignment and the tree, using Bayesian MCMC.
  - Output can be a multiple sequence alignment, a phylogeny, or both, and can give estimate of uncertainty in each one.

BAli-Phy: better than PASTA



Each dataset has 100 or 200 sequences. To run BAli-Phy on a single dataset, we used 32 Blue Waters processors and ran BAli-Phy on each processor for 24 hours. We then collected all the samples from all the processors, and computed the posterior decoding (PD) alignment.

**Averages over 10 replicates*

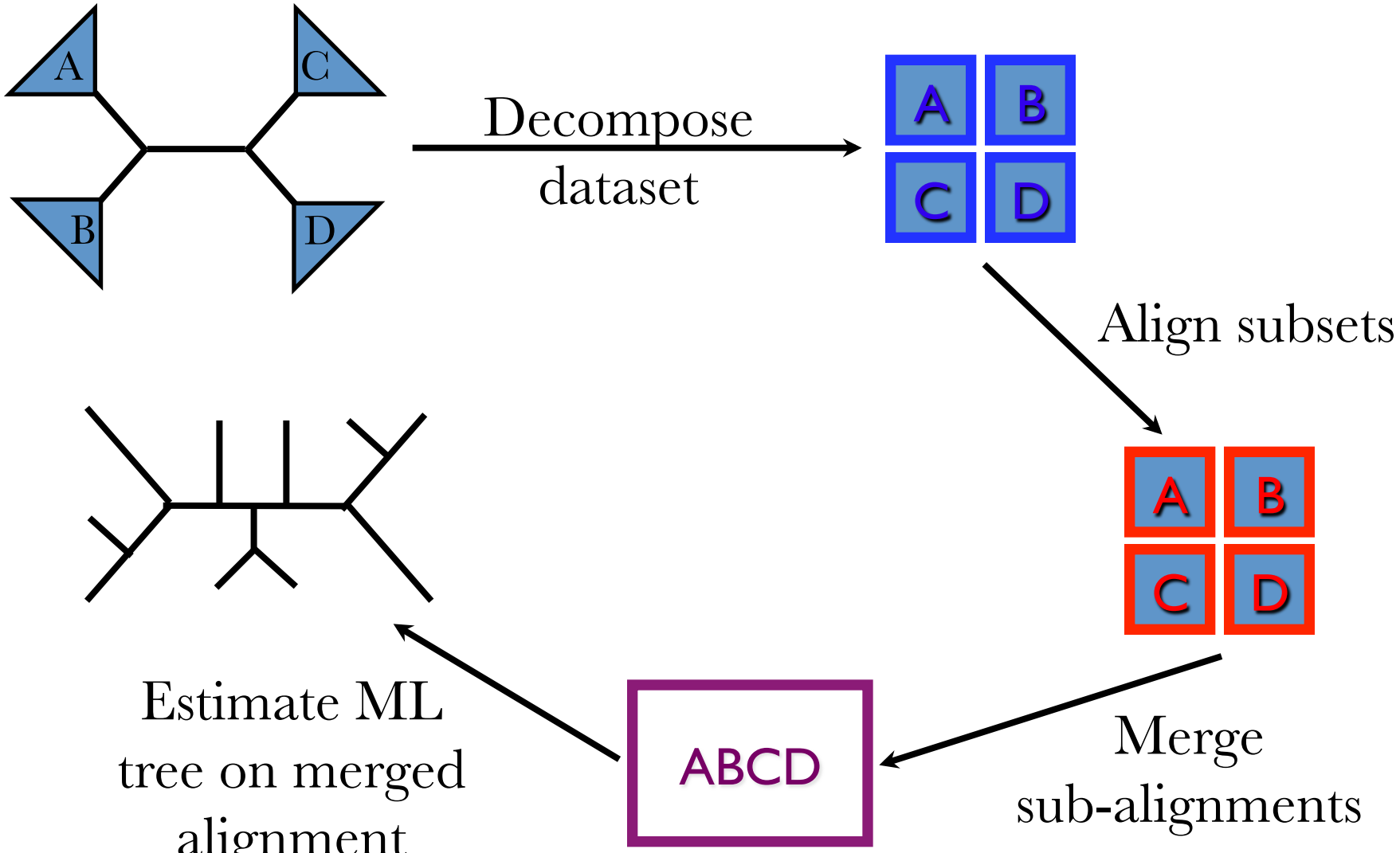
But: BAli-Phy is limited to small datasets

- BAli-Phy is computationally intensive:
  - 63 sequence dataset (Gaya et al., 2011) took 3 weeks
  - Largest dataset analyzed had 117 sequences (McKenzie et al., 2014)

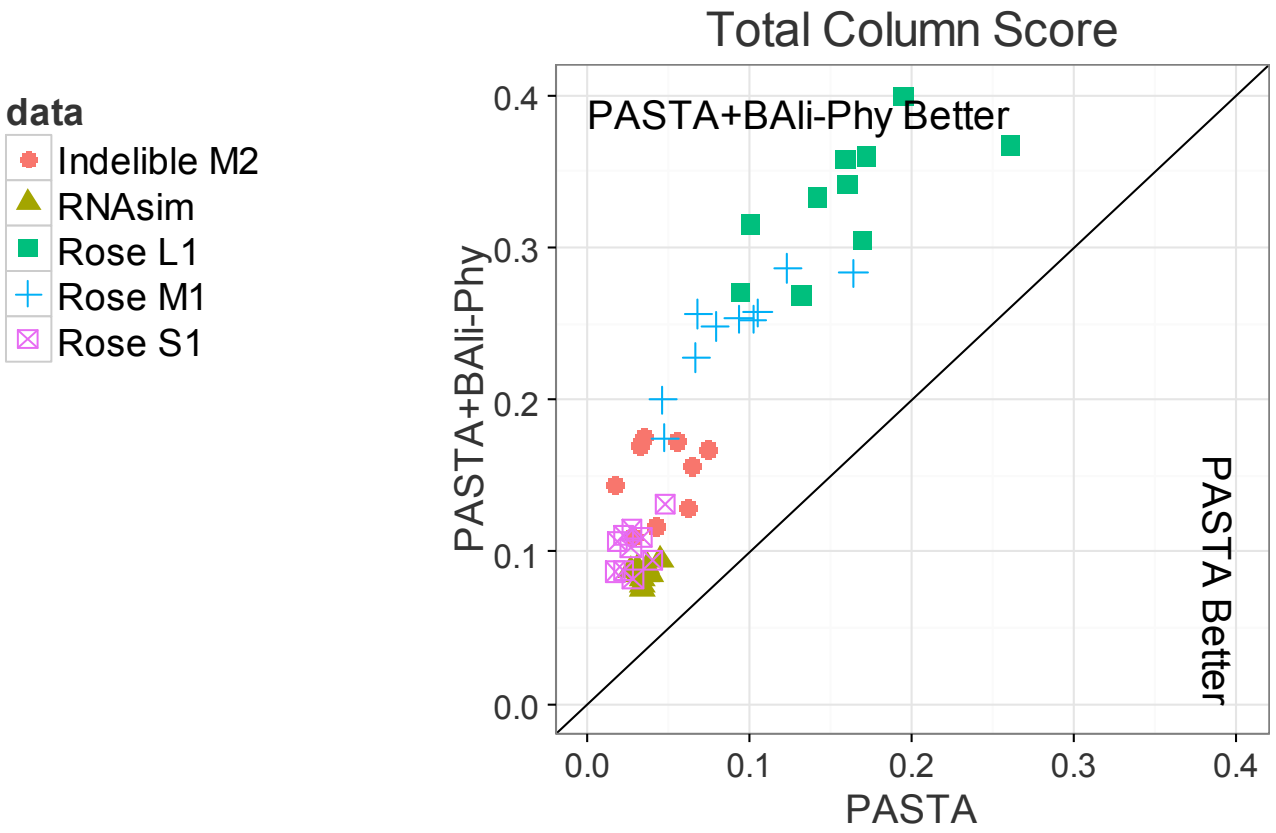
- BAli-Phy is not scalable:
  - Our study shows it breaks somewhere before 500 sequences (numerical issues possibly)

- From www.bali-phy.org/README.html
  - **5.2.1. Too many taxa?**
  - "BAli-Phy is quite CPU intensive, and so we recommend using 50 or fewer taxa in order to limit the time required to accumulate enough MCMC samples. (Despite this recommendation, data sets with more than 100 taxa have occasionally been known to converge.) We recommend initially pruning as many taxa as possible from your data set, then adding some back if the MCMC is not too slow."

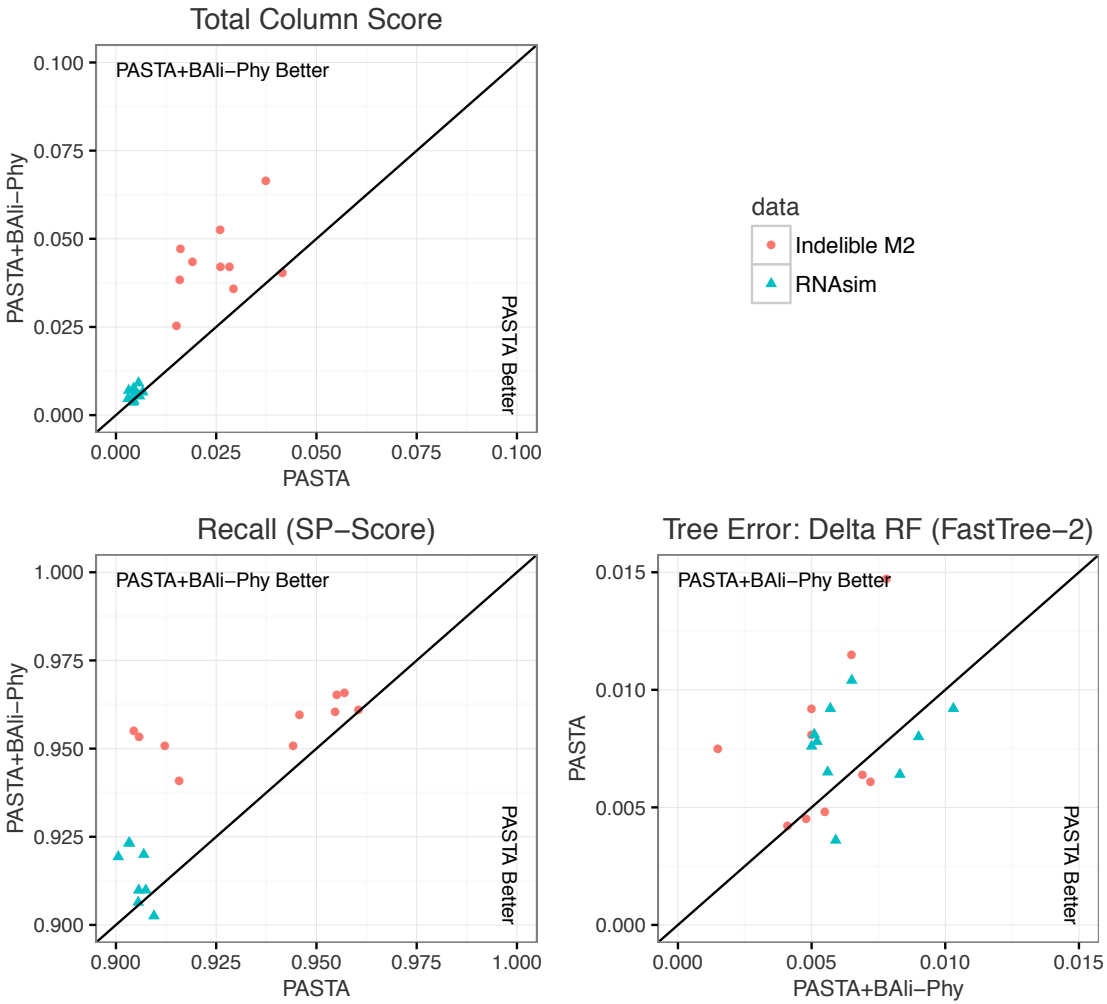
Use BAli-Phy as the subset aligner



Scaling BAli-Phy to 1K sequences (using PASTA)



UPP+BAli-Phy better than default UPP (Scaling BAli-Phy to 10K sequences)

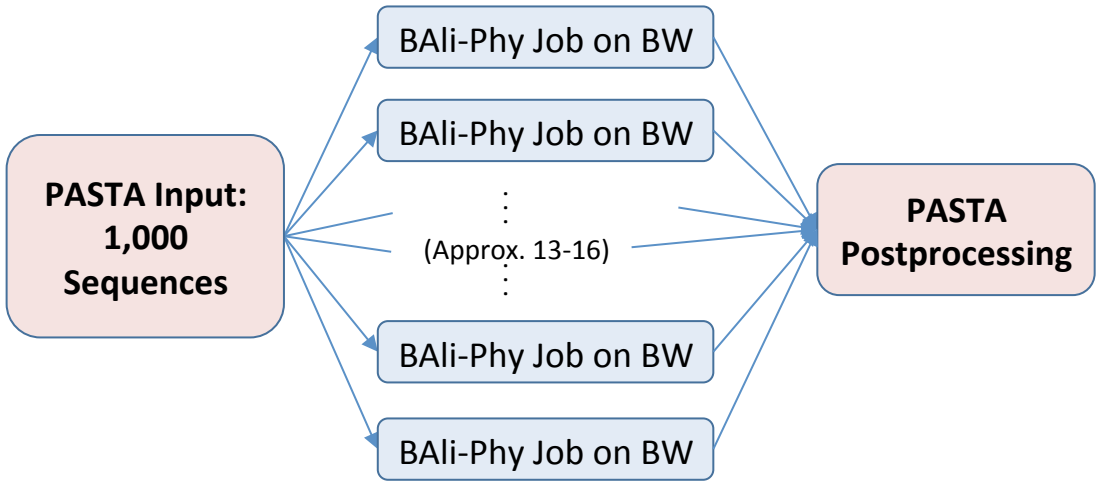


PASTA+BAli-Phy on Blue Waters

Each PASTA job with 1,000 sequences creates 13-16 BAli-Phy jobs

BAli-Phy jobs run asynchronously for 24 hours on single node

Results are downloaded to separate server and processed to yield a single alignment.



Each individual PASTA job takes 300-400 node-hours that can run using the low-priority backfill queue on blue waters.

Data:

- 5 Model Conditions
- 10 Replicates each
- 17.8k node-hours

Additional exploratory data used to develop the method in the paper required an additional 2,164 BAli-Phy processes and 52k node-hours.

Unused capacity on Blue Waters allows the divide-and-conquer alignment algorithm to use the large number of nodes for easily parallelized jobs.

Other projects on Blue Waters

- Coalescent-based species tree estimation
- Supertree methods
- Extensions of UPP (using ensembles of Hidden Markov Models):
  - Metagenomic taxon identification (collaboration with Bill Gropp and Mihai Pop)
  - Protein family classification (collaboration with Jian Peng)

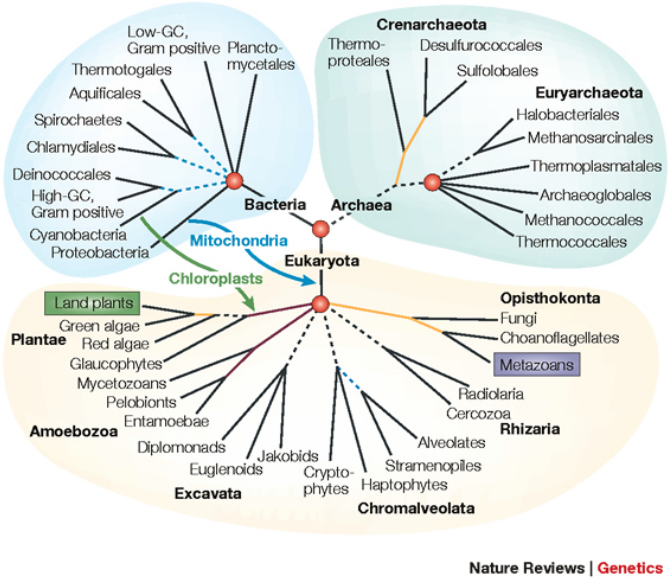
Blue Waters is essential for method development, testing, and refinement.

We also use Blue Waters to analyze biological datasets – when the otherwise available infrastructure is inadequate.

The Tree of Life: *Multiple Challenges*

Scientific challenges:

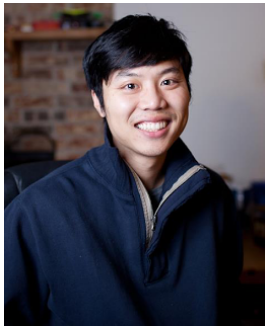
- Ultra-large multiple-sequence alignment
- Gene tree estimation
- Metagenomic classification
- Alignment-free phylogeny estimation
- Supertree estimation
- Estimating species trees from many gene trees
- Genome rearrangement phylogeny
- Reticulate evolution
- Visualization of large trees and alignments
- 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


Papers available at http://tandy.cs.illinois.edu/papers.html

PASTA and UPP at https://github.com/smirarab

Funding: NSF ABI-1458652 and III:AF:1513629, a Founder Professorship from the Grainger Foundation, and HHMI (to S.M.)

Computational support: BlueWaters (PASTA+BAliPhy) and TACC (PASTA and UPP)