Ensembles of Hidden Markov Models and their use in Bioinformatics

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

Supported by NSF grant 2006069

This talk

- Multiple Sequence Alignment (MSA): challenges and progress
- MSA using ensembles of HMMs
- Applications of eHMMs
- Statistical alignment (e.g., BAli-Phy)

This talk

- Multiple Sequence Alignment (MSA): challenges and progress
- MSA using ensembles of HMMs
- Applications of eHMMs
- Statistical alignment (e.g., BAli-Phy)

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
```

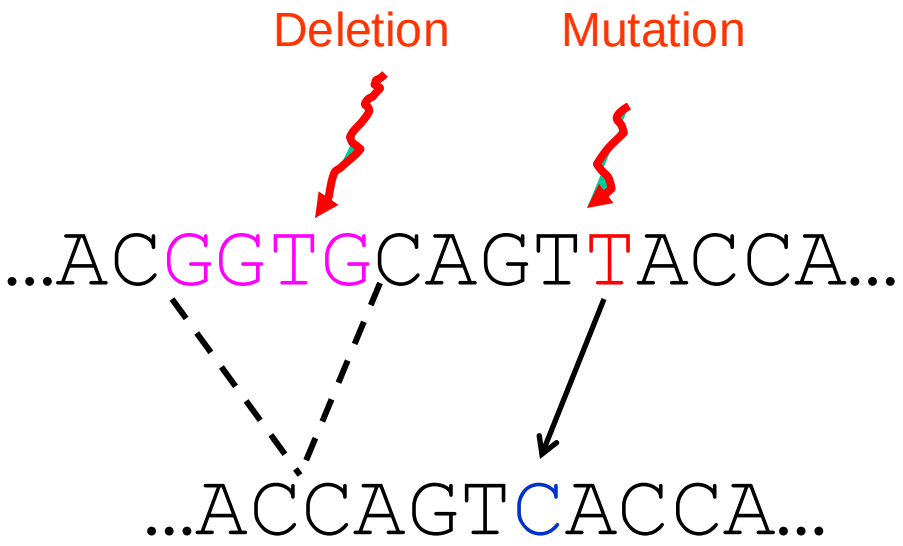
Novel techniques needed for scalability and accuracy

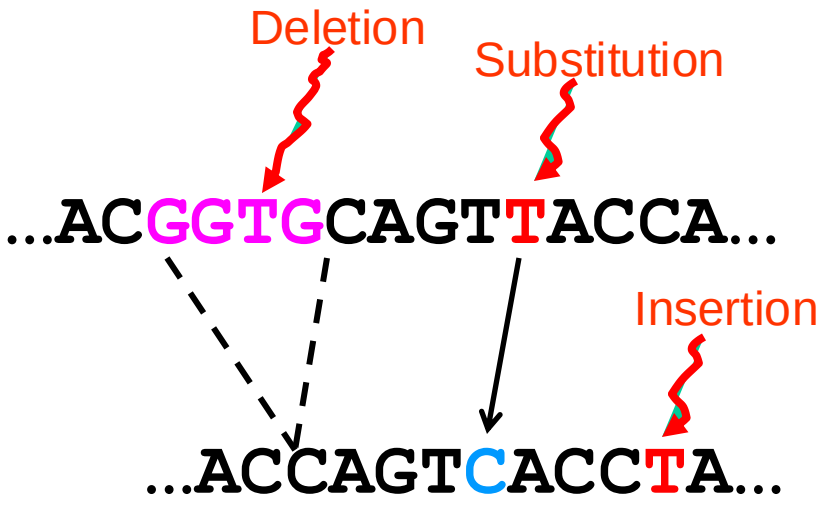
NP-hard problems and large datasets
Current methods do not provide good accuracy
Few methods can analyze even moderately large datasets

Many important applications besides phylogenetic estimation

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

Indels (insertions and deletions)





...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

What makes MSA difficult?

- Large numbers of sequences
- High evolutionary rates
- Sequence length heterogeneity (e.g., fragmentary sequences)
- Very long sequences (e.g., genome-scale)
- Rearrangement events

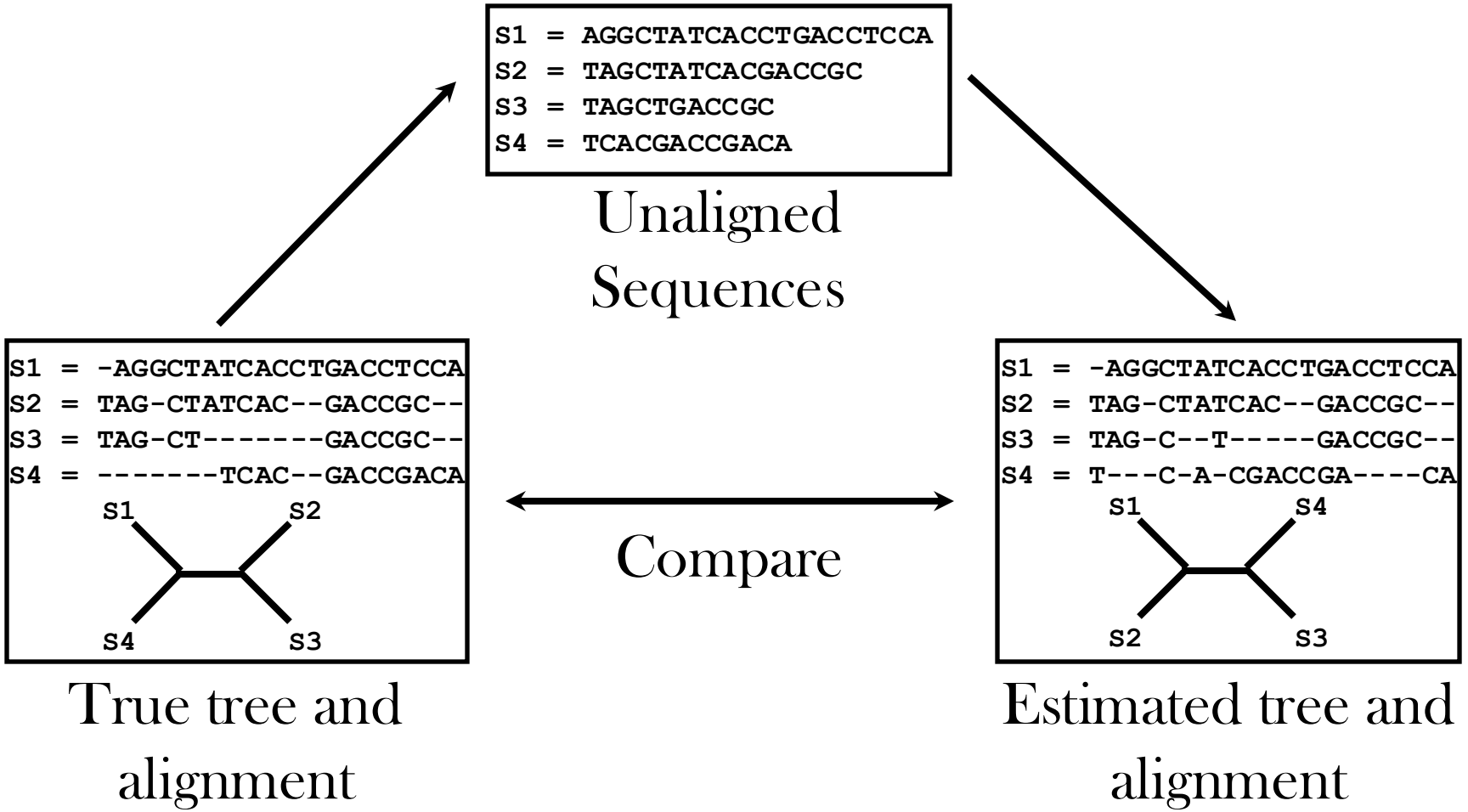
What makes MSA difficult?

- Large numbers of sequences
- High evolutionary rates
- Sequence length heterogeneity (e.g., fragmentary sequences)
- Very long sequences (e.g., genome-scale)
- Rearrangement events

This talk

- Multiple Sequence Alignment (MSA): challenges and progress
- MSA using ensembles of HMMs
- Applications of eHMMs
- Statistical alignment (e.g., BAli-Phy)

Simulation Studies



Two-phase estimation

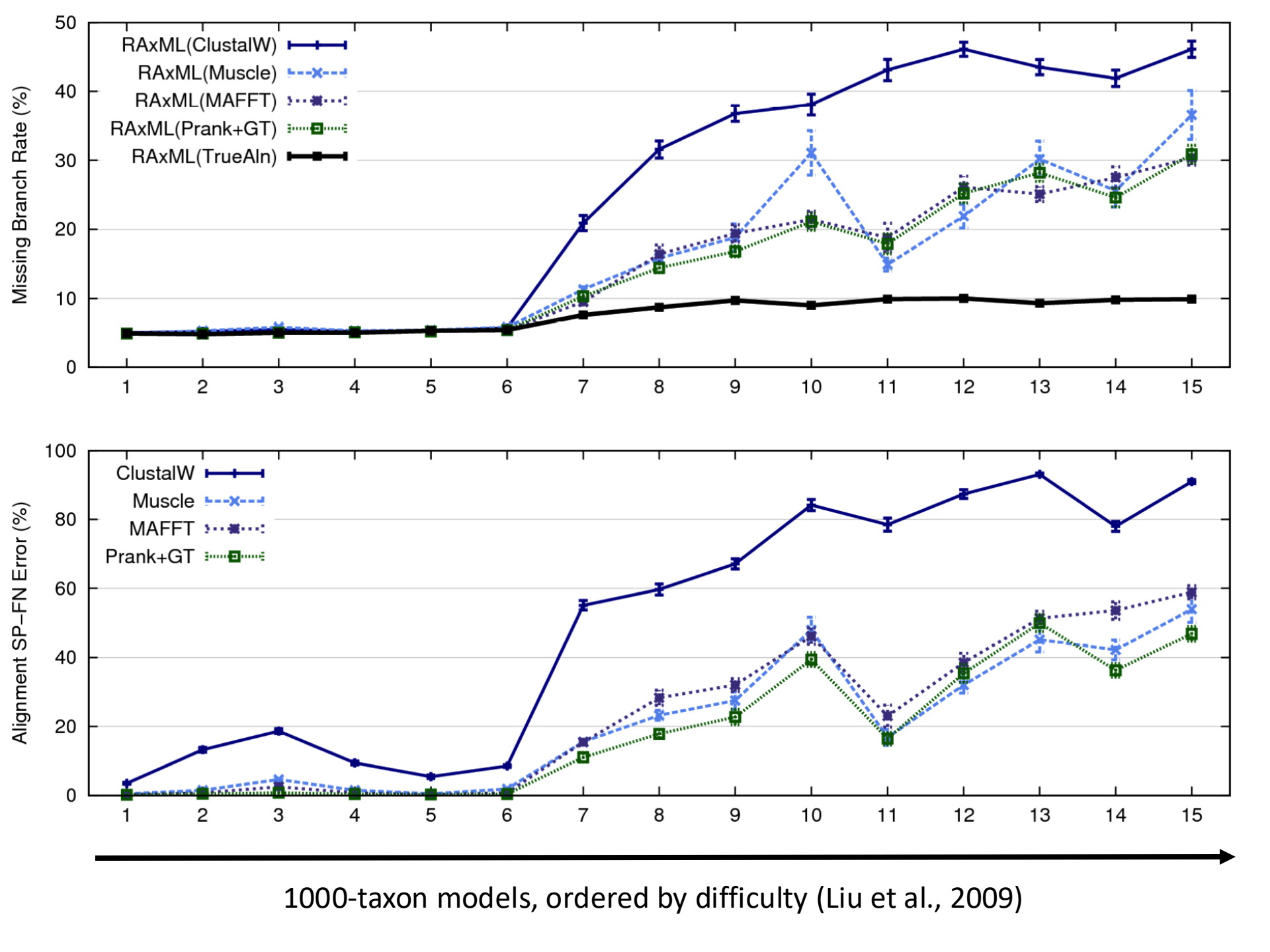
Alignment methods

- 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.

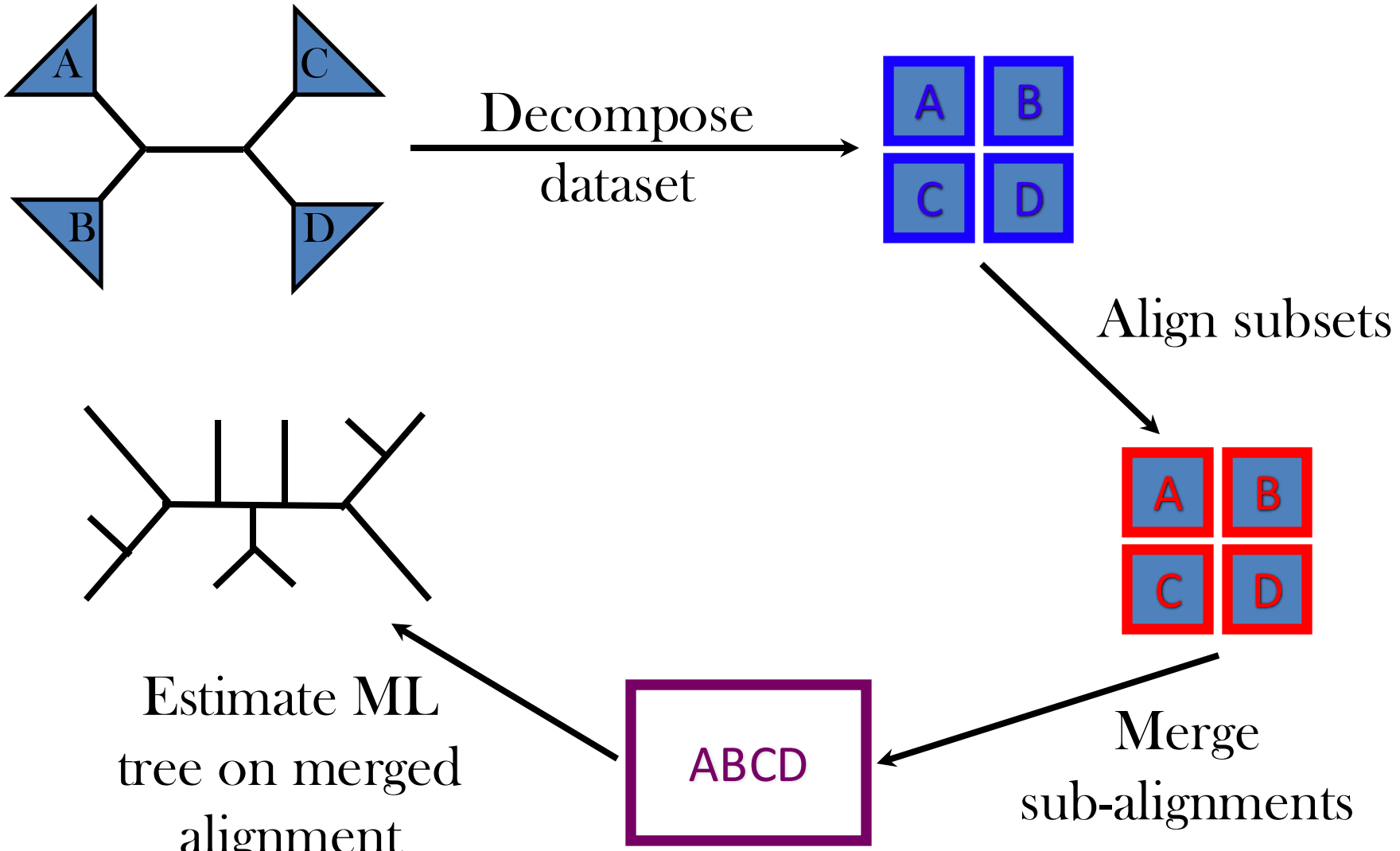
Phylogeny methods

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

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

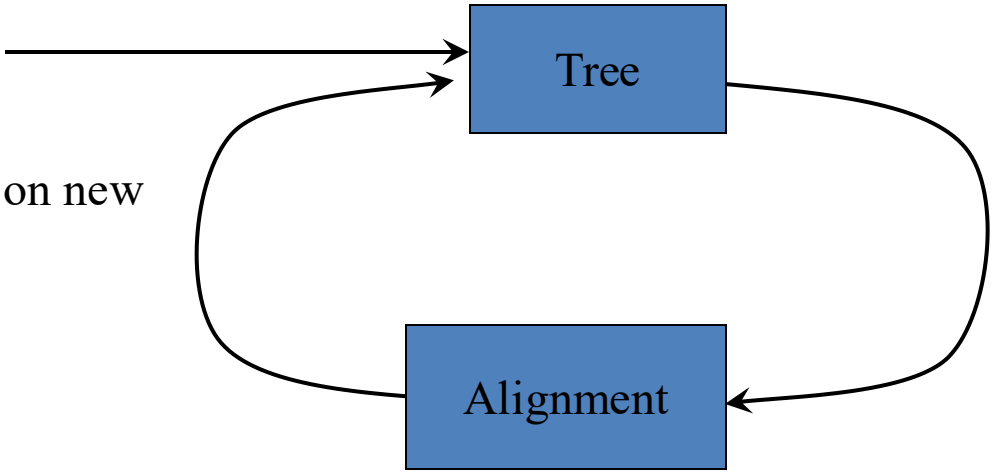

RAxML(ClustalW)
RAxML(Muscle)
RAxML(MAFFT)
RAxML(Prank+GT)
RAxML(TrueAln)
Missing Branch Rate (%)
0
10
20
30
40
50
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
ClustalW
Muscle
MAFFT
Prank+GT
Alignment SP–FN Error (%)
0
20
40
60
80
100
1000-taxon models, ordered by difficulty (Liu et al., 2009)

Re-aligning on a tree



SATé (Liu et al., Science 2009)

Obtain initial alignment and estimated ML tree



Estimate ML tree on new alignment

Use tree to compute new alignment

Repeat until termination condition, and

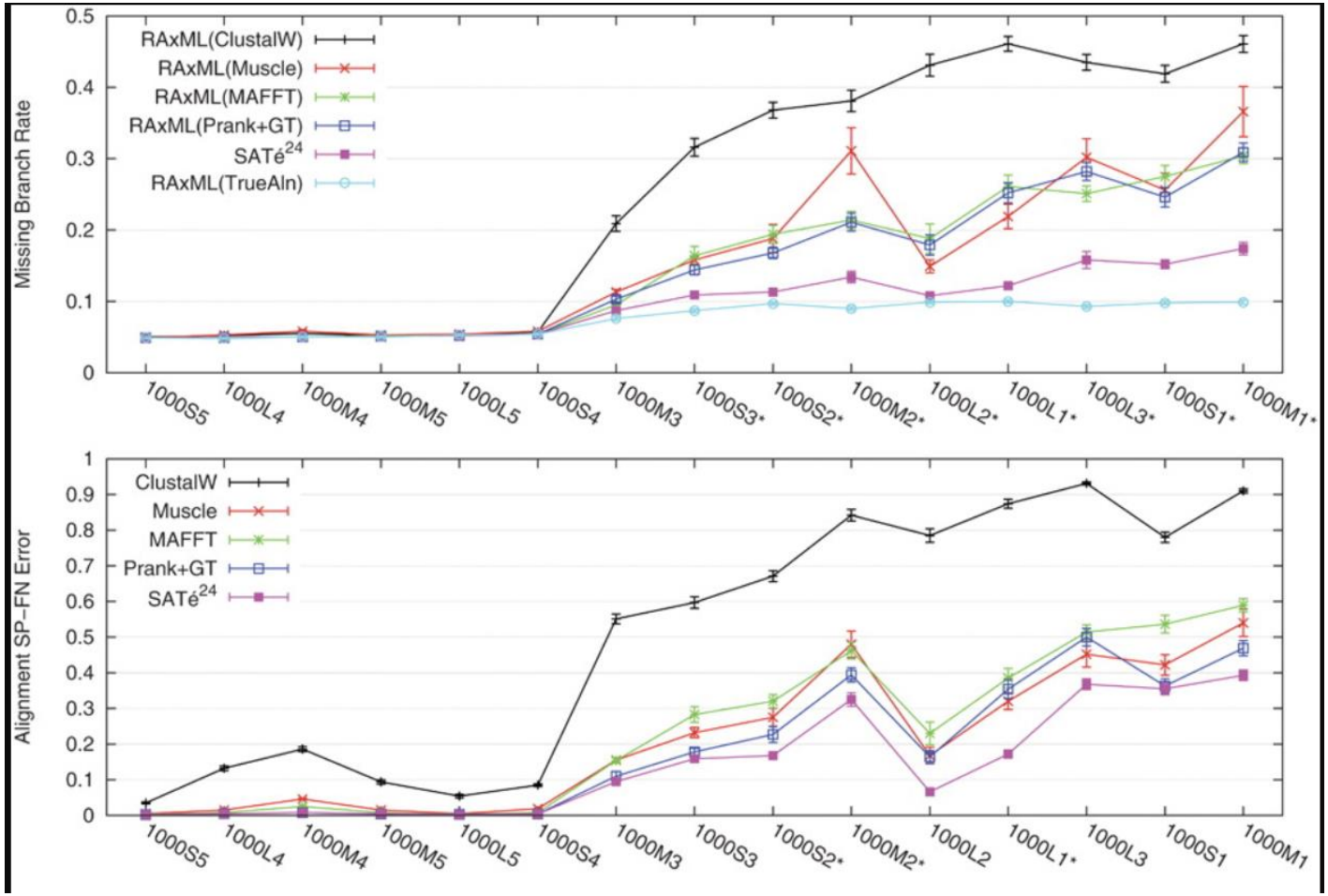
return the alignment/tree pair with the best ML score

REPORT

Rapid and Accurate Large-Scale Coestimation of Sequence Alignments and Phylogenetic Trees

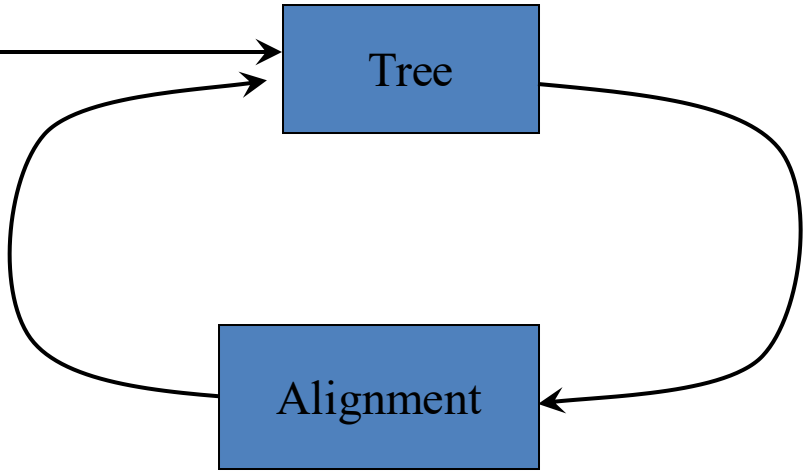
KEVIN LIU, SINDHU RAGHAVAN, SERITA NELESEN, C. RANDAL LINDER, AND TANDY WARNOW Authors Info & Affiliations

SCIENCE • 19 Jun 2009 • Vol 324, Issue 5934 • pp. 1561-1564 • DOI: 10.1126/science.1171243



SATé, PASTA, and MAGUS Algorithms

Obtain initial alignment and estimated ML tree



Estimate ML tree on new alignment

Use tree to compute new alignment

Repeat until termination condition, and

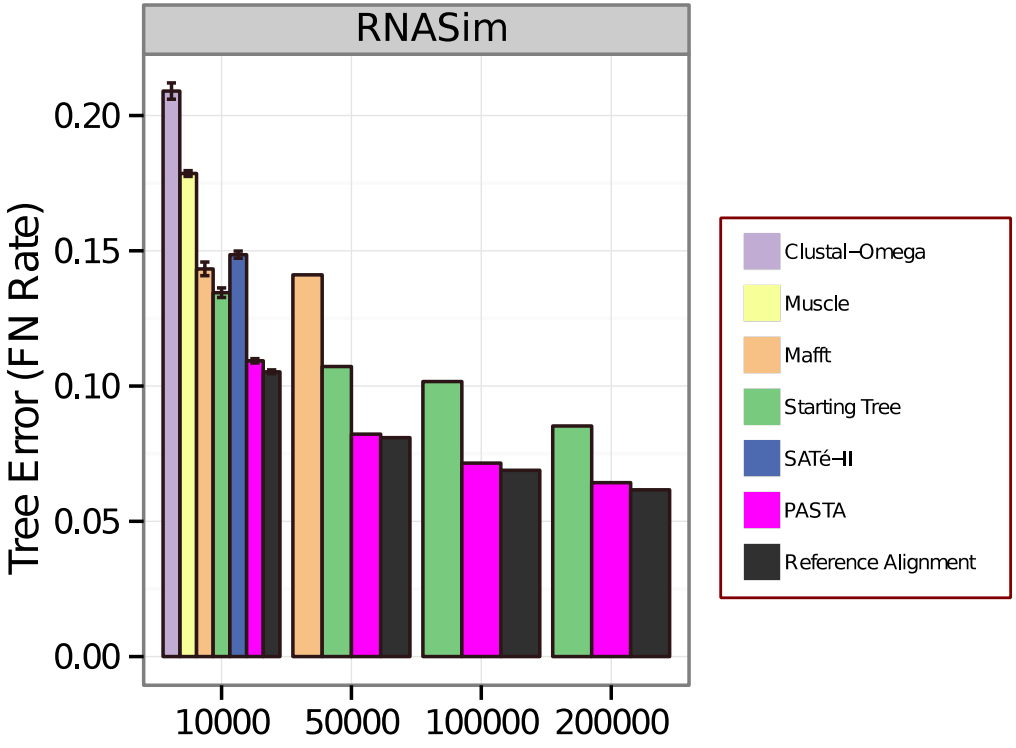
return the alignment/tree pair with the best ML score

Improvement over time

- SATe-1 (Science 2009): up to about 8,000
- SATe-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)

Each method improved on the previous with respect to accuracy, speed, and scalability

Tree accuracy



1 million sequences:

- PASTA finished one iteration in 15 days
- PASTA tree had 6% error, compared to 5.6% when using true alignment
- Starting tree had 8.4% error

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

1kp: Thousand Transcriptome Project

G. Ka-Shu Wong
U Alberta

J. Leebens-Mack
U Georgia

N. Wickett
Northwestern

N. Matasci
iPlant

T. Warnow,
UIUC

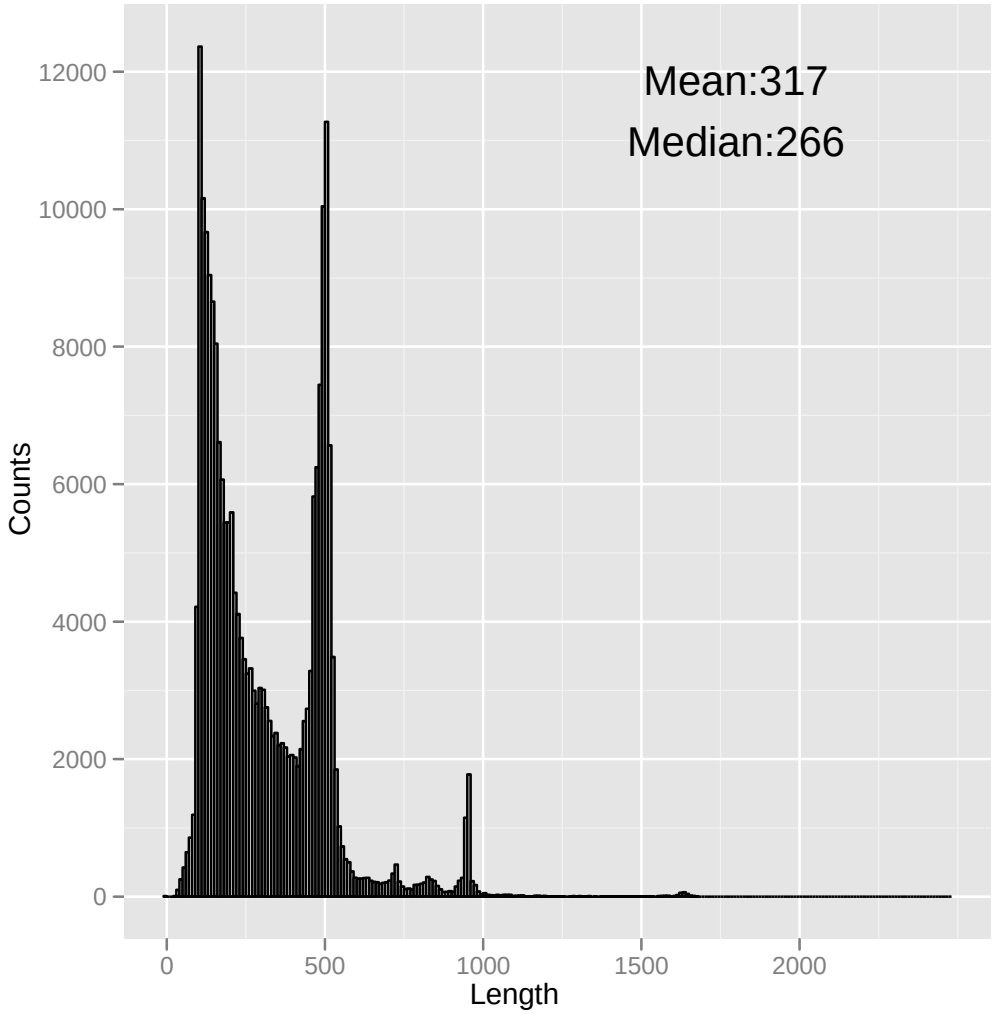
S. Mirarab,
UT-Austin

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.

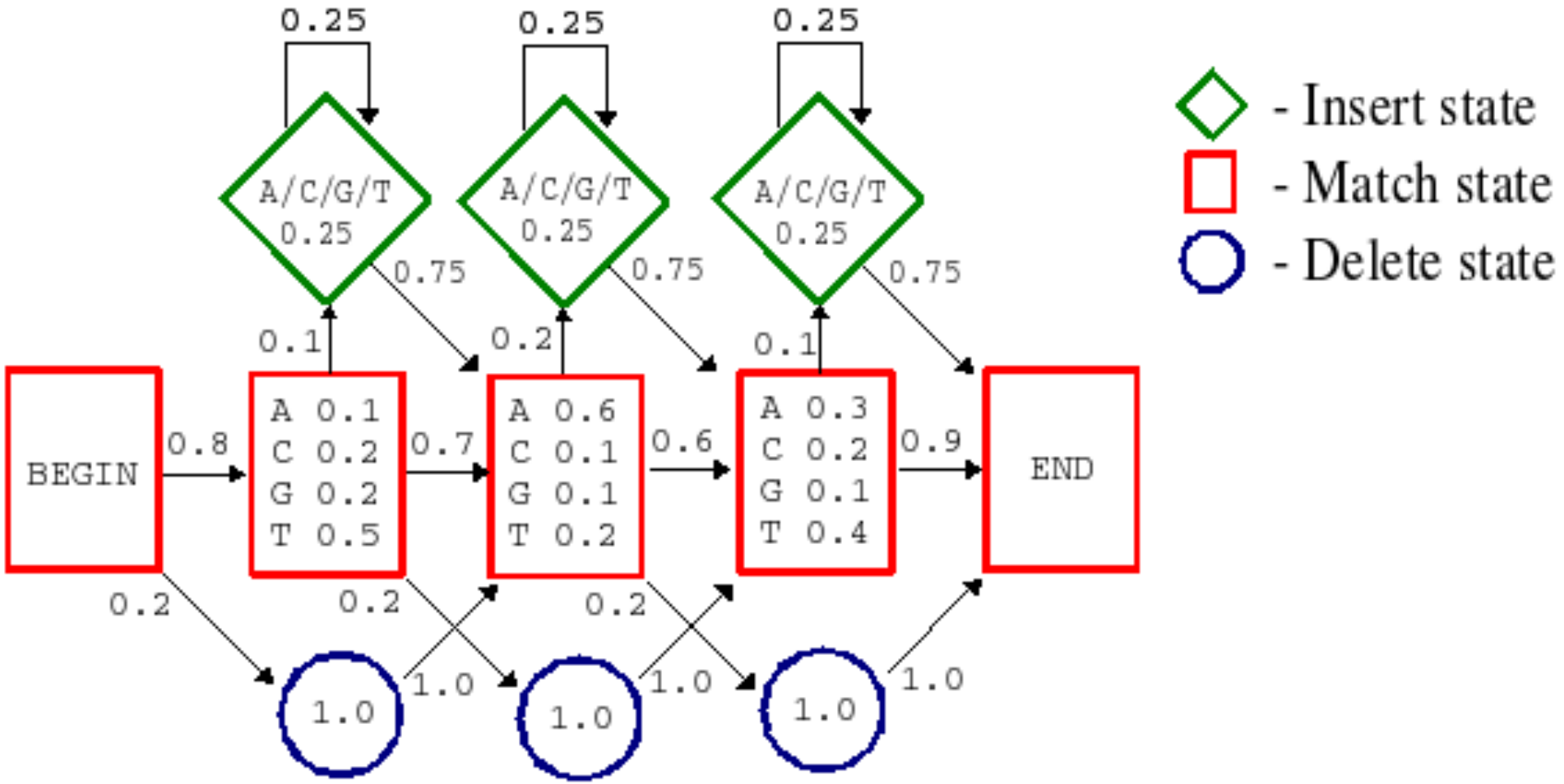
This talk

- Multiple Sequence Alignment (MSA): challenges and progress
- MSA using ensembles of HMMs
- Applications of eHMMs
- Statistical alignment (e.g., BAli-Phy)

Profile HMMs

- Generative model for representing a MSA
- Consists of:
  - Set of states (Match, insertion, and deletion)
  - Transition probabilities
  - Emission probabilities

Profile Hidden Markov Model for DNA sequence alignment



HMMs for MSA

- Given seed alignment (e.g., in PFAM) and a collection of sequences for the protein family:
  - Represent seed alignment using HMM
  - Align each additional sequence to the HMM
  - Use transitivity to obtain MSA
-

HMMs for MSA

- Given seed alignment (e.g., in PFAM) and a collection of sequences for the protein family:
  - Represent seed alignment using HMM
  - Align each additional sequence to the HMM
  - Use transitivity to obtain MSA
- Can we do something like this without a seed alignment?

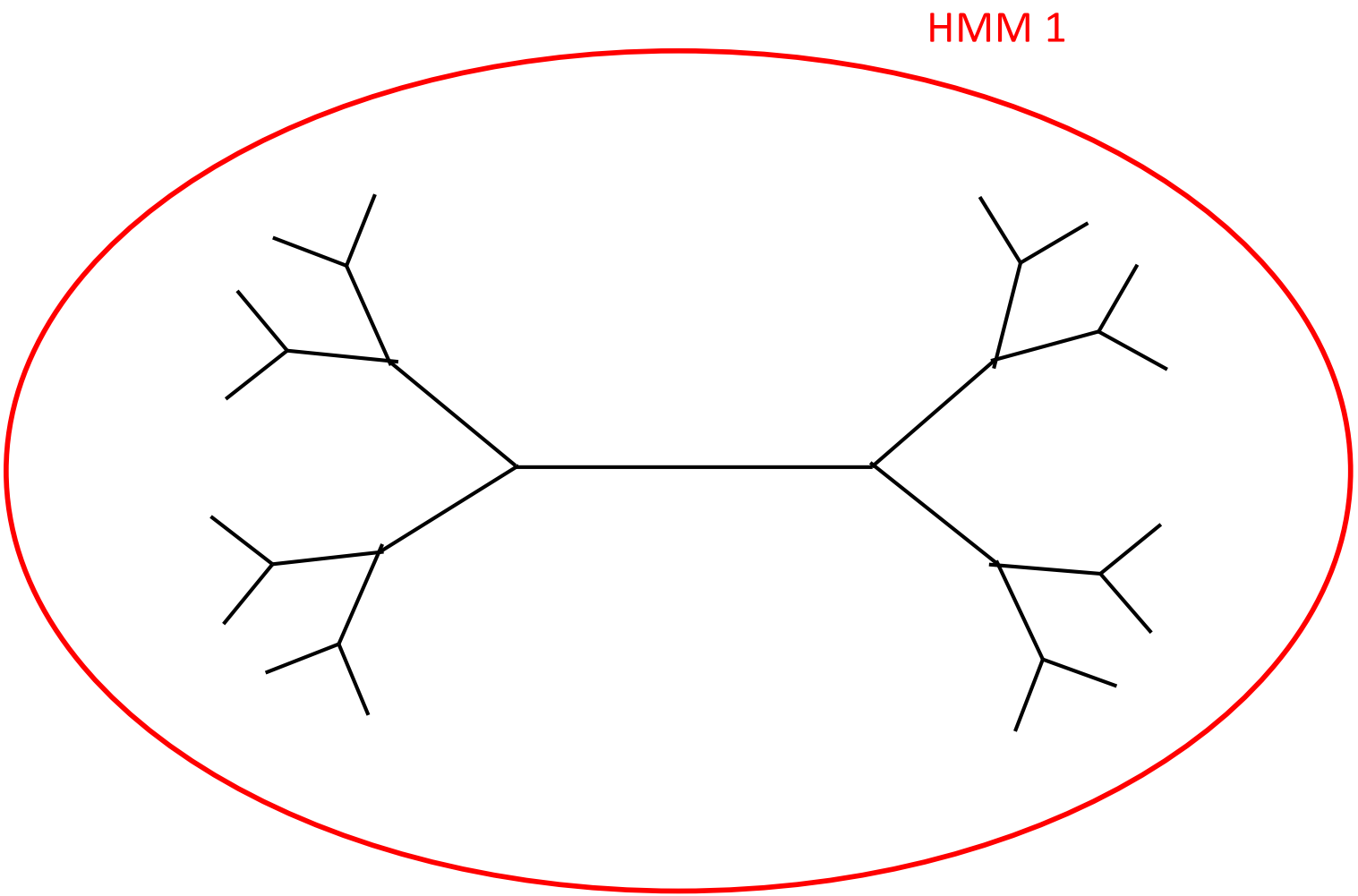
Simple idea (not UPP)

- Select random subset of sequences, and build "backbone alignment"
- Construct a Hidden Markov Model (HMM) on the backbone alignment
- Add all remaining sequences to the backbone alignment using the HMM

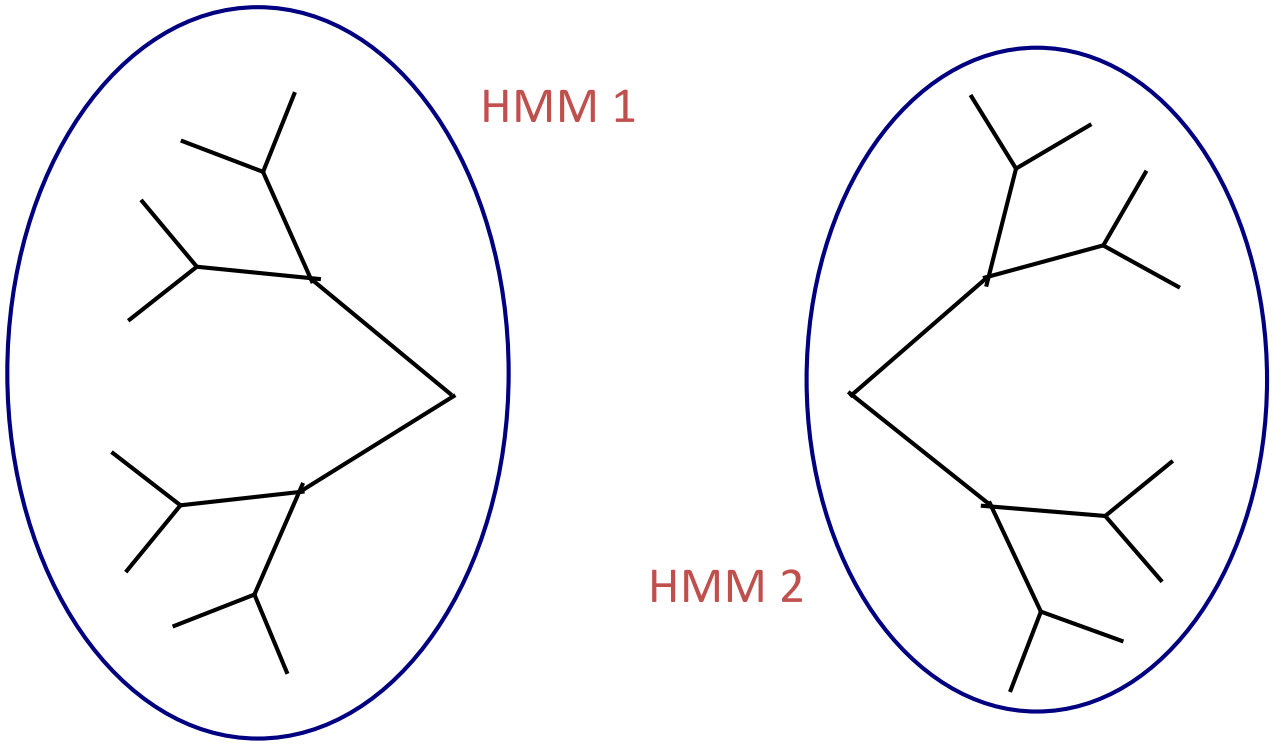
This approach works well if the dataset is small and has low evolutionary rates, but is not very accurate otherwise.

- Select random subset of sequences, and build "backbone alignment"
- Construct a Hidden Markov Model (HMM) on the backbone alignment
- Add all remaining sequences to the backbone alignment using the HMM

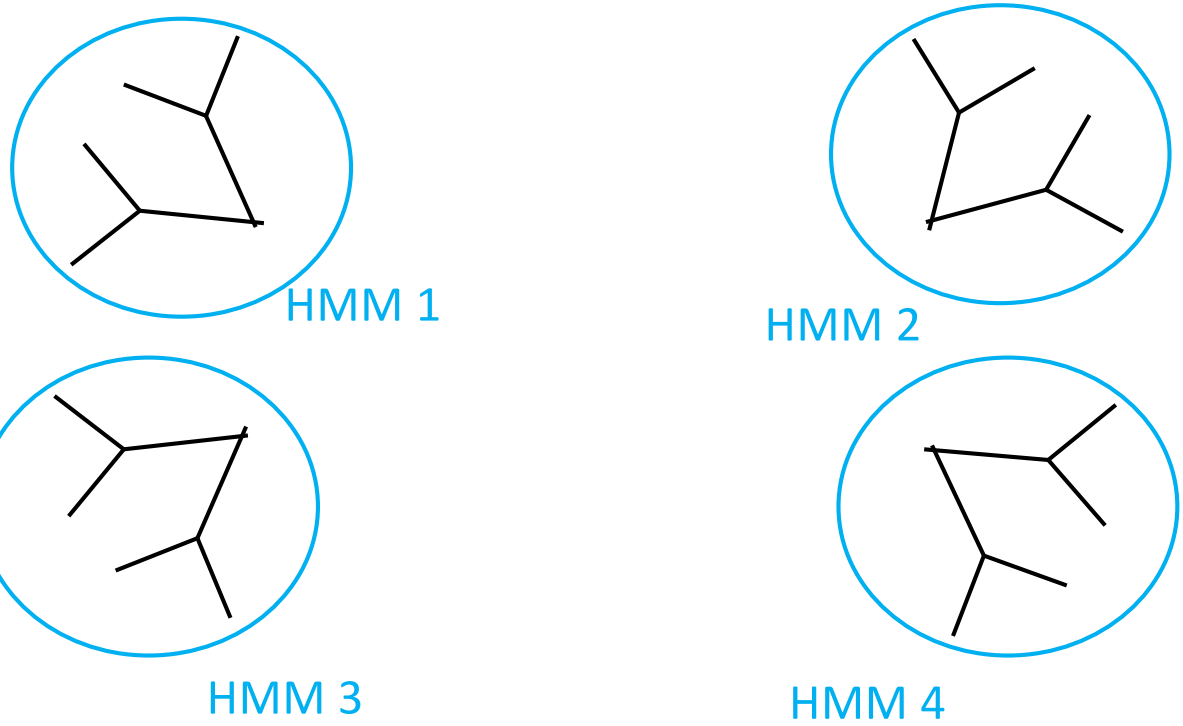
One Hidden Markov Model for the backbone alignment?



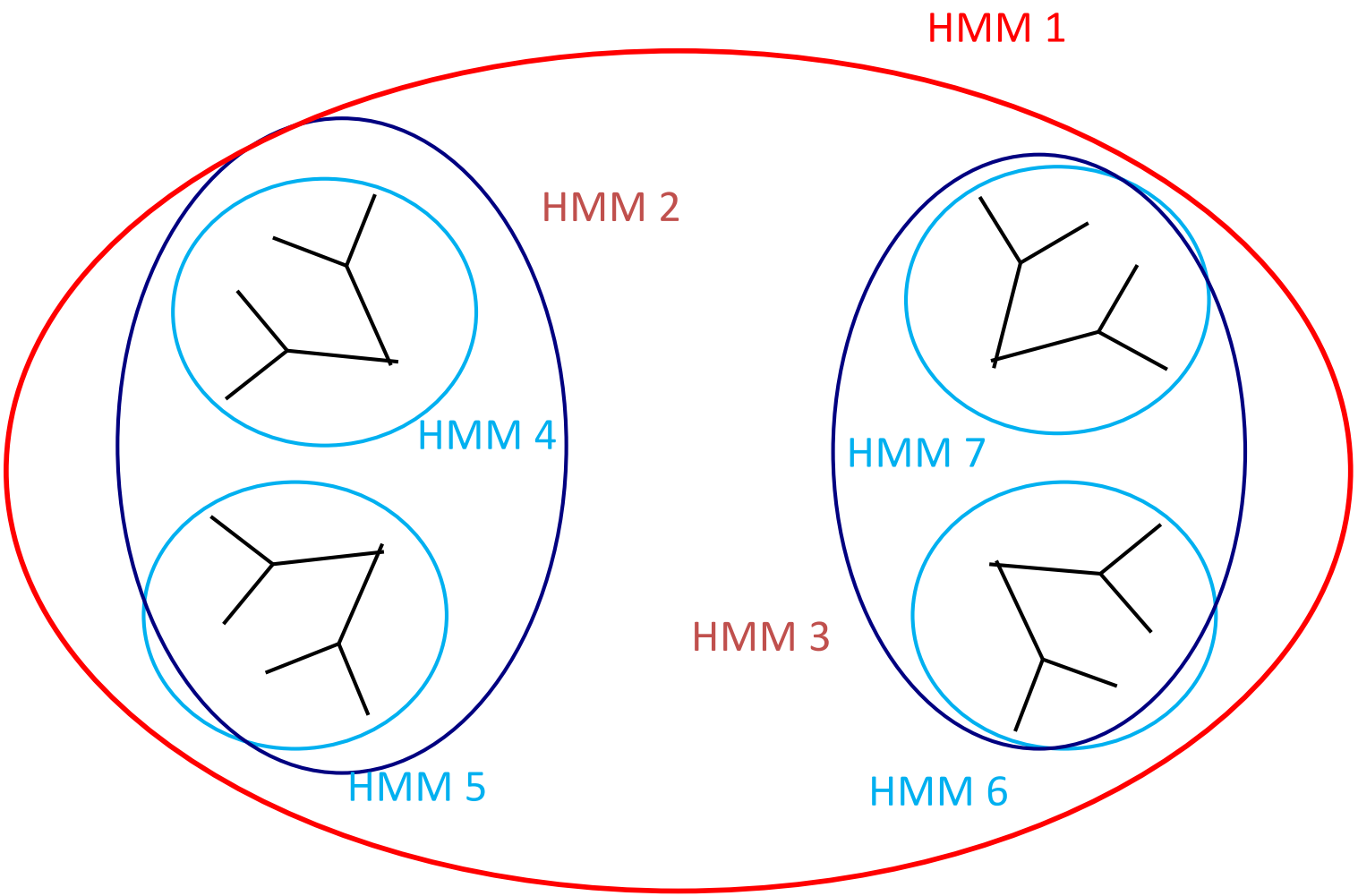
Or 2 HMMs?



Or 4 HMMs?



Or all 7 HMMs?



UPP Algorithmic Approach

1. Select 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

UPP

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

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

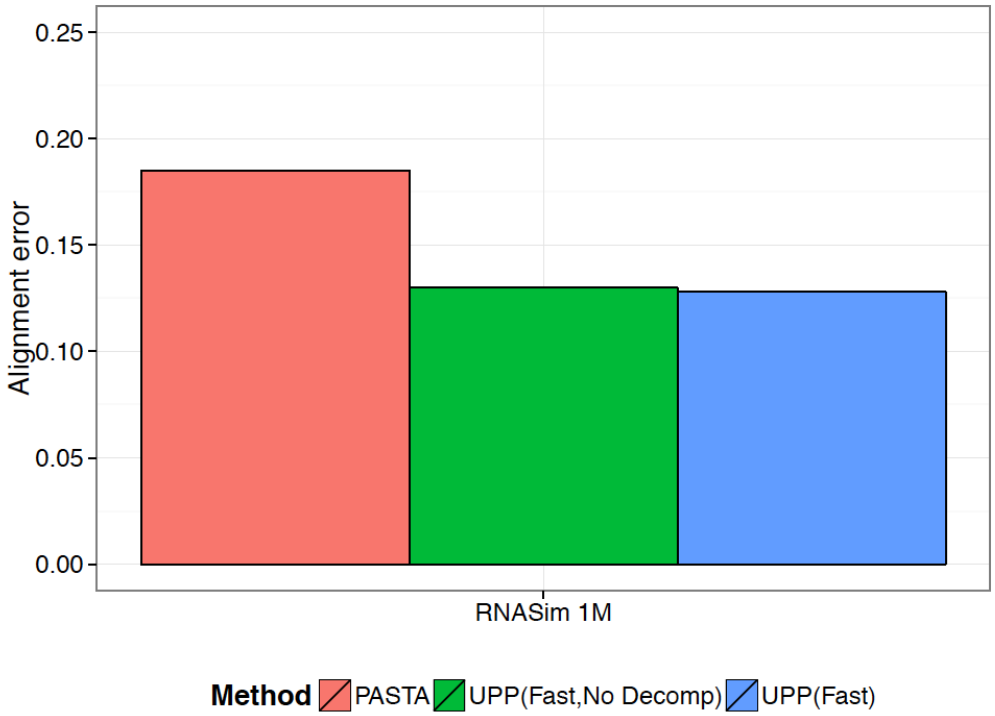
Purpose: highly accurate large-scale multiple sequence alignments, even in the presence of fragmentary sequences.

Uses an ensemble of HMMs

Evaluation

- Simulated datasets (some have fragmentary sequences):
  - 10K to 1,000,000 sequences in RNASim – complex RNA sequence evolution simulation
  - 1000-sequence nucleotide datasets from SATé papers
  - 5000-sequence AA datasets (from FastTree paper)
  - 10,000-sequence Indelible nucleotide simulation
- Biological datasets:
  - Proteins: largest BaliBASE and HomFam
  - RNA: 3 CRW datasets up to 28,000 sequences

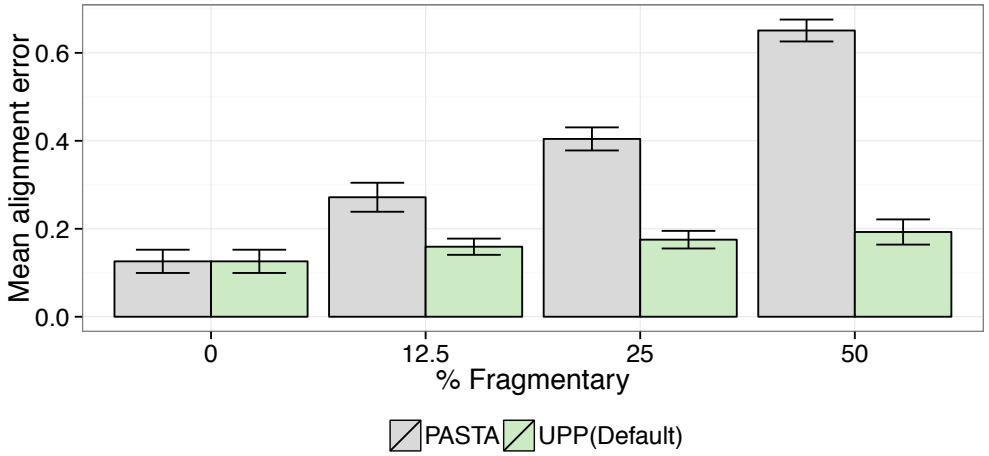
RNASim Million Sequences: alignment error



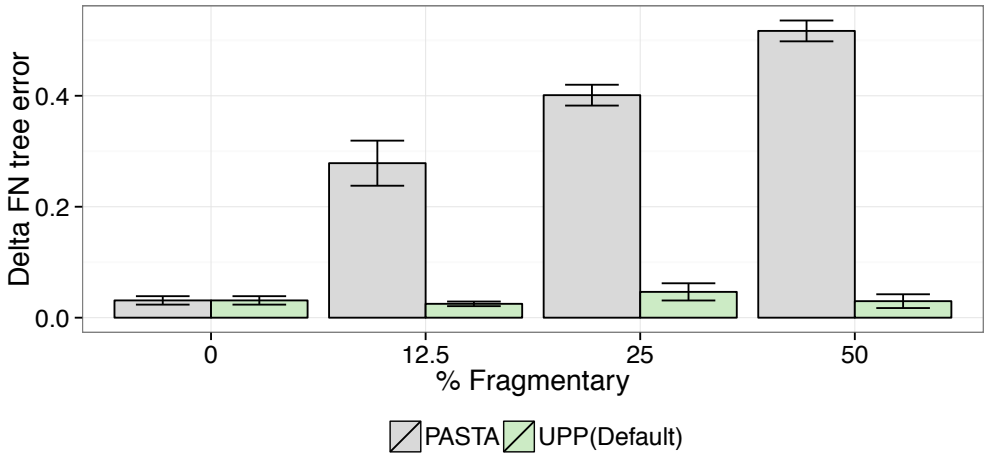
Notes:

- 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

UPP is very robust to fragmentary sequences



(a) Average alignment error



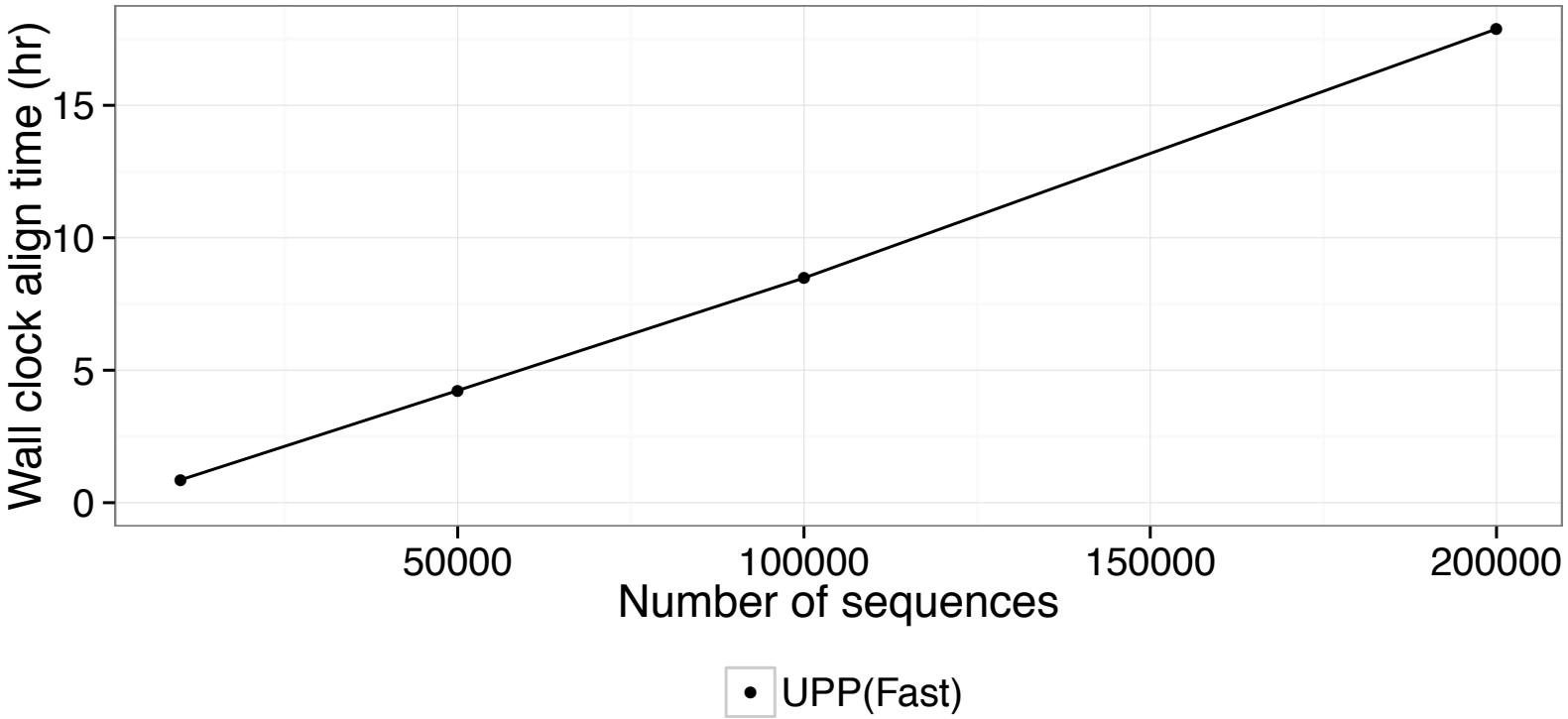
(b) Average tree error

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

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

UPP Running Time



Wall-clock time used (in hours) given 12 processors

Improvements on UPP

- WITCH (Shen, Park, and Warnow, J Comp Bio 2022) and WITCH-ng (Liu and Warnow, Bioinf Adv 2023) improve on UPP by combining alignments from different HMMs in the eHMM
- UPP2 (Park et al., Bioinformatics 2023): faster version of UPP
- HMMerge (Park and Warnow, 2023): creates a new profile HMM that combines the HMMs in the eHMM

Other applications of eHMMs

- Updating phylogenies: Phylogenetic Placement
- Microbiome analysis: Taxon identification and Abundance Profiling
- Protein classification

Phylogenetic Placement

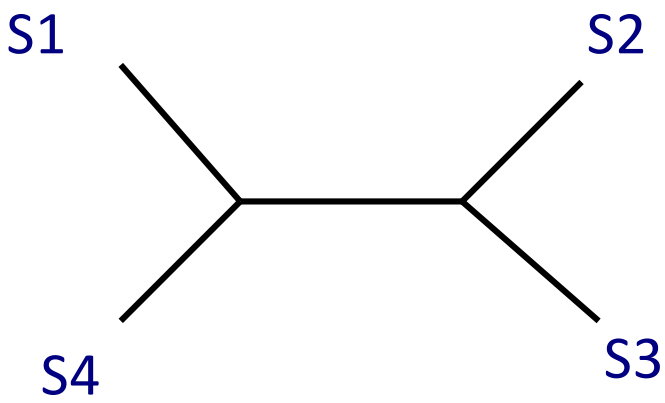
Input: Backbone alignment and tree on full-length sequences, and a set of homologous query sequences (e.g., reads in a metagenomic sample for the same gene)

Output: Placement of query sequences on backbone tree

Applications: Updating existing tree, microbiome analysis

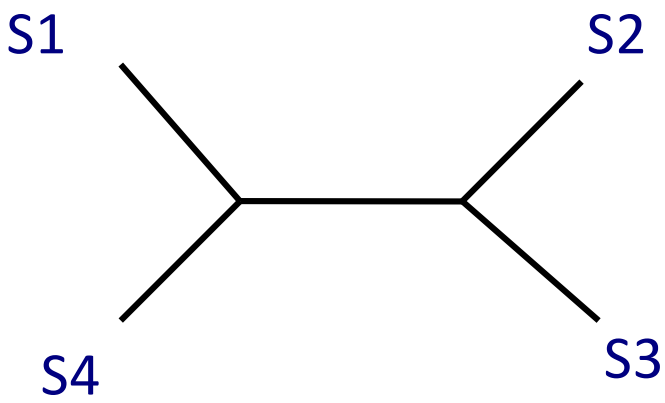
Input

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



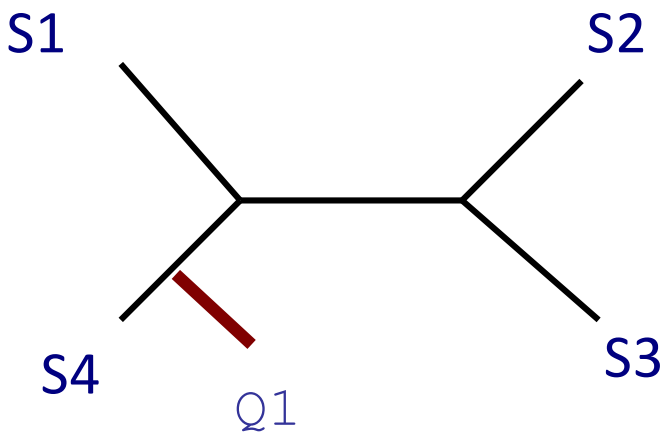
Align Sequence

```
S1  = -AGGCTATCACCTGACCTCCA-AA
S2  = TAG-CTATCAC--GACCGC--GCA
S3  = TAG-CT-------GACCGC--GCT
S4  = TAC----TCAC--GACCGACAGCT
Q1  = -------T-A--AAAC--------
```

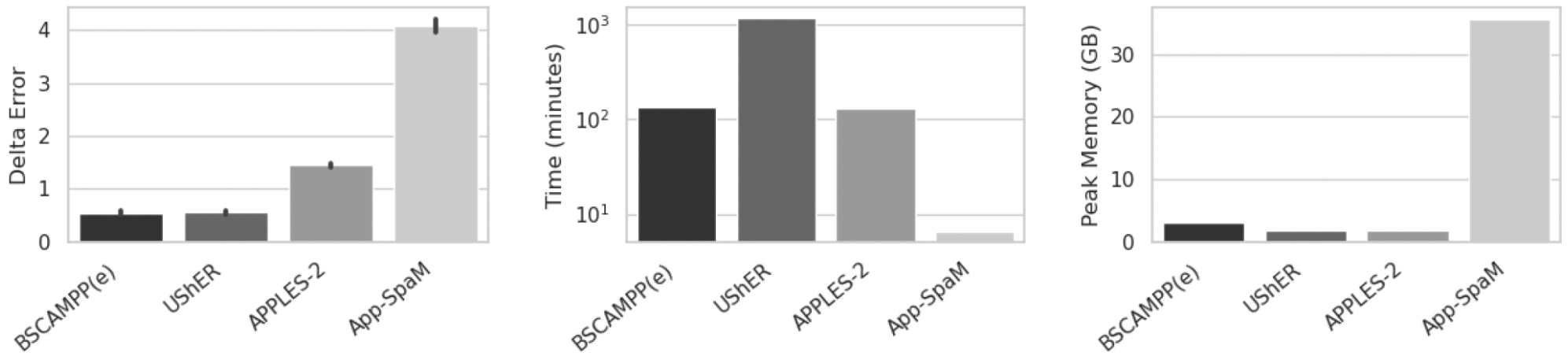


Place Sequence

```
S1   = -AGGCTATCACCTGACCTCCA-AA
S2   = TAG-CTATCAC--GACCGC--GCA
S3   = TAG-CT-------GACCGC--GCT
S4   = TAC----TCAC--GACCGACAGCT
Q1   = -------T-A--AAAC--------
```



Phylogenetic Placement using BSCAMPP(e)



(b) RNAsim: 10,000 Pacbio reads placed into a 50,000 leaf tree

- Batch-SCAMPP(e) (under review) aligns query sequences using UPP, then places into backbone tree using Batch-SCAMPP with EPA-ng (maximum likelihood placement method).
- UShER uses the same alignment, but does a maximum parsimony placement
- APPLES-2 uses the same alignment but places using distances
- App-SpaM is alignment-free

Phylogenetic Placement: BSCAMPP, UShER, APPLES-2, and App-SpaM

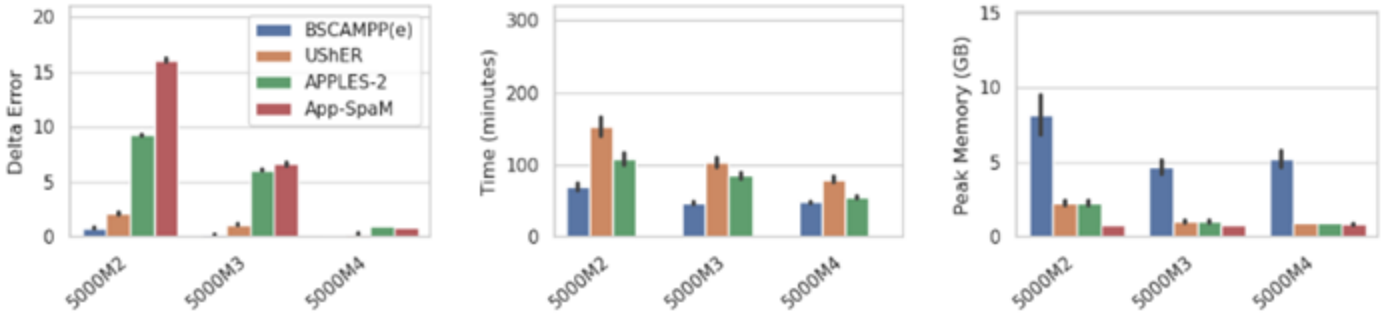
Relative rate of evolution:

- 5000M2 high
- 5000M3 med
- 5000M4 low

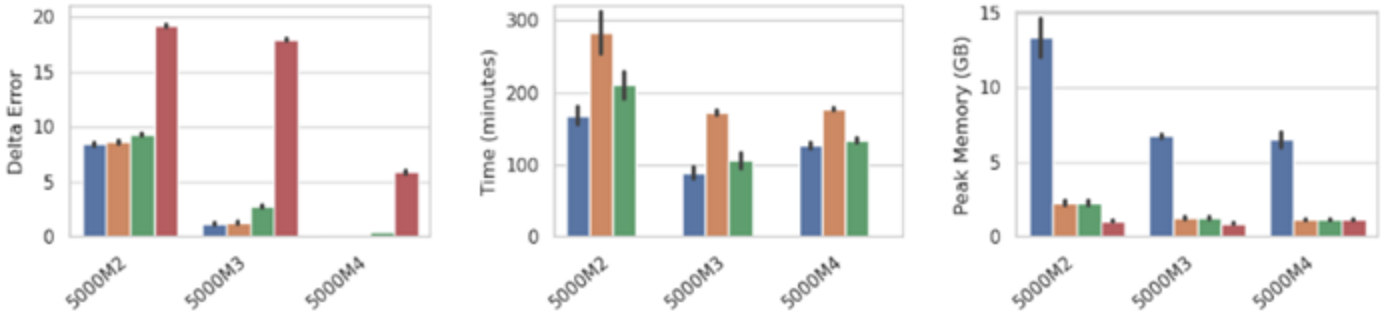
UPP alignment for all but App-SpaM

BSCAMPP clearly most accurate when rate of evolution is high

BSCAMPP faster than APPLES-2 and UShER on all conditions.



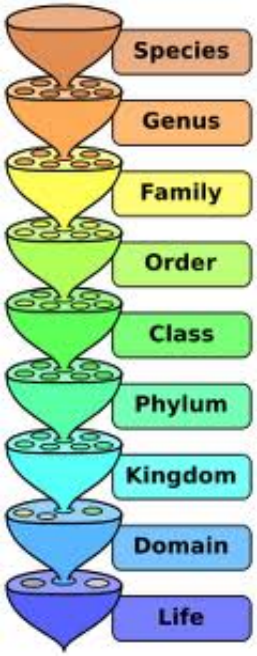
(a) 10,000 Illumina-style reads into 4,000-leaf tree

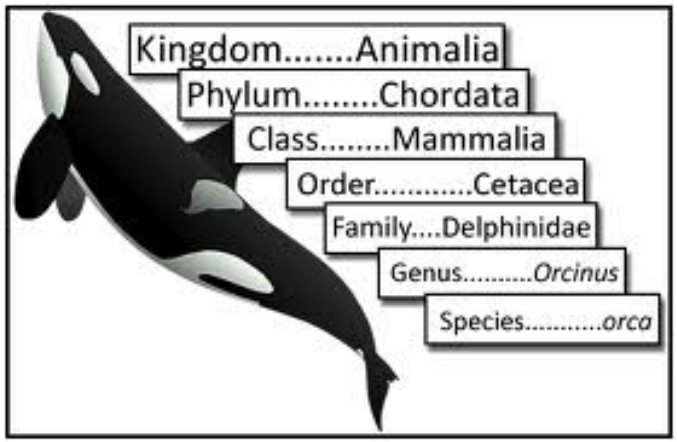


(b) 10,000 Pacbio-style reads into 4,000-leaf tree

Metagenomic Taxon Identification

Objective: classify short reads in a metagenomic sample

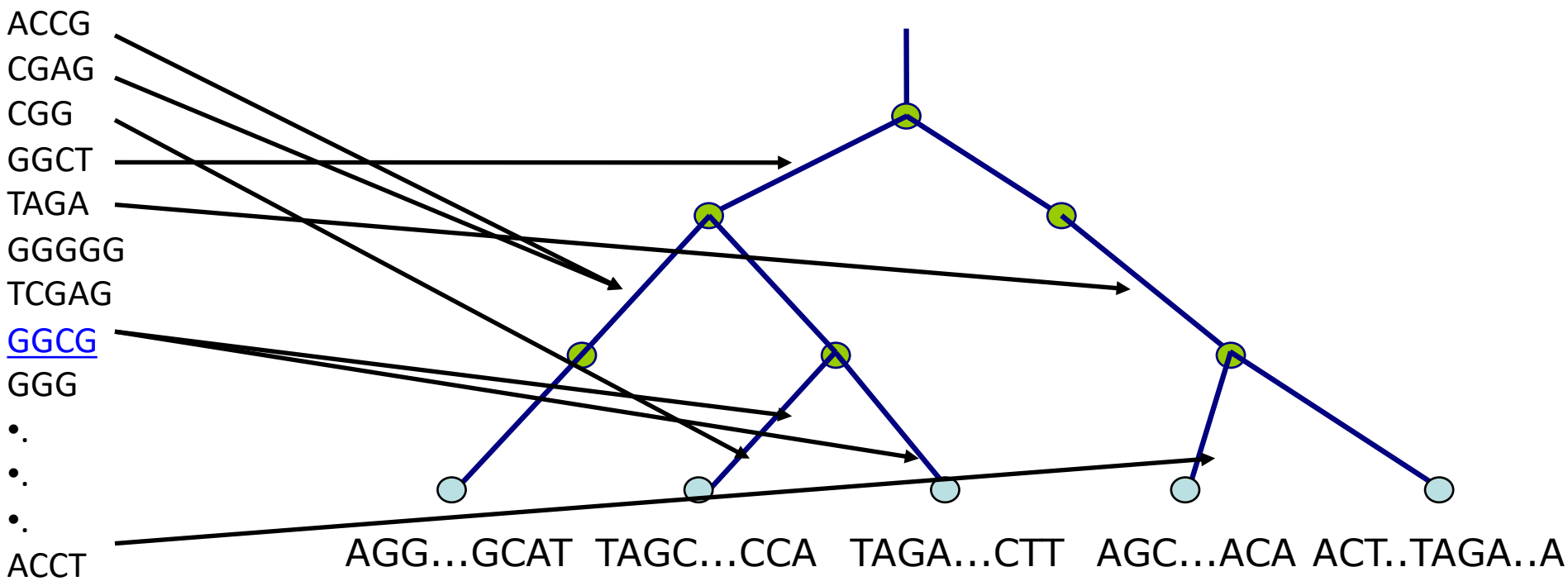




Marker-based Taxon Identification

Fragmentary sequences from some gene

Full-length sequences for same gene, and an alignment and a tree



TIPP family of methods

Can be used for taxon identification or abundance profiling

Basic approach:

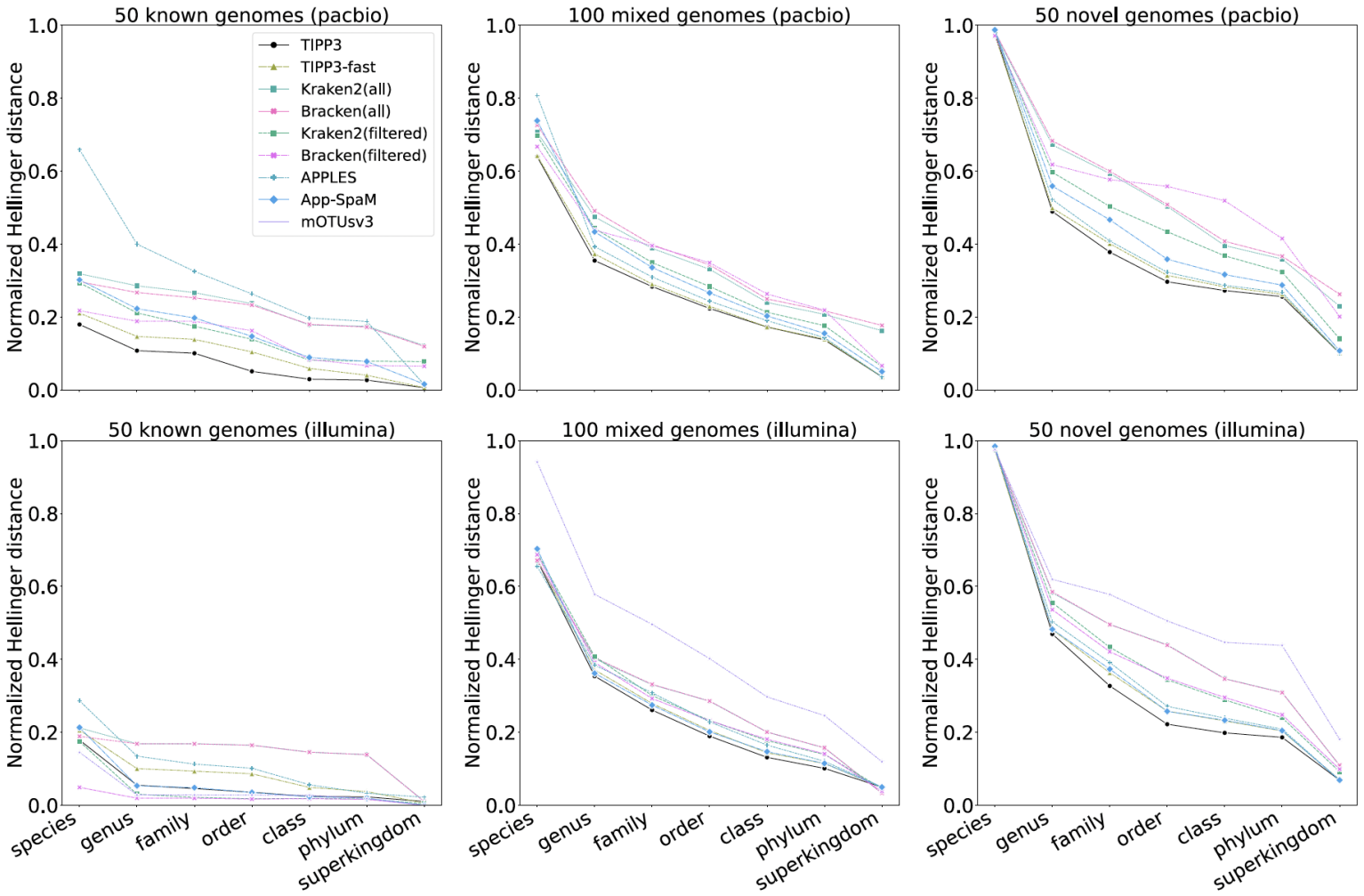
- Construct “library”: MSA and taxonomy for each marker gene (single-copy, universal)

Given set of reads from an environmental sample:

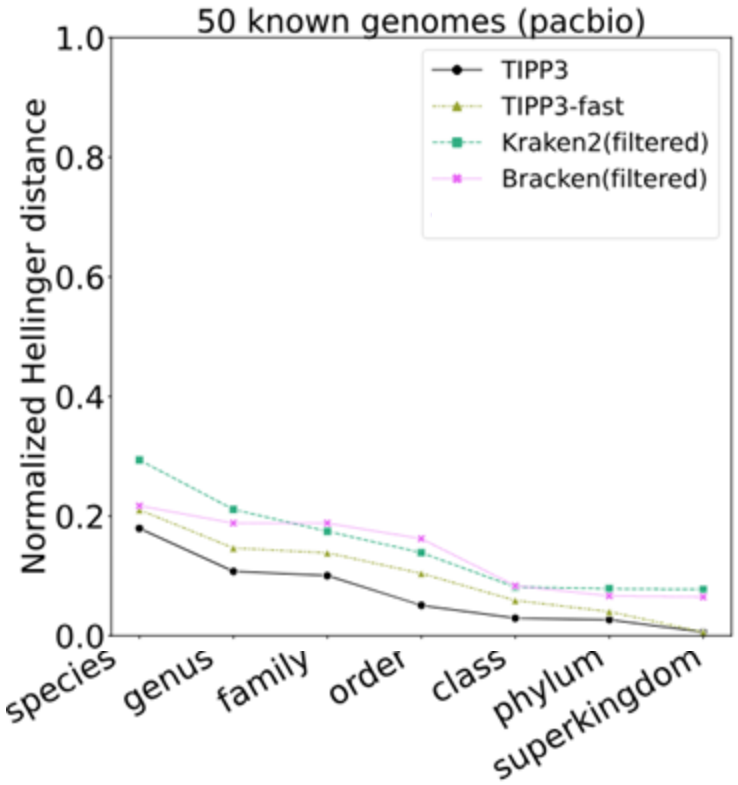
- Match each read to a marker gene
- For each marker gene,
  - Add assigned reads to the marker gene MSA (UPP or WITCH)
  - Place reads into marker gene taxonomy using Maximum Likelihood (pplacer, SCAMPP, BSCAMPP) to meet desired confidence threshold (95% is default)
  - Obtain taxonomic assignment for each read
- Combine all taxonomic assignments across all the reads for abundance profiling

TIPP: Bioinformatics 2014
TIPP2: Bioinformatics 2021
TIPP3: under development

Abundance Profiling using TIPP3



Abundance Profiling using TIPP3, TIPP3-fast, Kraken2, and Bracken



- marker gene-based methods: TIPP3/TIPP3-fast
- Kmer-based methods: Kraken2/Bracken

- When input reads have high sequencing error
  - Marker gene-based methods have more accurate abundance profile

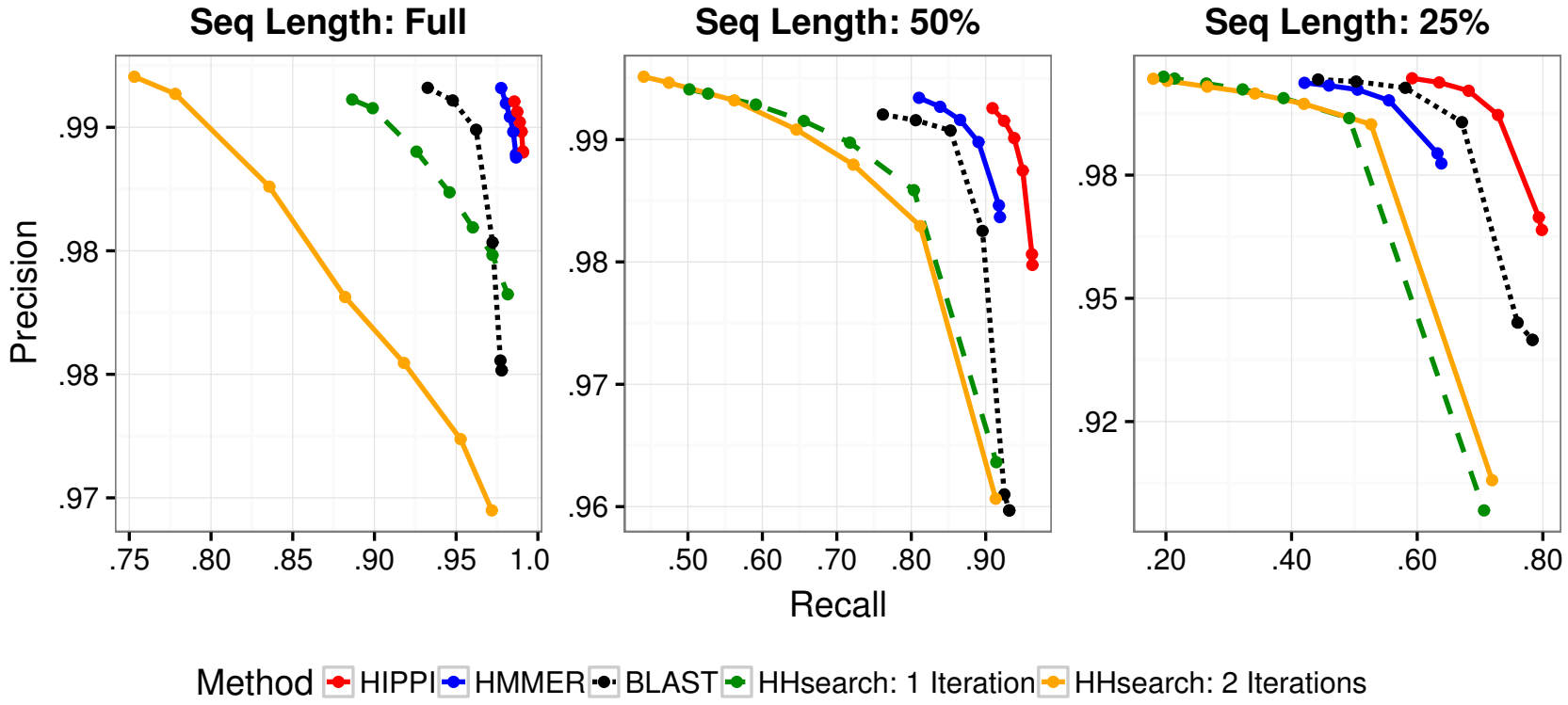
Protein Family Assignment

- Input: new AA sequence (might be fragmentary) and database of protein families (e.g., PFAM)
- Output: assignment (if justified) of the sequence to an existing family in the database

HIPPI

- HIerarchical Profile HMMs for Protein family Identification
- Nguyen, Nute, Mirarab, and Warnow, RECOMB-CG and BMC-Genomics 2016
- Uses an ensemble of HMMs to classify protein sequences
- Tested on HMMER

HIPPI for protein classification



Nguyen et al., BMC Genomics 2016

Summary: eHMMs

An ensemble of HMMs provides a better model of a multiple sequence alignment than a single HMM, and is better able to

- detect homology between full length sequences and fragmentary sequences
- add fragmentary sequences into an existing alignment

especially when there are many indels and/or substitutions.

Summary

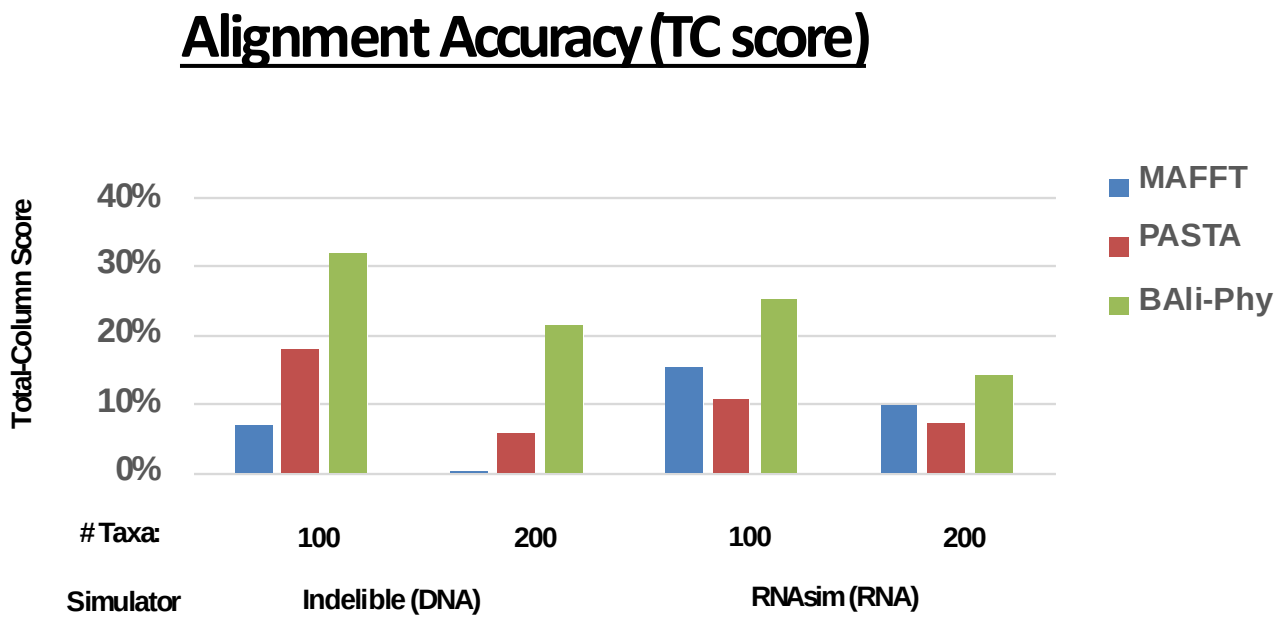
- Using an ensemble of HMMs tends to improve accuracy, for a cost of running time. Applications so far to taxonomic placement (SEPP), multiple sequence alignment (UPP), protein family classification (HIPPI). Improvements are mostly noticeable for large diverse datasets.
- Phylogenetically-based construction of the ensemble helps accuracy (note: the decompositions we produce are not clade-based), but the design and use of these ensembles is still in its infancy. (Many relatively similar approaches have been used by others, including FlowerPower by Sjolander)
- The basic idea can be used with any kind of probabilistic model, doesn't have to be restricted to profile HMMs.
- Basic question: why does it help?

What about Statistical Alignment?

BAli-Phy (Redelings and Suchard): leading method for statistical co-estimation of alignments and trees

Like Bayesian phylogeny estimation, it is expected to be the most rigorous and accurate technique for estimating trees and alignments!

BAli-Phy: Better than PASTA!



Simulated nucleotide datasets with 100 or 200 sequences (unpublished data from Mike Nute's PhD dissertation).

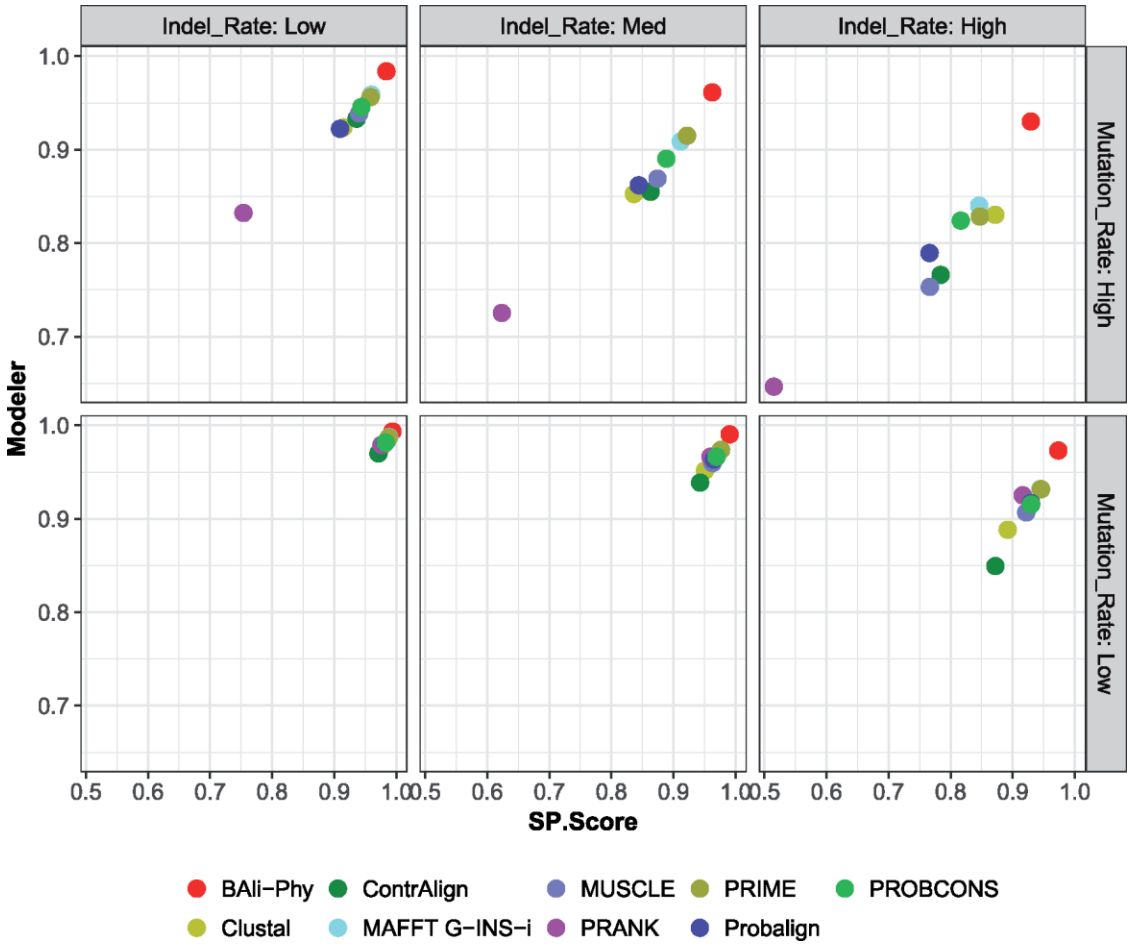
**Averages over 10 replicates*

But: BAli-Phy is limited to small datasets

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."

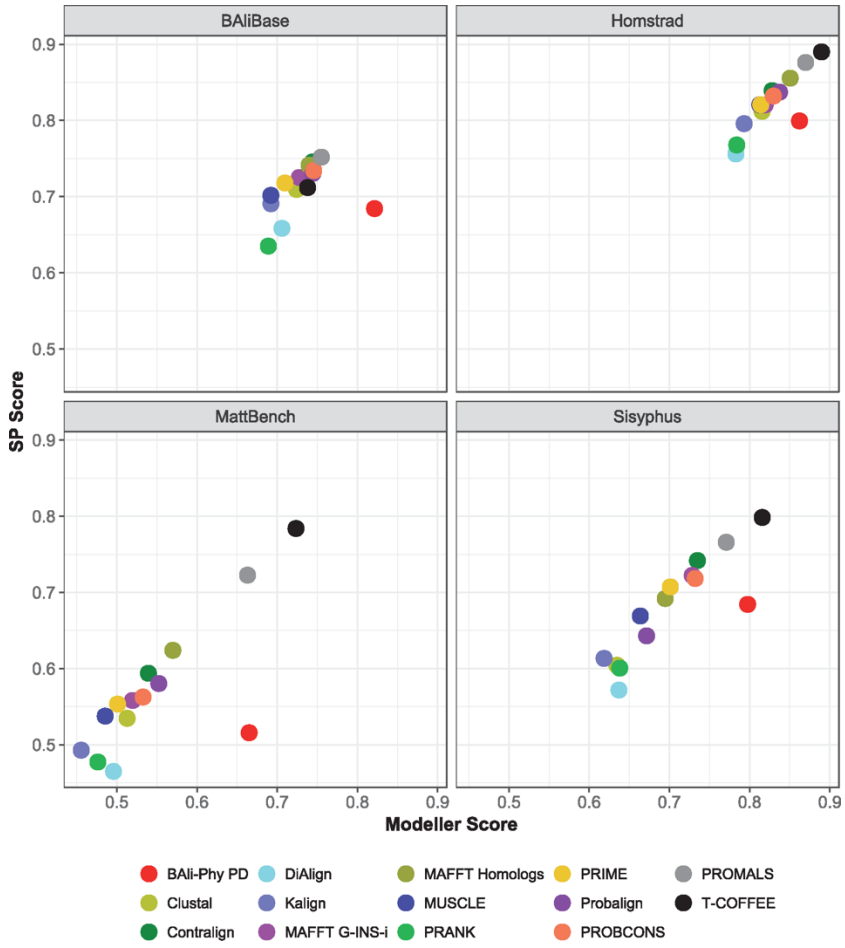
Modeler vs SP-Score on 120 Simulated Datasets



BAli-Phy is best!

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

Modeler score vs SP-score on 1192 biological 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

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

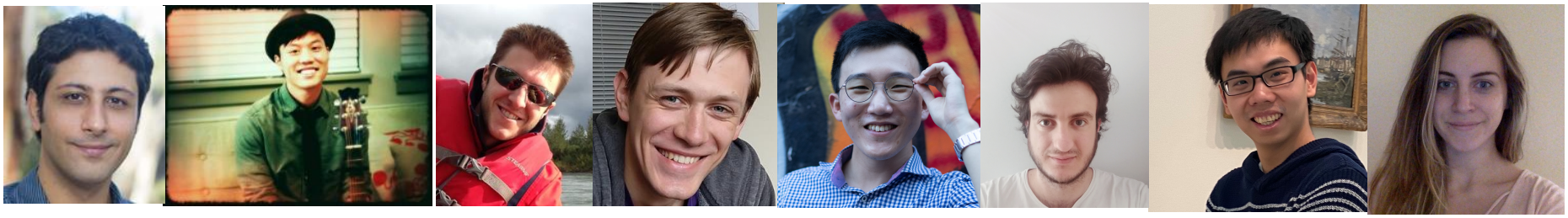
All these explanations are likely true, but the relative contributions are unknown.

Final comments

- 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.

- Yet the differences between performance on biological and simulated datasets is very troubling – and depending on the resolution, there are potential implications for phylogeny estimation as well.

Acknowledgments



Left to right: Siavash Mirarabbaygi, Nam-phuong Nguyen, Mike Nute, Vlad Smirnov, Minhyuk Park, Paul Zaharias, Chengze Shen, Eleanor Wedell

NSF grant: 2006069 Advancing bioinformatics methods using ensembles of Hidden Markov Models)
Grainger Foundation (at UIUC)
TACC, UTCS, Blue Waters, and UIUC campus cluster

Links to papers and software at http://tandy.cs.illinois.edu/MSAproject.html

Summary of new methods

- Improved methods for MSA
  - PASTA and MAGUS
  - UPP, WITCH, WITCH-ng
  - EMMA (for adding full-length sequences)
- Applications that benefit from better MSAs
  - Protein sequence classification: HIPPI
  - Taxon identification and abundance profiling: TIPP, TIPP2, TIPP3 (in development)
  - Phylogenetic placement: SCAMPP, BATCH-SCAMPP