Advances in Large-scale Multiple Sequence Alignment

Tandy Warnow
Grainger Distinguished Chair in Engineering
The University of Illinois at Urbana-Champaign
http://tandy.cs.illinois.edu

This PPTX is available online at http://tandy.cs.Illinois.edu/warnow-msa-CGSI-2023.pptx

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

```
S1 = AGGCTATCACCTGACCTCCA        S1 = -AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC            S2 = TAG-CTATCAC--GACCGC--
S3 = TAGCTGACCGC                 S3 = TAG-CT-------GACCGC--
  …                              …
Sn = TCACGACCGACA           →    Sn = -------TCAC--GACCGACA
```

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

What are MSAs used for?

- Inferring evolutionary histories
- Predicting biomolecular (RNA, protein) structure
- Genome annotation and assembly
- And others

Phylogenomic pipeline

- Select taxon set and markers
- Gather and screen sequence data, possibly identify orthologs
- Compute multiple sequence alignments for each locus, and construct gene trees
- Compute species tree or network:
  - Combine the estimated gene trees, OR
  - Estimate a tree from a concatenation of the multiple sequence alignments
- Get statistical support on each branch (e.g., bootstrapping)
- Estimate dates on the nodes of the phylogeny
- 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
- Compute multiple sequence alignments for each locus, and construct gene trees
- Compute species tree or network:
  - Combine the estimated gene trees, OR
  - Estimate a tree from a concatenation of the multiple sequence alignments
- Get statistical support on each branch (e.g., bootstrapping)
- Estimate dates on the nodes of the phylogeny
- Use species tree with branch support and dates to understand biology

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…

- Plant Tree of Life based on transcriptomes of ~1200 species
- More than 13,000 gene families (most not single copy)

Challenge:
Alignment of datasets with > 100,000 sequences with many very short sequences

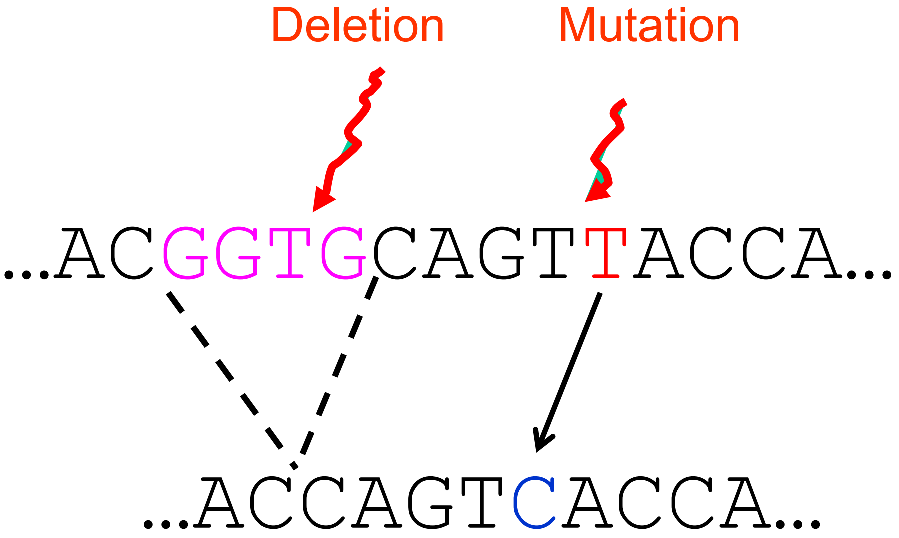
Outline

- Part I: Basic concepts and techniques
  - Computing pairwise alignments
  - Computing multiple sequence alignments
- Part II: Techniques for Large-scale MSA
  - Divide-and-conquer MSA
  - “Two-phase” (compute backbone, then add remaining sequences)
- Part III: Adding to alignments
- Part IV: Discussion and Future Work

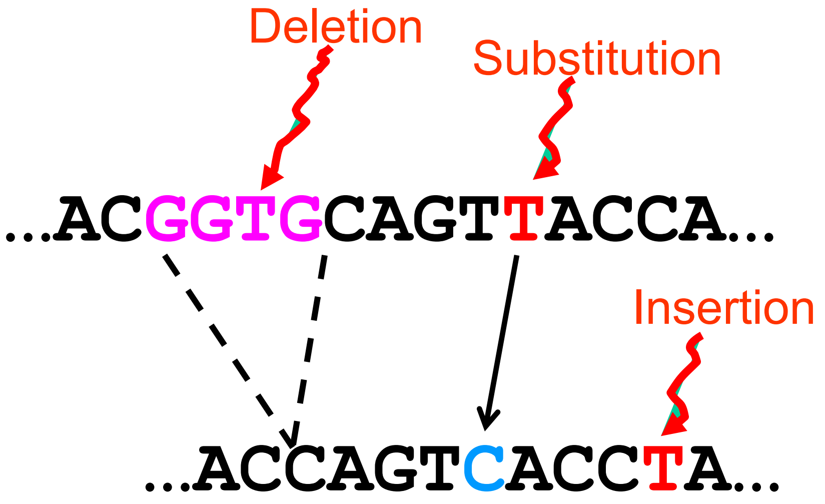
Part I: Basic Concepts

- Homology
- Indels (insertions and deletions)
- True pairwise alignment
- Edit transformation
- Global vs local alignment
- Transitivity

Indels (insertions and deletions)



Homology: two letters (nucleotides or amino-acids) that are related by descent from a common ancestor



Deletion

Substitution

...ACGGTGCAGTTACCA...

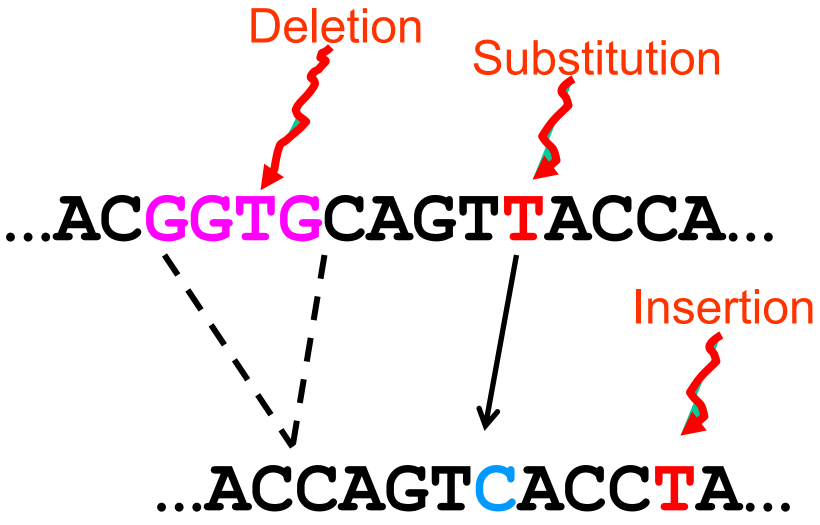
Insertion

...ACCAGTCACCTA...

```
...ACGGTGCAGTTACC-A...
...AC----CAGTCACCTA...
```

The **true pairwise alignment**

- Reflects historical substitution, insertion, and deletion events
- Letters (nucleotides or amino acids) in the same column are supposed to be homologs



Deletion

Substitution

…ACGGTGCAGTTACCA…

Insertion

…ACCAGTCACCTA…

Then two deletions (one at front, long one at end)

CCAGT

```
…ACGGTGCAGTTACC-A…
…AC----CAGTCACCTA…
…-C----CAGT------…
```

The **true multiple alignment**

- Reflects historical substitution, insertion, and deletion events
- Defined using transitive closure of pairwise alignments defined on the edges of the true tree

Pairwise alignment

- Global alignment: finding the lowest-cost edit transformation, solved using Needleman-Wunsch (dynamic programming)
- Polynomial time!
- Allows for variations in cost function and similarity scores, still polynomial time

Examples

For each pair of sequences, what is the best global pairwise alignment?

Suppose that indels and substitutions each have cost 1.

- S1 = ACTAG
- S2 = GCTAG

- S3 = ACTAG
- S4 = TTACTAGGA

- S5 = TTAAGAGAACTATGGACCTA
- S6 = GAGAAGGTAGGTTTAAGTAAGCCATTA

DP algorithm for the edit distance

Input: sequences a and b of lengths m and n, respectively.
Output: minimum number of indels and substitutions needed to transform a into b.
A two-dimensional matrix, F[0..m,0..n] is used to hold the edit distance values:

$F(i,j) = d(a[1..i], b[1..j])$ (Definition of what we want)

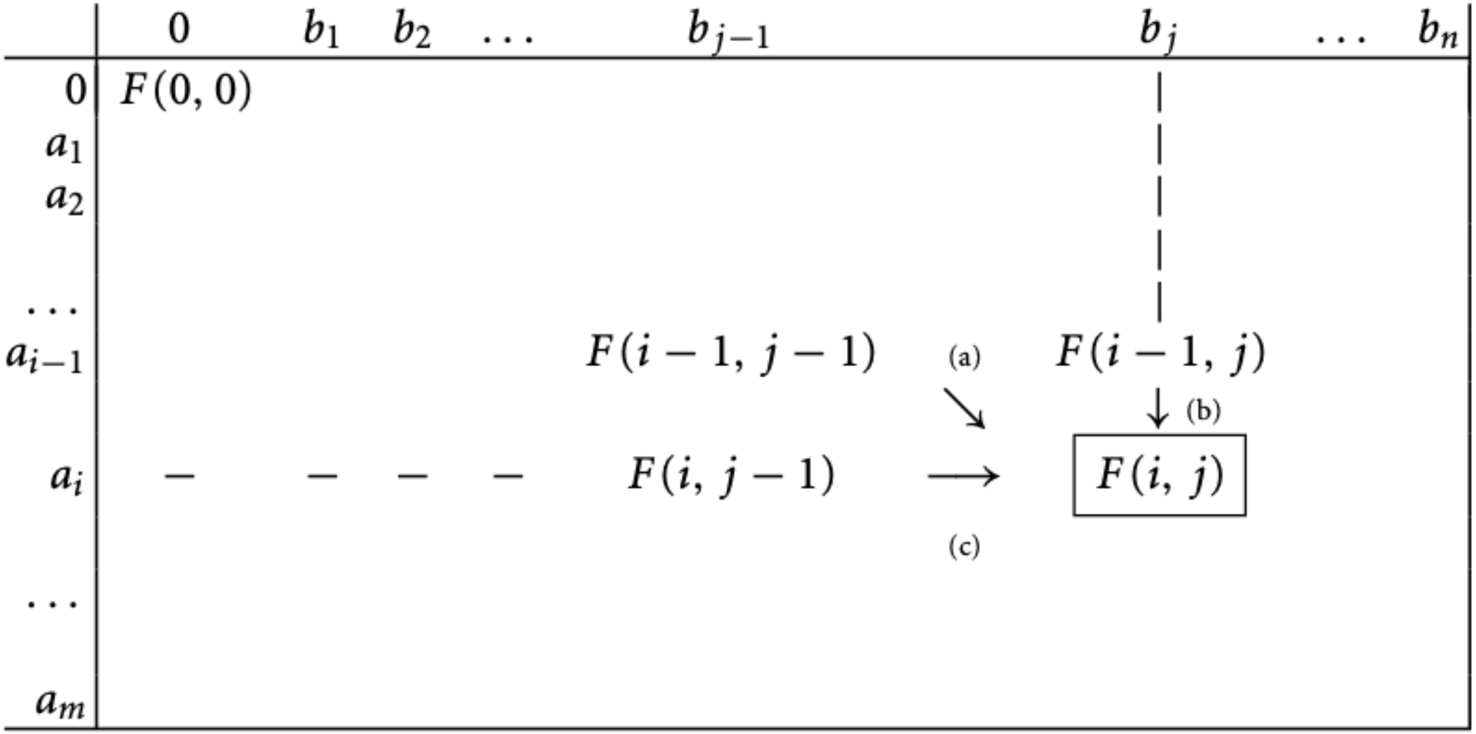
$F(0,0) = 0$

$F(i,0) = i, i = 1..m$

$F(0,j) = j, j = 1..n$

```
For $i,j \geq 1$, $F[i,j] = min\{$
    $F[i-1, j-1]+$ if a[i]=b[j] then 0 else 1,
    $F[i-1, j] + 1$,
    $F[i, j-1] + 1$
    $\}$
```

Needleman-Wunsch



From Huson et al., (2010)

Pairwise alignment

- Global alignment: finding the lowest-cost edit transformation, solved using Needleman-Wunsch (dynamic programming)
- Local alignment: finding the two substrings of highest similarity, solved using Smith-Waterman (also dynamic programming)
- Polynomial time!
- Both allow for variations in cost function and similarity scores, still polynomial time

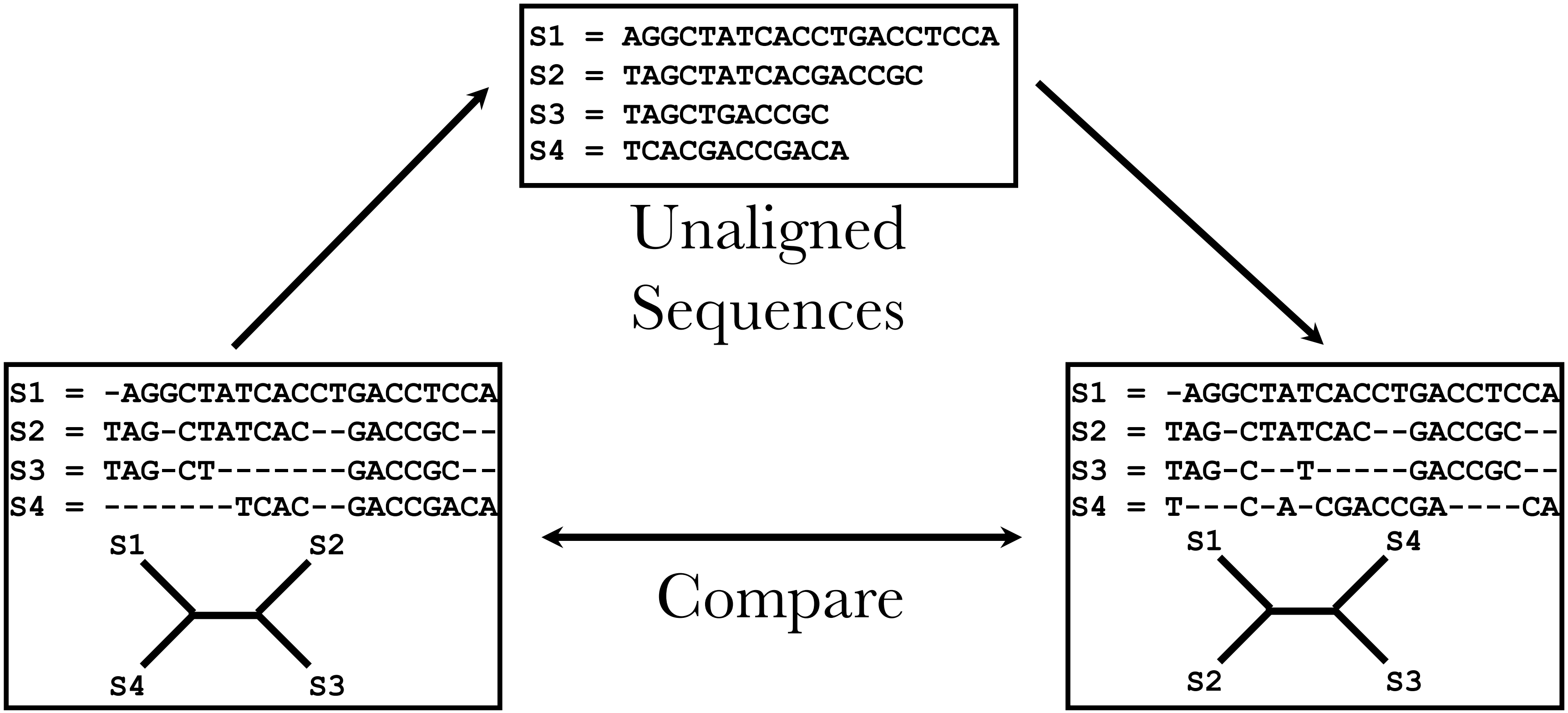
Multiple Sequence Alignment

- Optimization problems extend pairwise alignment
  - Minimizing sum-of-pairs costs
  - Minimizing tree length
  - Likelihood-based approaches (e.g., Bayesian estimation)
- Optimization problems are NP-hard
- Bayesian estimation is even less scalable

Standard approaches?

- Standard methods use a variety of techniques, such as extending pairwise alignments with:
  - Star alignment
  - Progressive alignment
  - Ensemble methods, including “Consistency”
  - Supervised learning

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

MSA+Tree estimation

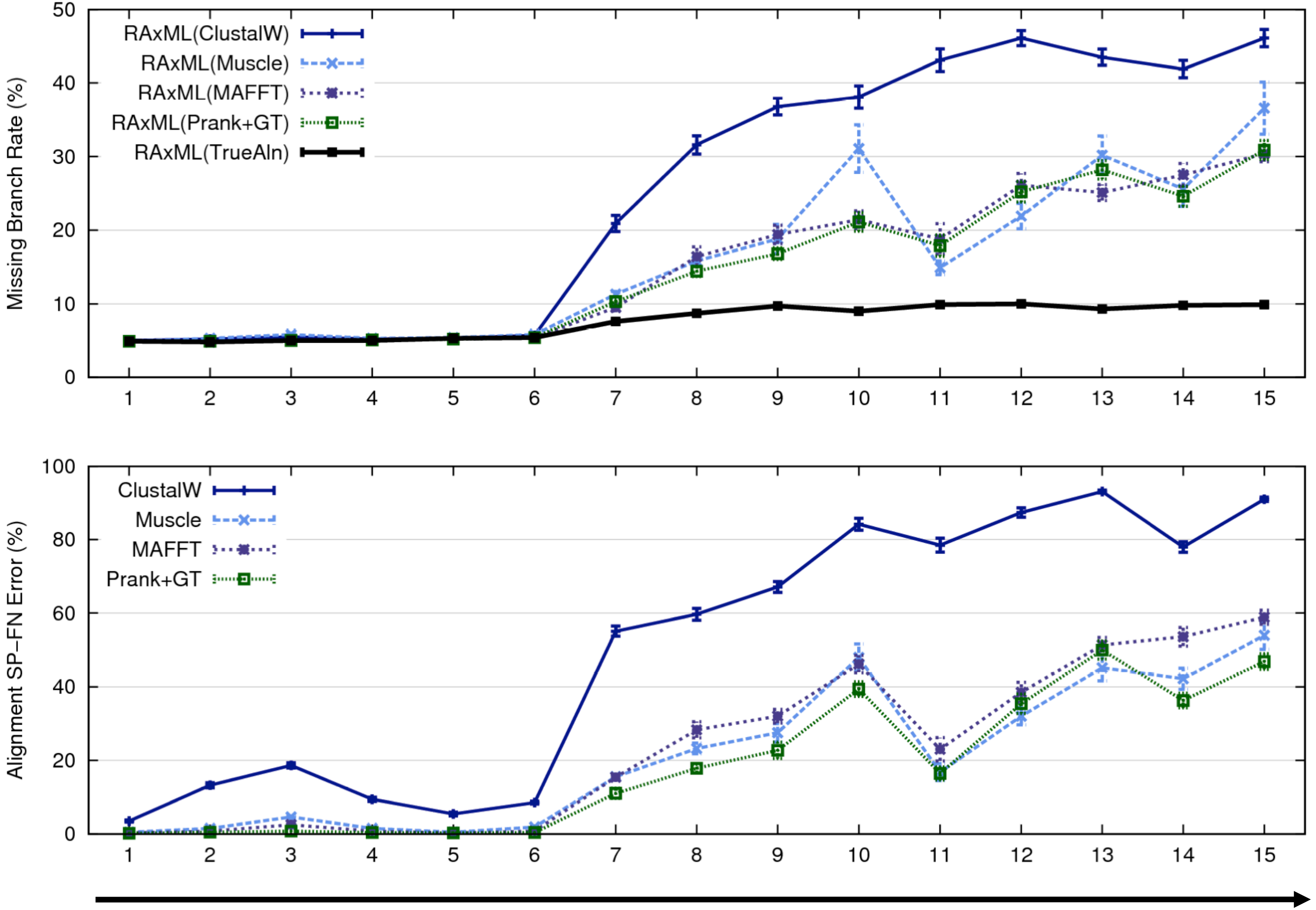
Alignment methods

- 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*



1000-taxon models, ordered by difficulty (Liu et al., 2009)

What makes for an “easy” MSA?

- MSA is easy when the input is a small set of very similar sequences
    - All nearly the same length
    - Very few substitutions
    - Very few “indels”
- But large datasets are difficult, even when they are otherwise relatively “easy”

Part II: Techniques for Large-Scale MSA

Large-scale MSA

Challenges

- High evolutionary rates
- Sequence length heterogeneity (e.g., fragments)
- Very long sequences

Techniques for *de novo* MSA

- Divide-and-conquer
  - Divide dataset (sequences) into disjoint subsets, align subsets, merge subset alignments
- "Two-phase" (mainly for datasets with sequence length heterogeneity)
  - Construct "backbone alignment" for the full-length sequences
  - Add remaining sequences into the backbone alignment

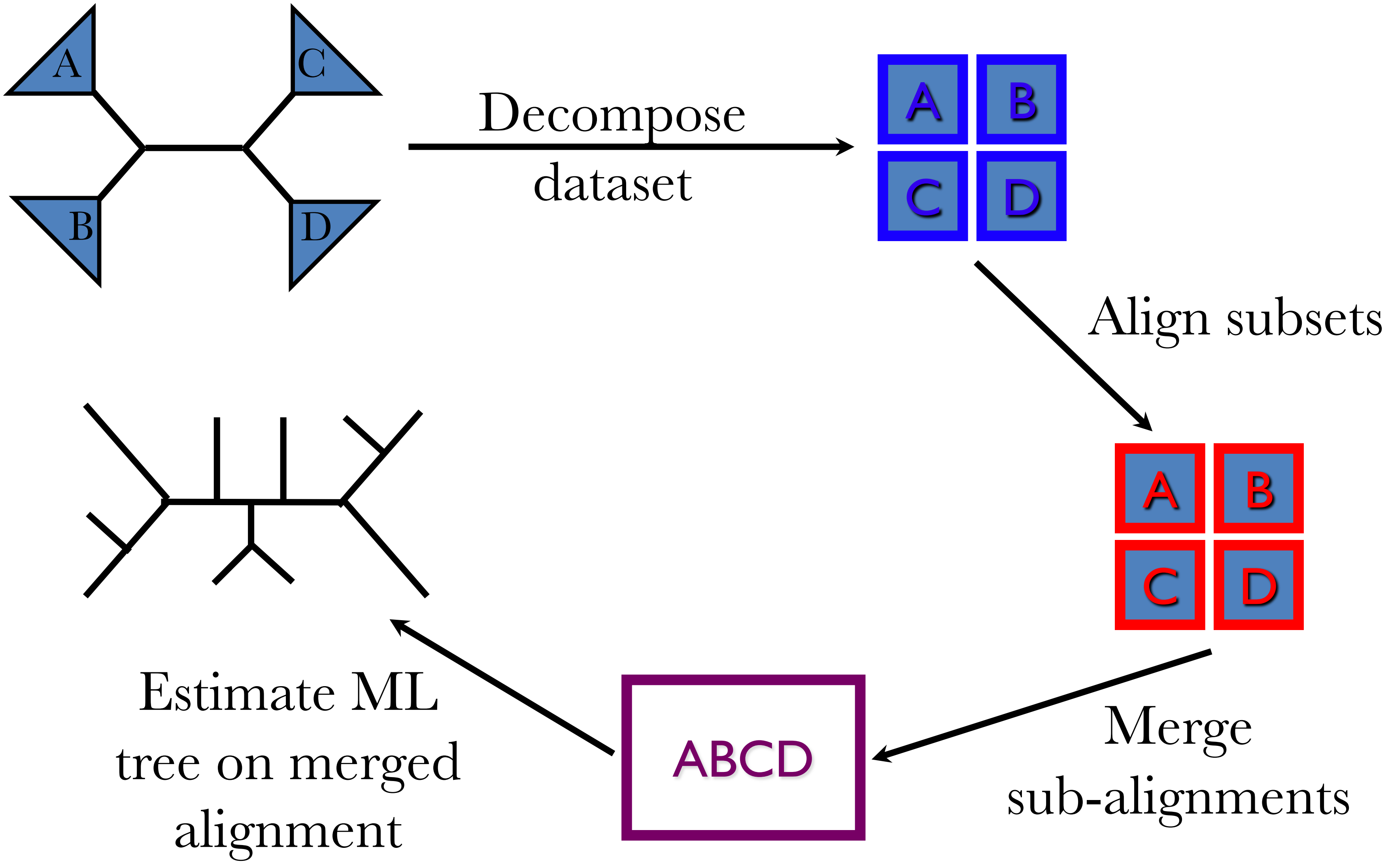
Techniques for *de novo* MSA

- Divide-and-conquer
  - Divide dataset (sequences) into disjoint subsets, align subsets, merge subset alignments
- “Two-phase” (mainly for datasets with sequence length heterogeneity)
  - Construct “backbone alignment” for the full-length sequences
  - Add remaining sequences into the backbone alignment

Divide-and-conquer

- Divide-and-conquer “meta-methods” for large numbers of sequences and high evolutionary rates:
    - SATé, PASTA, and MAGUS

Re-aligning on a tree



Decompose dataset

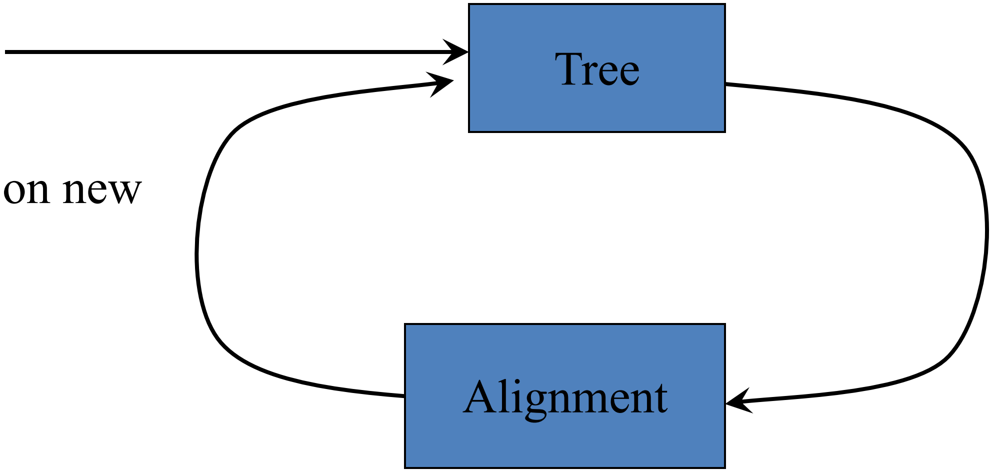
Align subsets

Merge sub-alignments

Estimate ML tree on merged alignment

SATé, PASTA, and MAGUS Algorithms

Obtain initial alignment and estimated ML tree



Tree

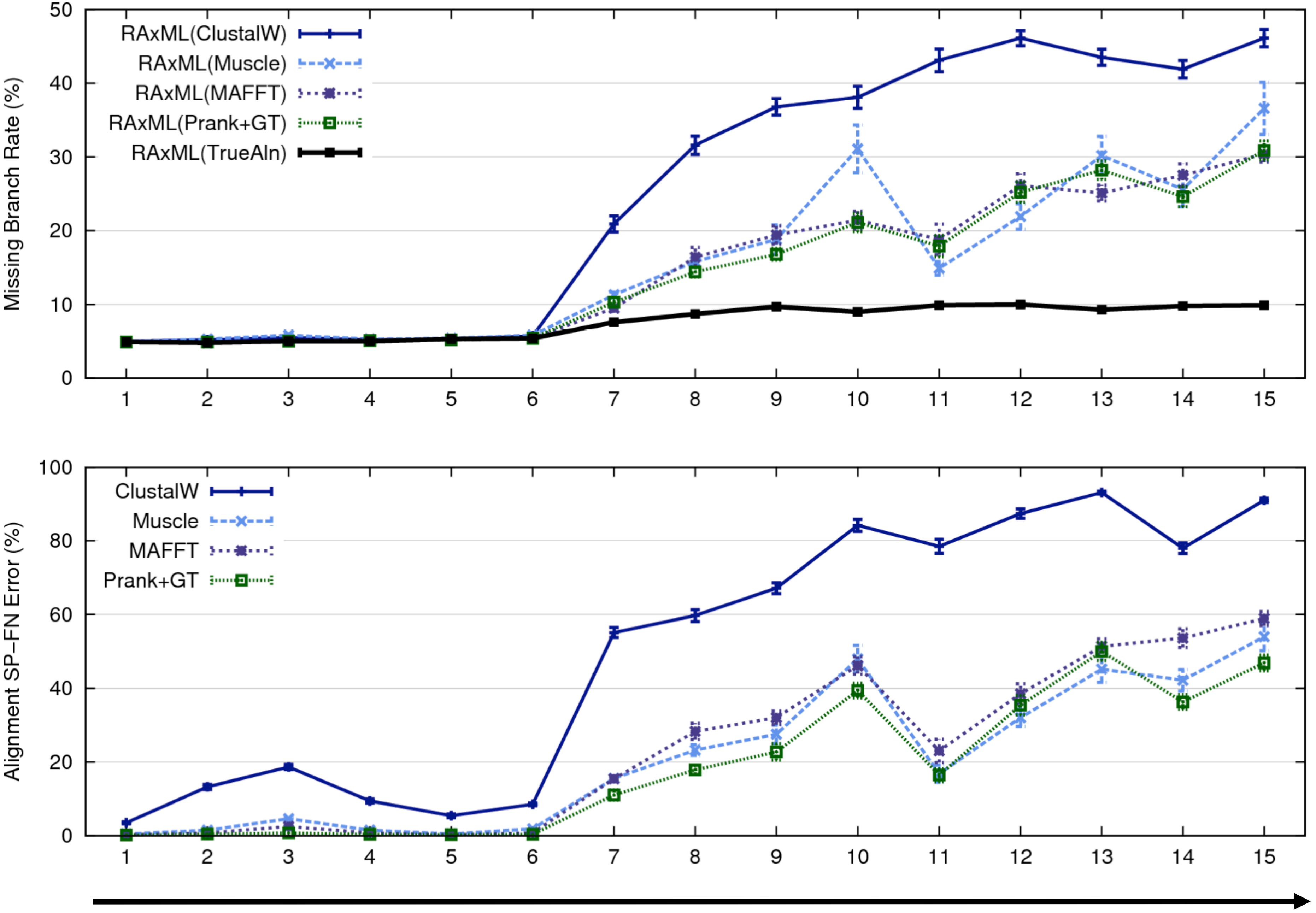
Estimate ML tree on new alignment

Use tree to compute new alignment

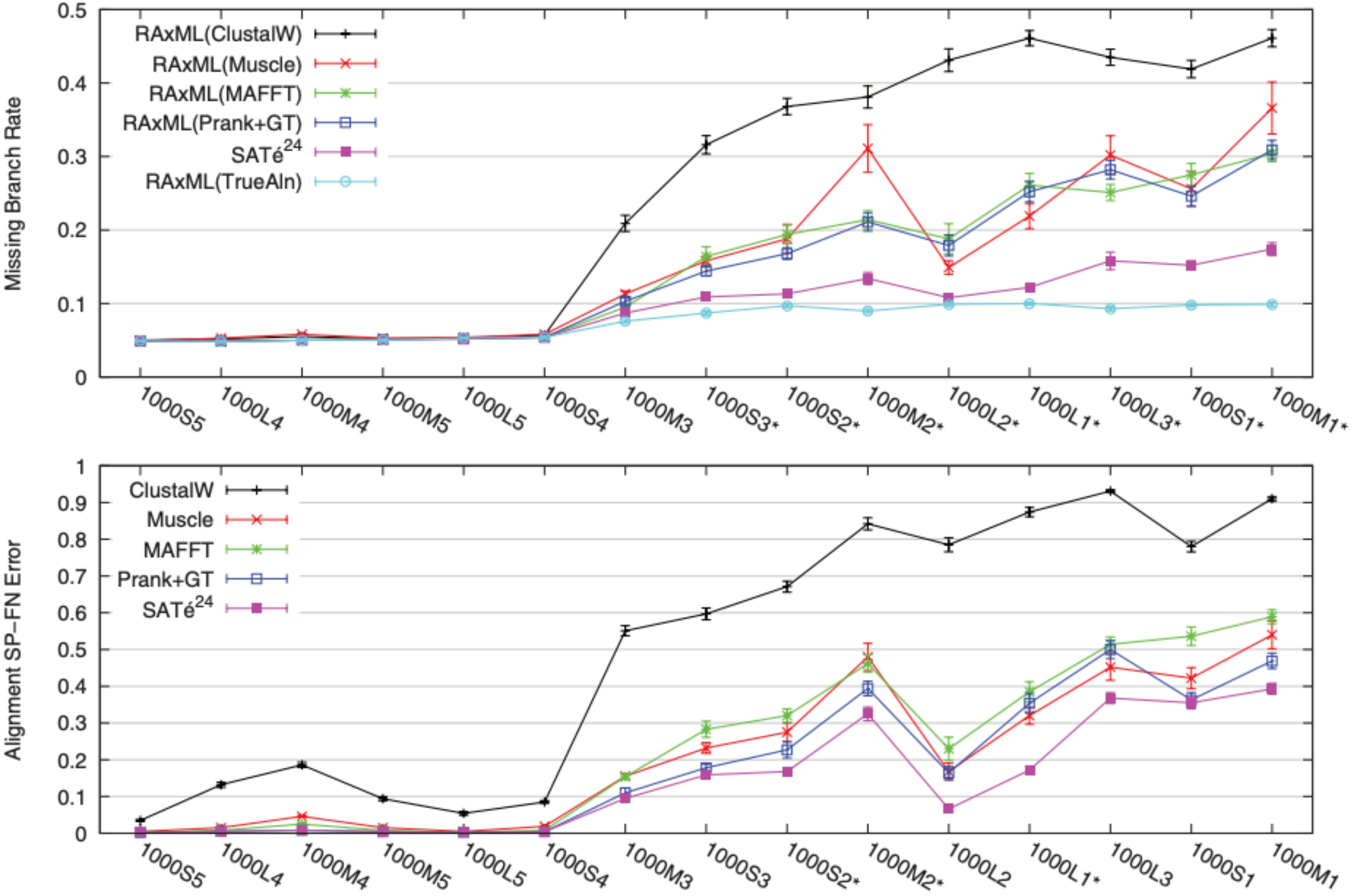
Alignment

Repeat until termination condition, and

return the alignment/tree pair with the best ML score



1000-taxon models, ordered by difficulty (Liu et al., 2009)


Missing Branch Rate
0
0.1
0.2
0.3
0.4
0.5
RAxML(ClustalW)
RAxML(Muscle)
RAxML(MAFFT)
RAxML(Prank+GT)
SATé24
RAxML(TrueAln)
1000S5
1000L4
1000M4
1000M5
1000L5
1000S4
1000M3
1000S3*
1000S2*
1000M2*
1000L2*
1000L1*
1000L3*
1000S1*
1000M1*
Alignment SP−FN Error
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
ClustalW
Muscle
MAFFT
Prank+GT
SATé24
1000S5
1000L4
1000M4
1000M5
1000L5
1000S4
1000M3
1000S3
1000S2*
1000M2*
1000L2
1000L1*
1000L3
1000S1
1000M1

Improvement over time

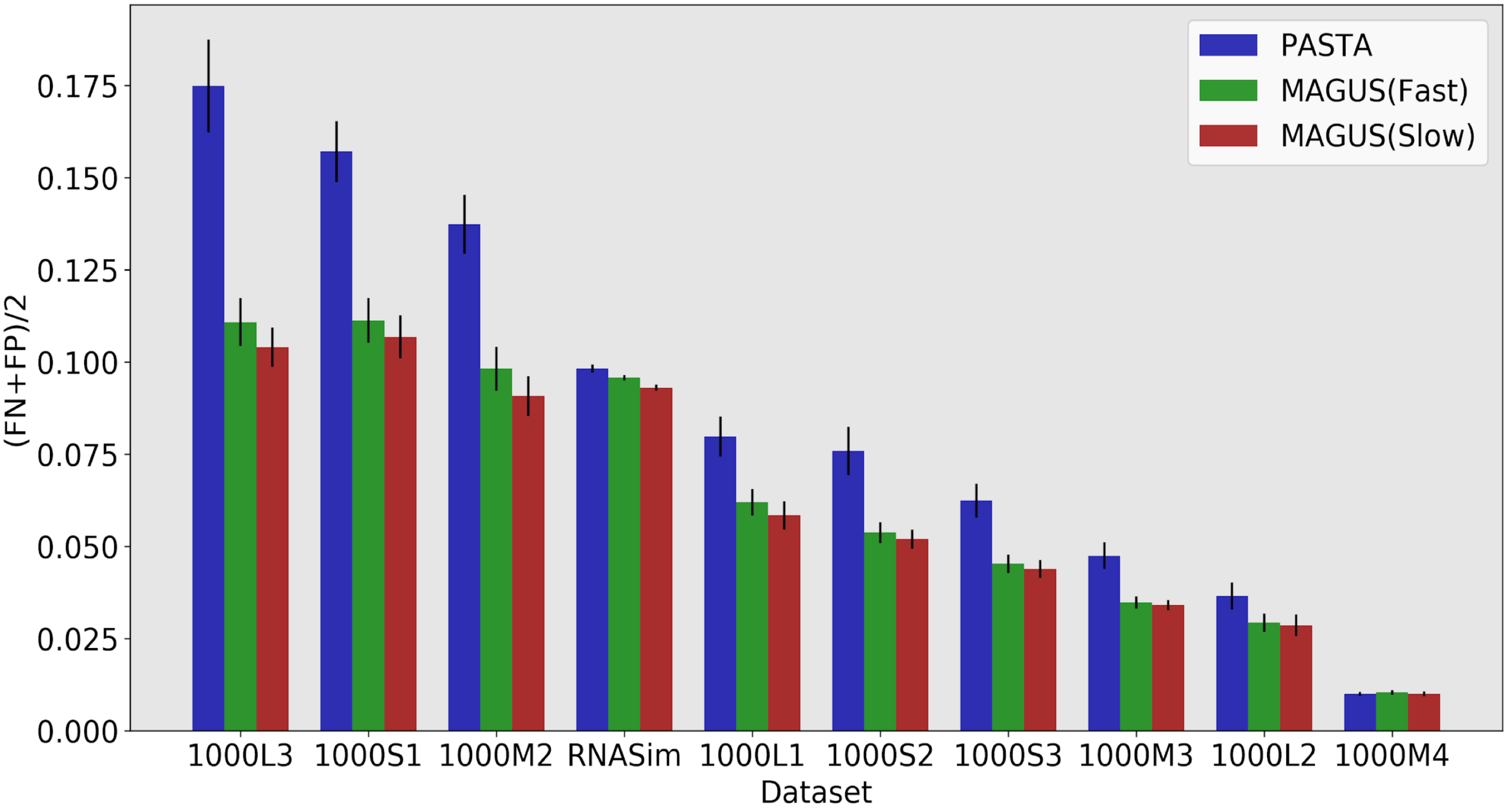
- SATé-1 (Science 2009): up to about 8,000
- SATé-2 (Syst Biol 2012): up to 50,000
- PASTA (J Comp Biol 2014): up to 1,000,000
- MAGUS (Bioinformatics 2021): more accurate than PASTA (and one iteration suffices) – up to 1,000,000

Each method improved on the previous with respect to MSA and Tree accuracy, speed, and scalability to large datasets

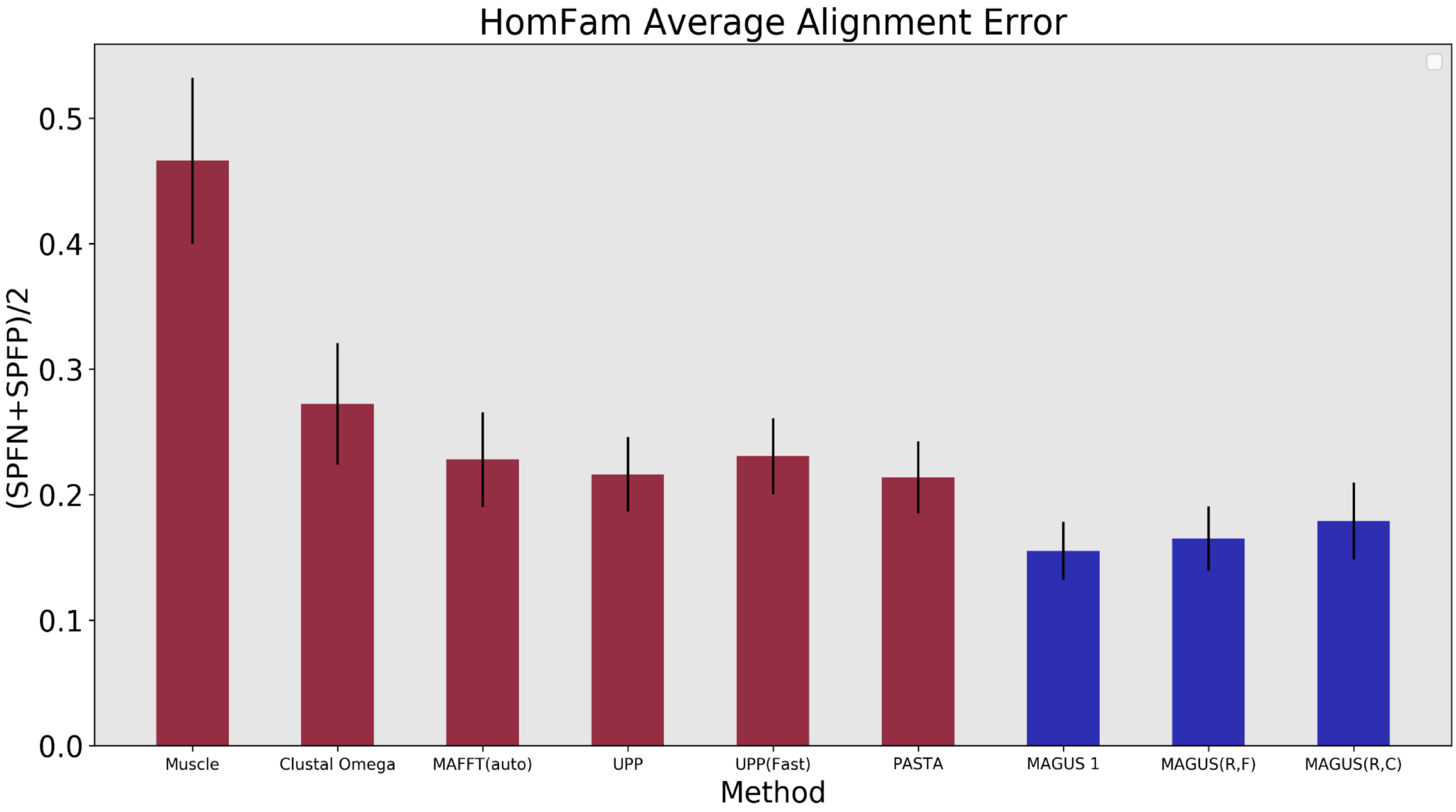
SATé-II vs PASTA vs MAGUS

- Decomposition: the same technique (delete centroid edges)
- Subset alignments: the same (all computed MAFFT-linsi alignments)
- Merging:
  - SATé-II uses a guide tree to merge the subset alignments up the tree
  - PASTA aligns all “adjacent pairs” of alignments, and then finishes with transitivity
  - MAGUS aligns all subset alignments *at once* (using a complex pipeline involving Markov Clustering)

MAGUS: More Accurate Alignments than PASTA



MAGUS: excellent on protein benchmarks too



Summary for Divide-and-Conquer

- Can be used with any base MSA method (we showed results with MAFFT-linsi, but improvements also found for other methods, including BAli-Phy)
- Iteration can help
- Merging alignments “all at once” promising; related to John Kececioglu’s “Maximum Weight Trace” problem

Sequence Length Heterogeneity

- The next challenge is sequence length heterogeneity (especially fragmentary sequences)

1kp: Thousand Transcriptome Project

G. Ka-Shu Wong
U Alberta

J. Leebens-Mack
U Georgia

N. Wickett
Northwestern

N. Matasci
iPlant

T. Warnow,
UIUC

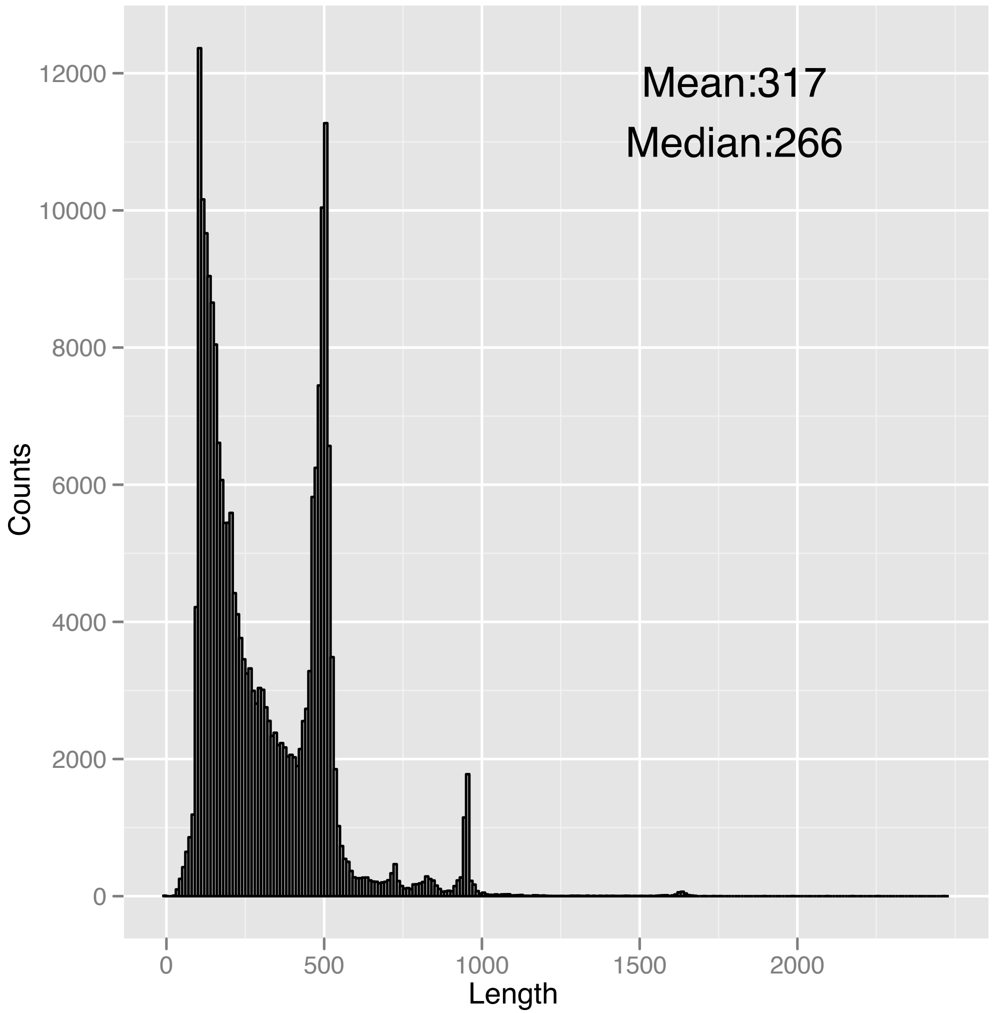
S. Mirarab,
UT-Austin

N. Nguyen
UT-Austin

Plus many many other people…

- Plant Tree of Life based on transcriptomes of ~1200 species
- More than 13,000 gene families (most not single copy)

Challenge:
Alignment of datasets with > 100,000 sequences with many very short sequences



1KP dataset: more than 100,000 p450 amino-acid sequences, many fragmentary

All standard multiple sequence alignment methods we tested performed poorly on datasets with fragments.

Solution: Two-phase approach

- Phase 1: Select a collection of “full-length” sequences, and compute a “backbone” alignment on them.
- Phase 2: Add the remaining sequences into the backbone alignment.

Note: Each stage matters!

- Depends on which sequences are in the backbone, and how the backbone alignment is computed (but can use expensive methods)
- Depends on how the remaining sequences are added to the backbone (can use “local alignment” techniques)

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

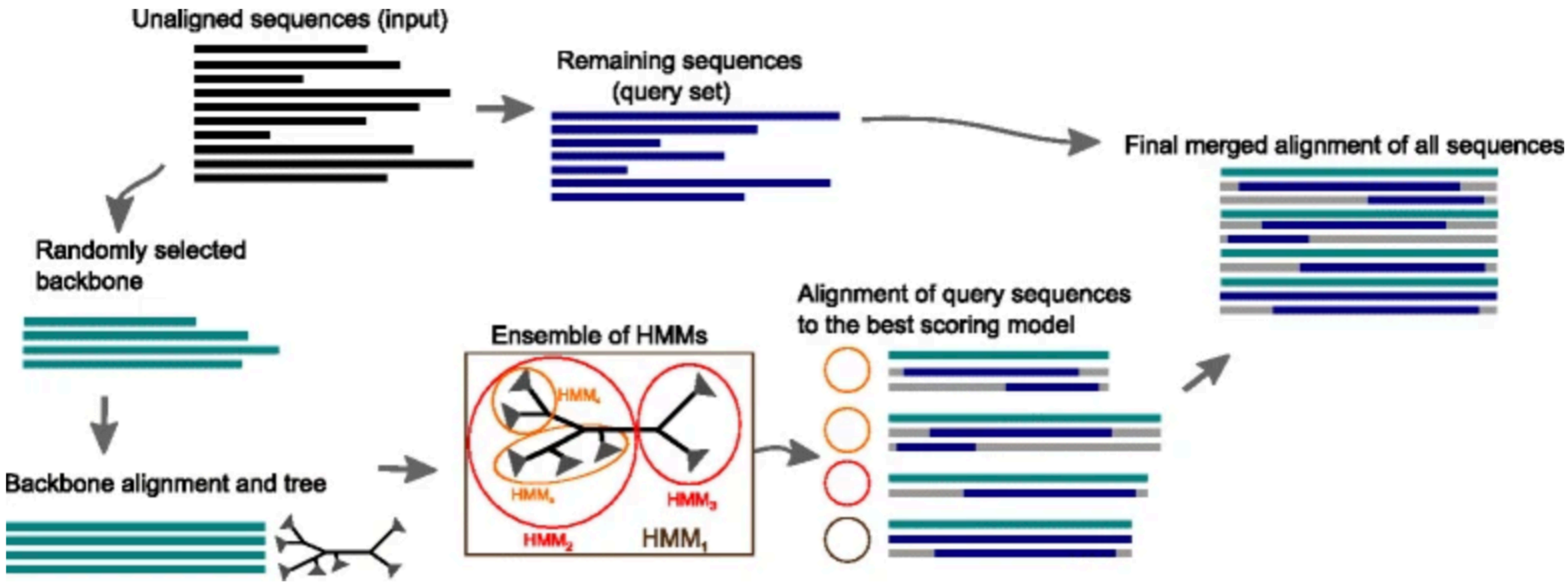
UPP = “Ultra-large multiple sequence alignment using Phylogeny-aware Profiles”

Nguyen, Mirarab, and Warnow. Genome Biology, 2015

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

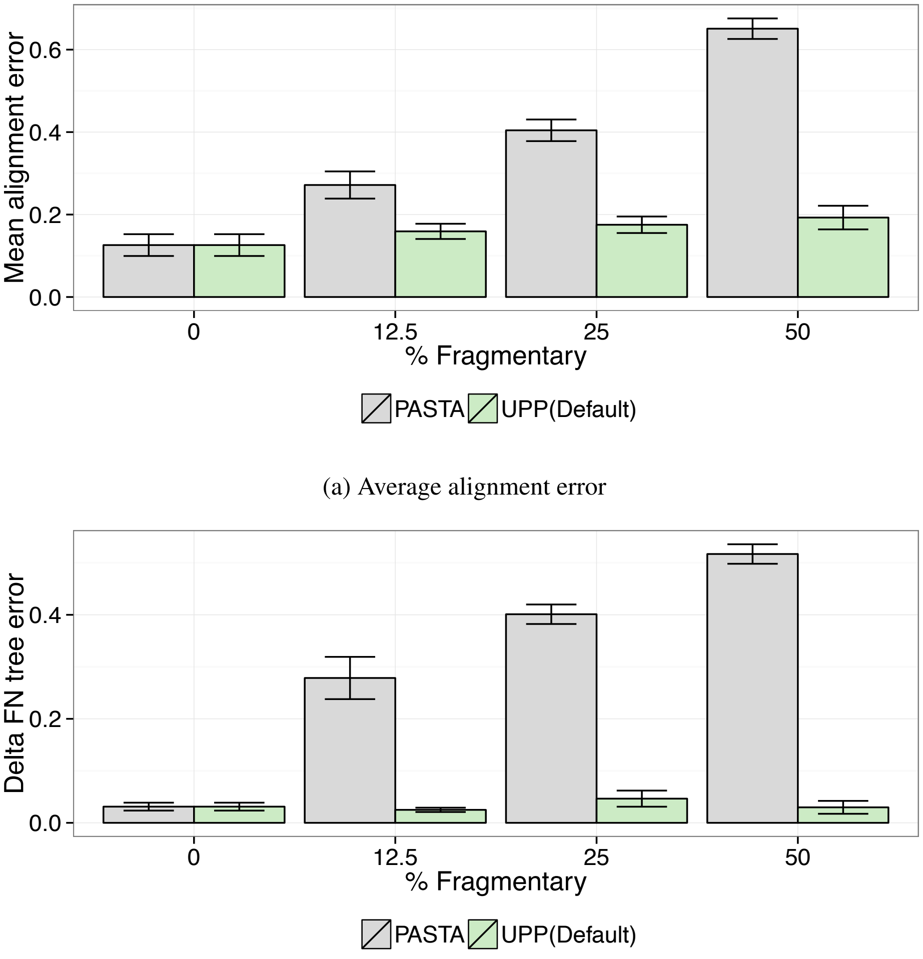
Uses an ensemble of HMMs

UPP (Nguyen et al. 2015)


Unaligned sequences (input)
Remaining sequences (query set)
Final merged alignment of all sequences
Randomly selected backbone
Alignment of query sequences to the best scoring model
Ensemble of HMMs
HMM$_4$
HMM$_5$
HMM$_3$
HMM$_2$
HMM$_1$
Backbone alignment and tree

The Ensemble of Hidden Markov Models is a “model” for the backbone alignment.
The HMMs are built on subset alignments, may not be clades in the backbone tree.

UPP vs. PASTA: impact of fragmentation



(a) Average alignment error

(b) Average tree error

Under high rates of evolution, PASTA is badly impacted by fragmentary sequences (the same is true for other methods).

Under low rates of evolution, PASTA can still be highly accurate (data not shown).

UPP continues to have good accuracy even on datasets with many fragments under all rates of evolution.

Performance on fragmentary datasets of the 1000M2 model condition

Other two-phase methods

These methods start the same as UPP (extract backbone alignment, build ensemble of HMMs on backbone), but then do things differently to add the query sequences to the backbone

- WITCH (Chengze Shen et al, J. Comp Biol 2022.) and WITCH-ng (Baqiao Liu and T. Warnow, Bioinformatics Advances 2022.): weights the HMMs, computes extended alignment for each HMM, merges the extended alignments using “consensus alignment” technique
- HMMerge (Minhyuk Park and T. Warnow, Bioinformatics Advances.): weights the HMMs, combines them into new Hidden Markov Model (not profile HMM), and uses that new HMM to add query sequences

All are more accurate than UPP

Part III: Adding to MSAs

Problem:

- Input: MSA A and set Q of additional (unaligned) sequences
- Output: add the sequences in Q to A (without changing A)

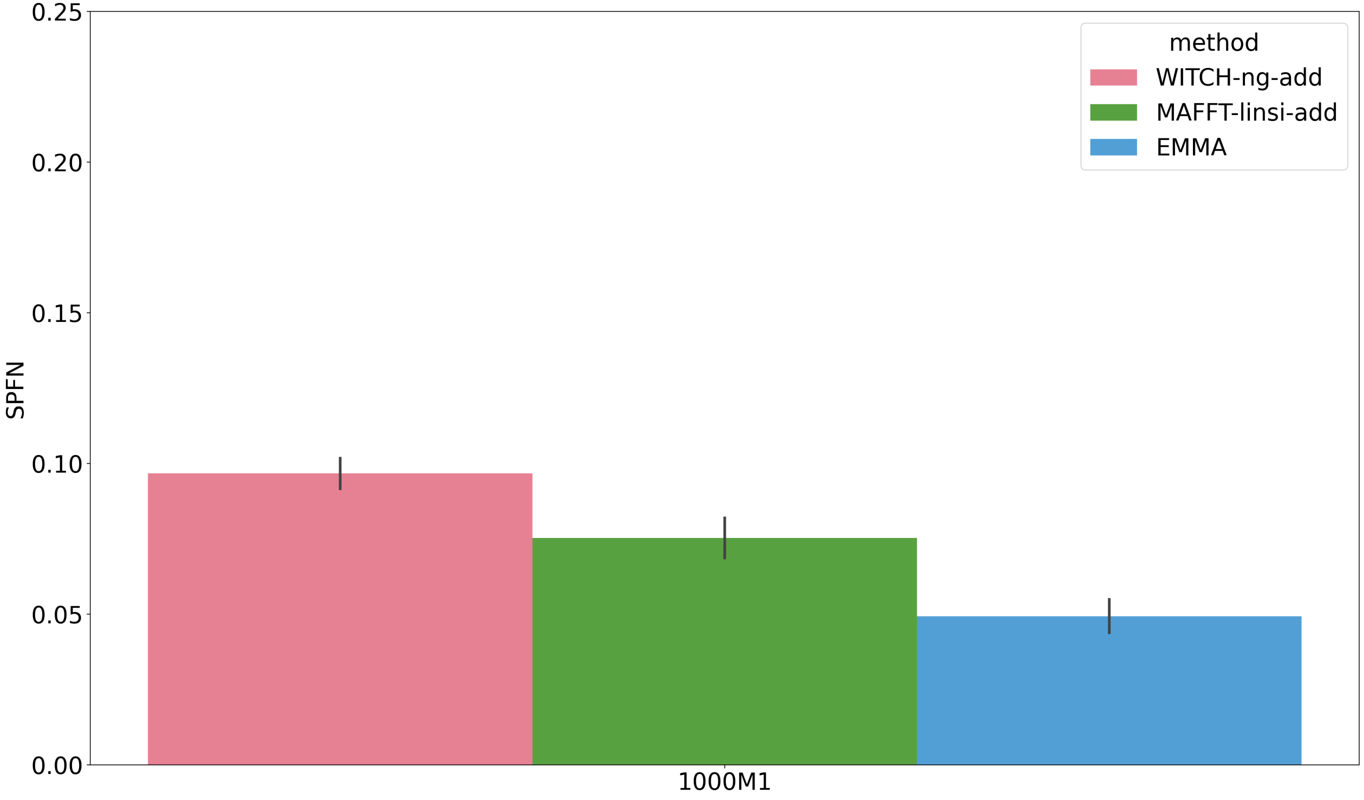
Applications:

- Two-phase methods (like UPP, WITCH, etc)
- Taxon identification (e.g., in metagenomics)
- Updating existing alignment

Methods for updating MSAs

- HMM-based methods: UPP-add, WITCH-add, WITCH-ng-add, HMMerge-add
- MAFFT-add (and its most accurate setting, MAFFT-linsi-add)
- EMMA (Chengze Shen et al., WABI 2023):
  - Extends the ideas in UPP-add, but follows by running MAFFT-linsi-add. (New version under development)

Comparison of EMMA-add, WITCH-ng-add, and MAFFT-linsi-add: The benefit of ***not*** using HMMs to align query sequences



SPFN of different methods (fraction of missing true pairwise homologies)

Dataset: 1000M1 with a high rate of evolution; all sequences are full-length
Backbone: 250 randomly selected sequences from full set.

Part IV: Discussion

Progress in MSA has been made

- MSA is challenging, but algorithmic techniques can improve accuracy and scalability:
  - Dataset size can be addressed using good divide-and-conquer approaches.
  - Heterogeneity in sequence length can be addressed using “local alignment” approaches, such as profile HMMs, with ensembles of profile HMMs providing improved accuracy.

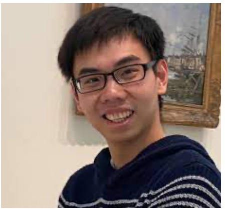
Algorithmic challenges

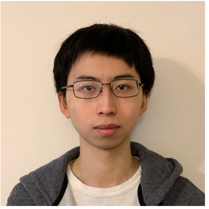
- How can we assess alignment uncertainty and use it in downstream analyses?
- Can we use a set of MSAs to advantage, instead of a single MSA? For example, can we develop effective and efficient “ensemble” methods?
- What are the best ways to merge disjoint alignments?
- How can we efficiently perform statistical alignment?

Summary

- Multiple sequence alignment (MSA) has large downstream consequences in bioinformatics analyses.
- MSA is far from solved – esp. (but not only) on large datasets with high rates of evolution, sequence length heterogeneity, and streaming data.
- New techniques show promise
- Not discussed: multiple whole genome alignment, MSA with rearrangements, statistical alignment

Acknowledgments





PASTA and UPP: Siavash Mirarab and Nam-phuong Nguyen
MAGUS: Vlad Smirnov
WITCH: Chengze Shen and Minhyuk Park
EMMA: Chengze Shen and Baqiao Liu

NSF grants: 1458652 and 2006069
Grainger Foundation (at UIUC)
TACC, Blue Waters, and UIUC campus cluster

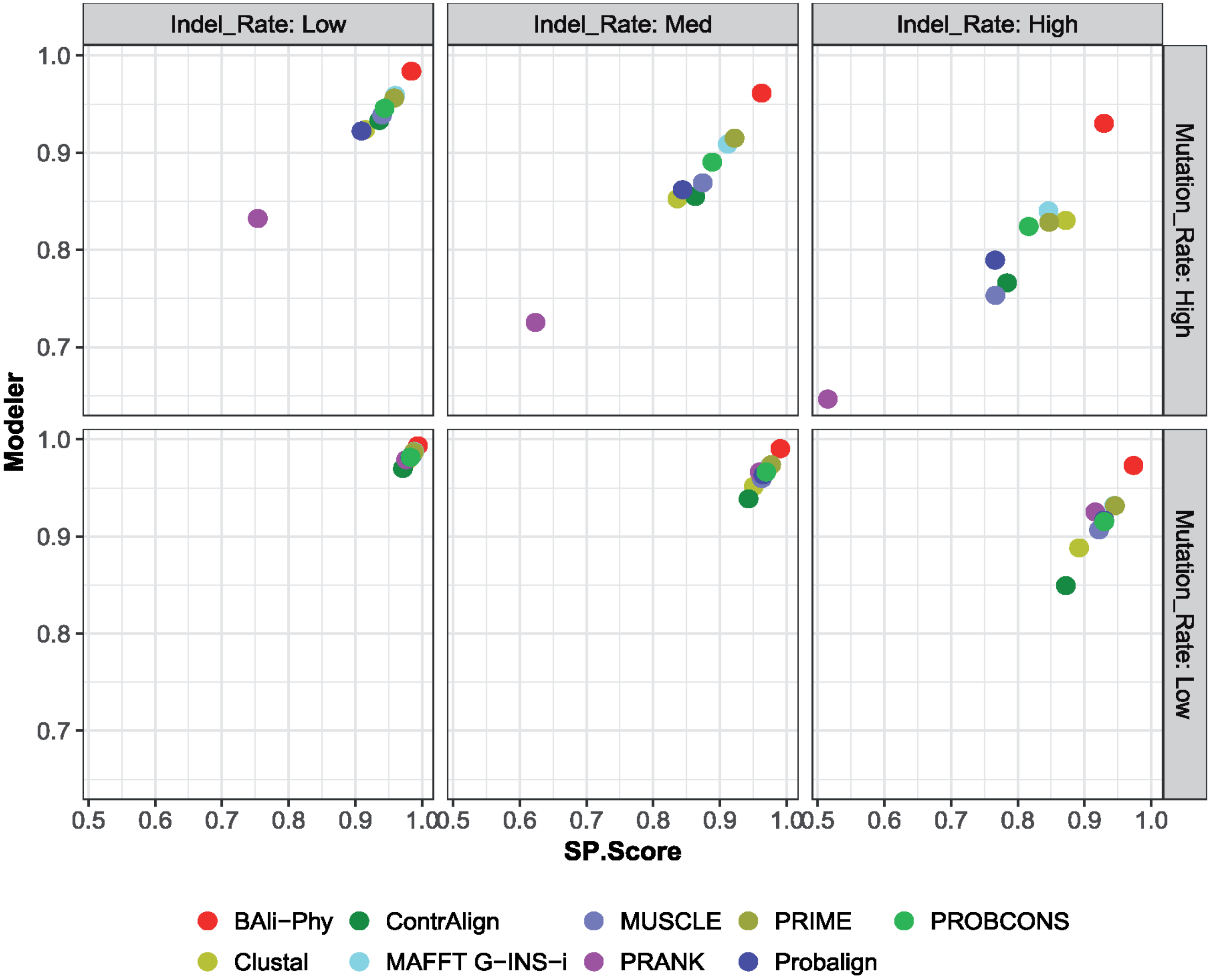
PASTA, UPP, SEPP, and TIPP are available on github at https://github.com/smirarab/
PASTA+BAli-Phy at http://github.com/MGNute/pasta
MAGUS: at https://github.com/vlasmirnov/MAGUS
WITCH: at https://github.com/c5shen/WITCH
WITCH-ng at https://github.com/RuneBlaze/WITCH-NG.
EMMA at https://github.com/c5shen/EMMA

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

Part IV: Statistical Alignment

- Since MSA and tree estimation are both about evolution (recognition of homologies due to evolution), can we co-estimate them together, using a statistical model of evolution?
- BAli-Phy (Redelings and Suchard) is the main method for this problem.

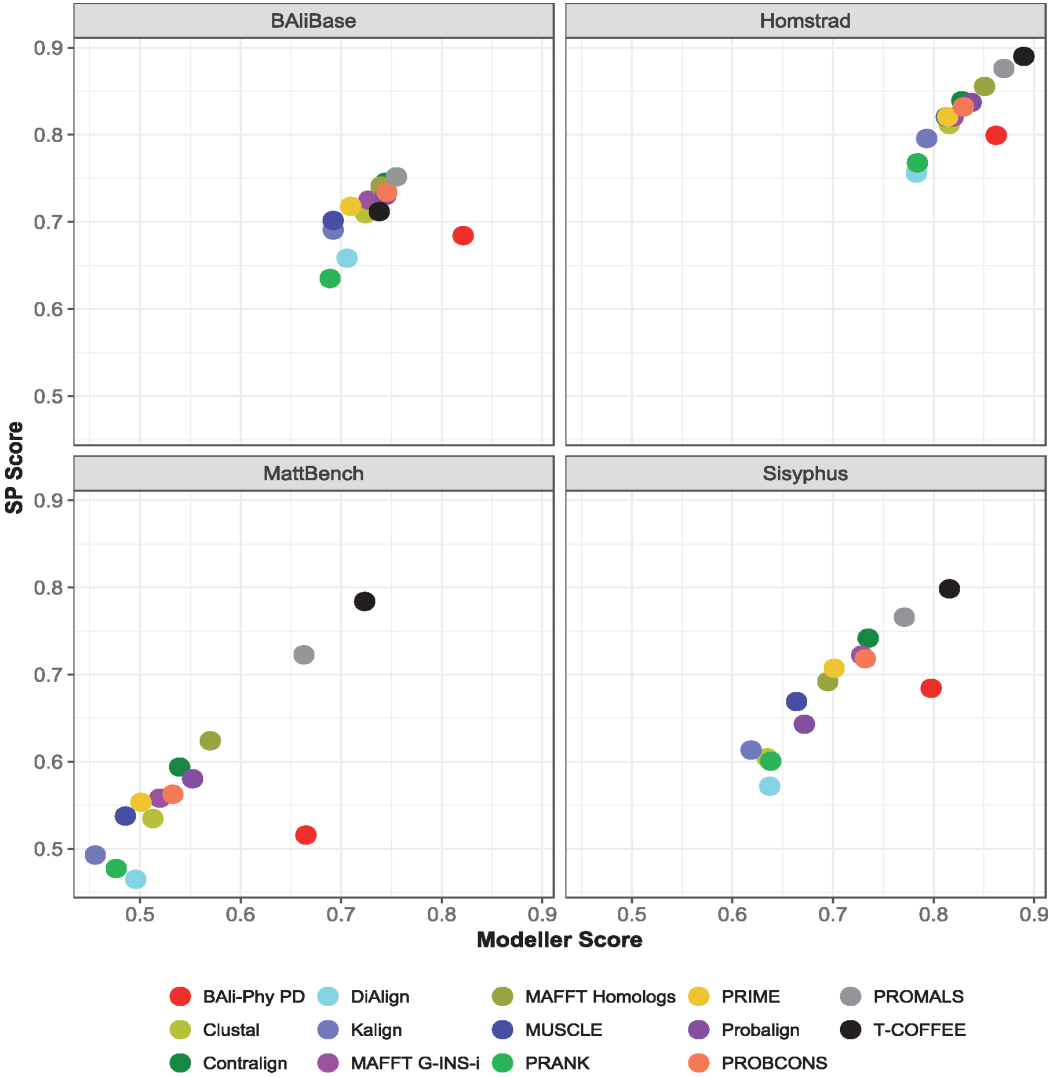
BAli-Phy is best on small simulated protein datasets!



BAli-Phy is best!

Systematic Biology, Volume 68, Issue 3, May 2019, Pages 396–411, https://doi.org/10.1093/sysbio/syy068

BAli-Phy not so great on on 1192 small biological protein datasets



T-Coffee and PROMALS are best!

BAli-Phy good for Modeler score, but not so good for SP-Score (e.g., MAFFT better)

Systematic Biology, Volume 68, Issue 3, May 2019, Pages 396–411, https://doi.org/10.1093/sysbio/syy068

OXFORD
UNIVERSITY PRESS

Observations

- Simulated data: Bali-Phy is the best!
- Protein benchmarks: BAli-Phy in middle
  - Good for Modeler score (so low false positives)
  - Not good for SP-score (so high false negatives)
- BAli-Phy under-aligns on biological datasets, but not on simulated datasets

What is going on?

Most likely not an issue of failure of the MCMC analyses to converge (48 hours, 32 processors, < 30 sequences).

Possible explanations:

1. Model misspecification (i.e., BAli-Phy model not appropriate)
2. Structural alignments and evolutionary alignments different
3. The structural alignments are not correct

All these explanations are likely true, but the relative contributions are unknown.

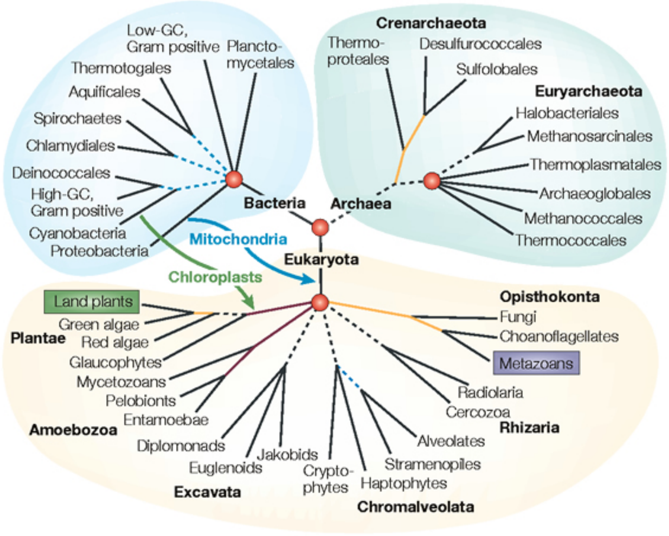
Other questions

- Large datasets can produce extremely large alignments.   How should we handle this?
- Can we predict impact of alignment error on the downstream analysis?
- What are the differences in desirable properties for alignments for different downstream purposes (e.g., protein structure prediction vs. tree estimation)?

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



Nature Reviews | Genetics

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

Testing: simulations and real data

List of papers

- Smirnov, V. & Warnow, T. (2020). MAGUS: Multiple Sequence Alignment using Graph Clustering. Bioinformatics, Volume 37, Issue 12, 15 June 2021, Pages 1666-1672,
- Shen, C., Park, M. & Warnow, T. (2022). WITCH: Improved Multiple Sequence Alignment through Weighted Consensus HMM alignment. J. Computational Biology, Vol. 29, issue 8, pages:782-801
- Nute, MG, Saleh, E., & Warnow, T. (2019). Evaluating statistical multiple sequence alignment in comparison to other alignment methods on protein data sets. Systematic Biology, Volume 68, Issue 3, May 2019, Pages 396-411
- Gupta, M., Zaharias, P., & Warnow, T. (2021). Accurate large-scale phylogeny-aware alignment using BAli-Phy. Bioinformatics, 37(24), 4677-4683.
- Warnow, T. (2021). Revisiting Evaluation of Multiple Sequence Alignment Methods. In: Katoh K. (eds) Multiple Sequence Alignment. Methods in Molecular Biology, vol 2231.