TIPP: Taxon Identification and Phylogenetic Profiling

Tandy Warnow

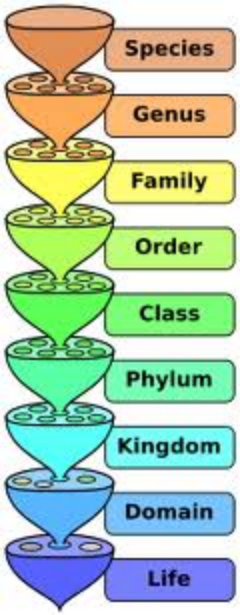
The Department of Computer Science

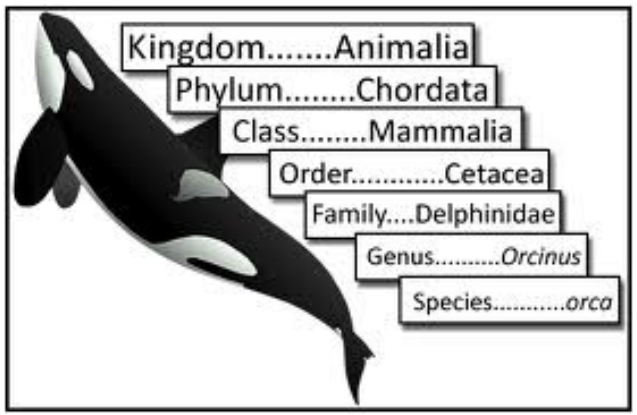
TIPP

- Submitted for publication
- Developers Nam Nguyen and Siavash Mirarab (PhD students in Computer Science at UTCS)
- Other co-authors: Mihai Pop and Bo Liu (University of Maryland, College Park)

Metagenomic Taxon Identification

Objective: classify short reads in a metagenomic sample





Two Basic Questions

1. What is this fragment? (Classify each fragment as well as possible.)

2. What is the taxonomic distribution in the dataset? (Note: helpful to use marker genes.)

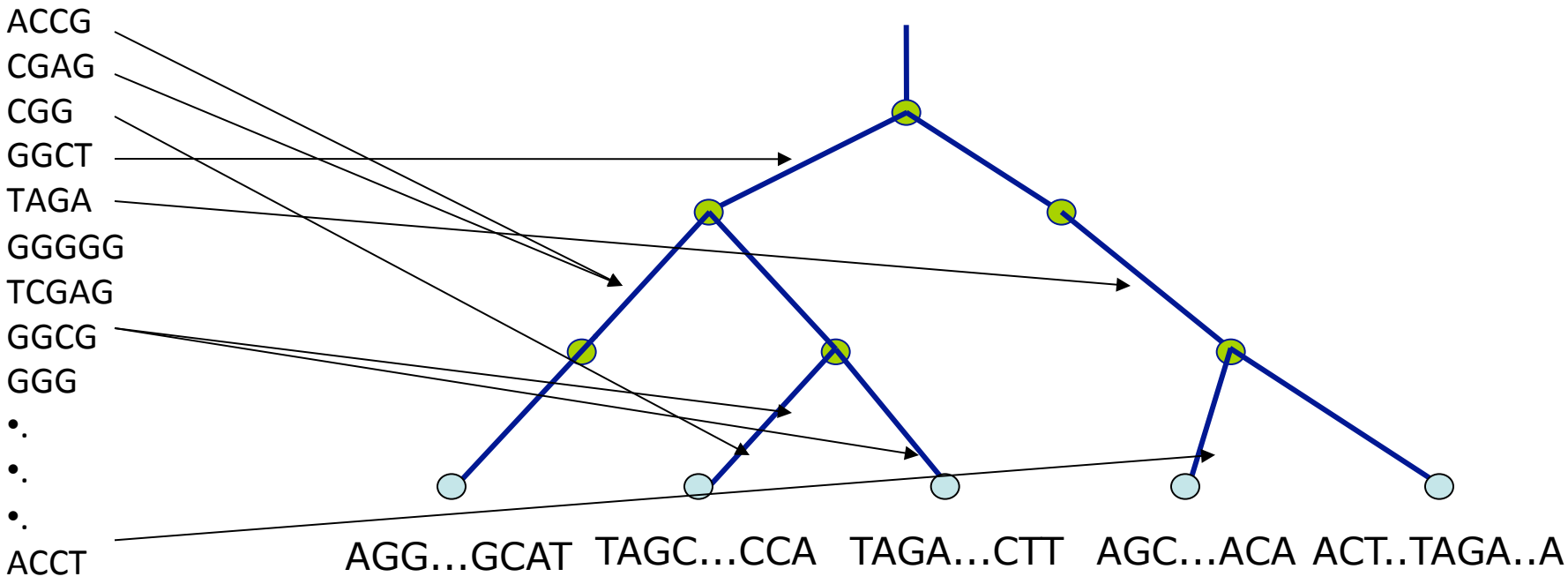
Identifying Fragments using Marker Genes

- Approach:
  - Determine the gene for the fragment (if possible), thus producing a set of "bins" (one for each gene, and a bin for "unclassified")
  - For each gene, classify each fragment:
    - Construct a reference alignment and tree for full-length sequences for that gene
    - Place each fragment within the tree
    - Predict taxon identification (species, genus, etc.) from the placement

Phylogenetic Placement

Fragmentary sequences from some gene

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



Phylogenetic Placement

Step 1: Align each query sequence to backbone alignment

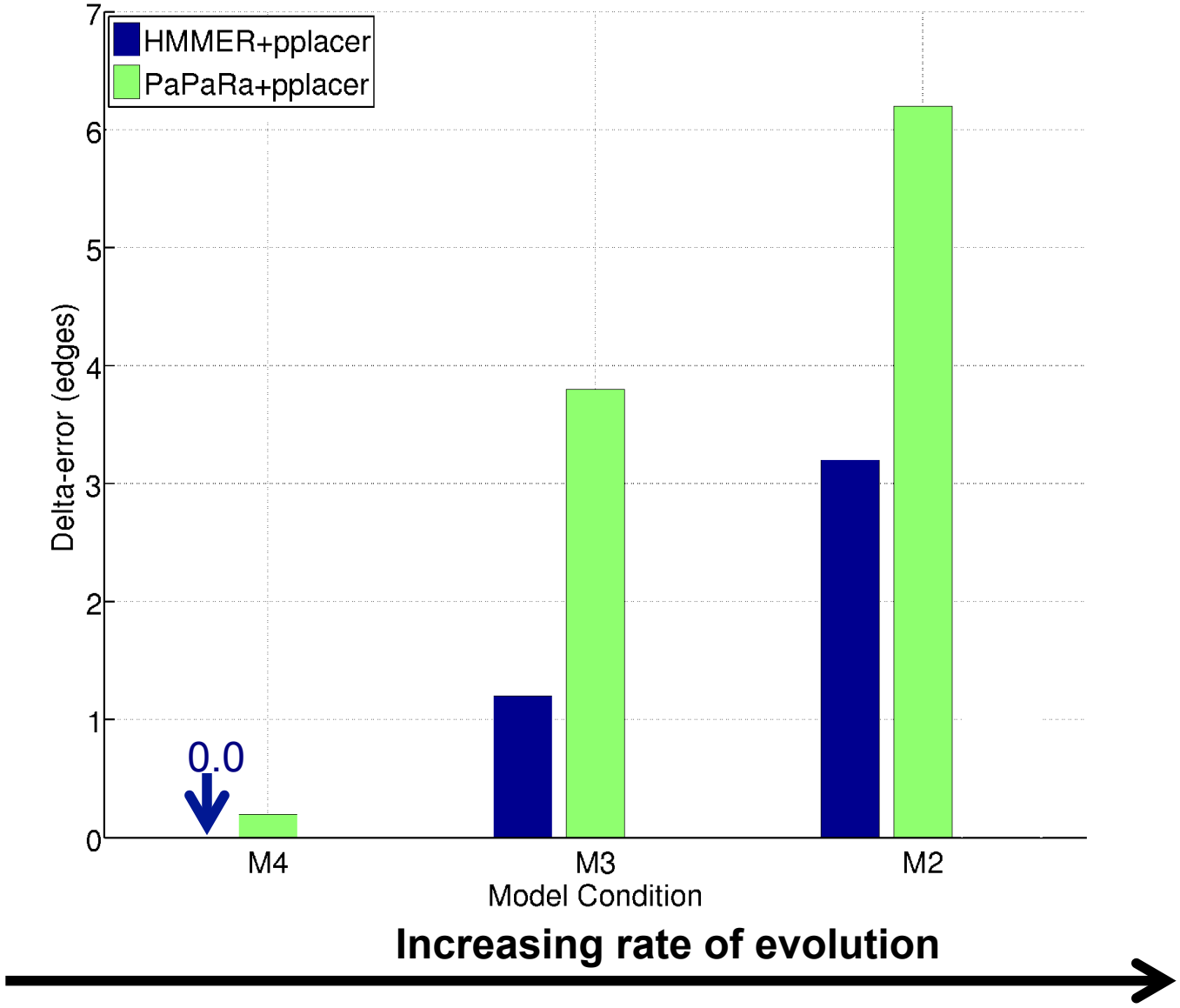
Step 2: Place each query sequence into backbone tree, using extended alignment

Phylogenetic Placement

- Align each query sequence to backbone alignment
  - HMMALIGN (Eddy, Bioinformatics 1998)
  - PaPaRa (Berger and Stamatakis, Bioinformatics 2011)
- Place each query sequence into backbone tree
  - Pplacer (Matsen et al., BMC Bioinformatics, 2011)
  - EPA (Berger and Stamatakis, Systematic Biology 2011)

Note: pplacer and EPA use maximum likelihood, and are reported to have the same accuracy.

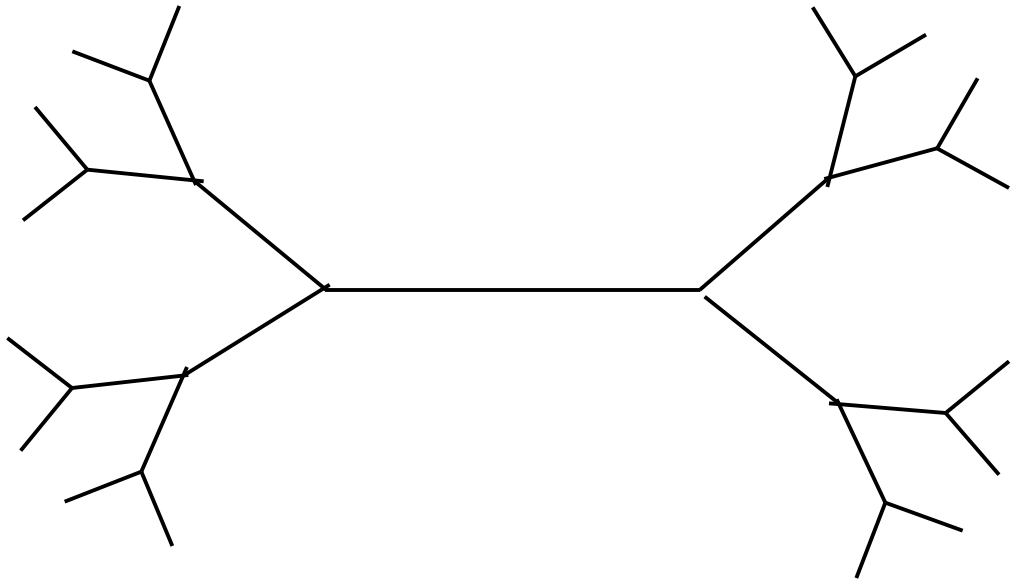
HMMER vs. PaPaRa placement error



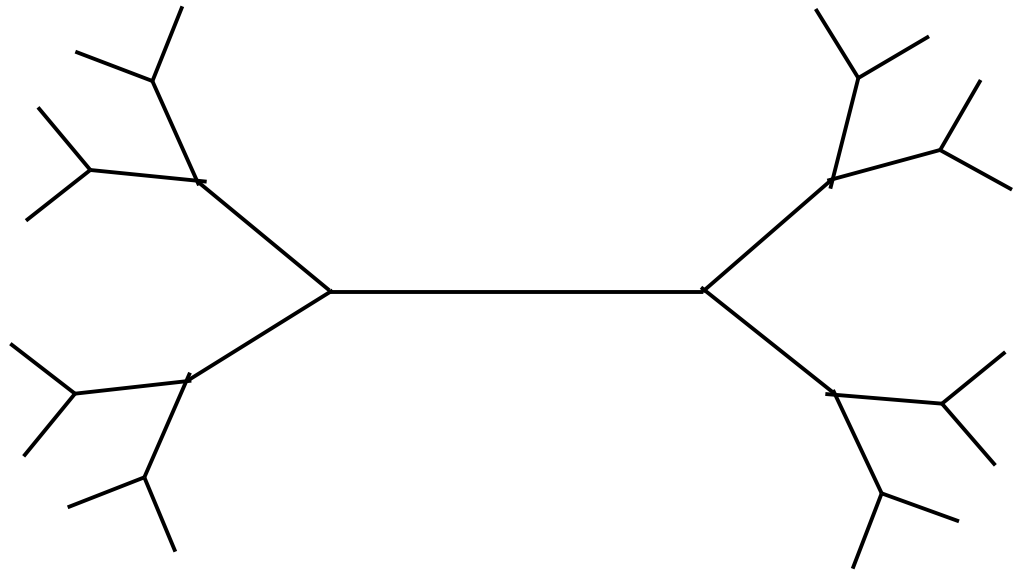
HMMER+pplacer

Steps:

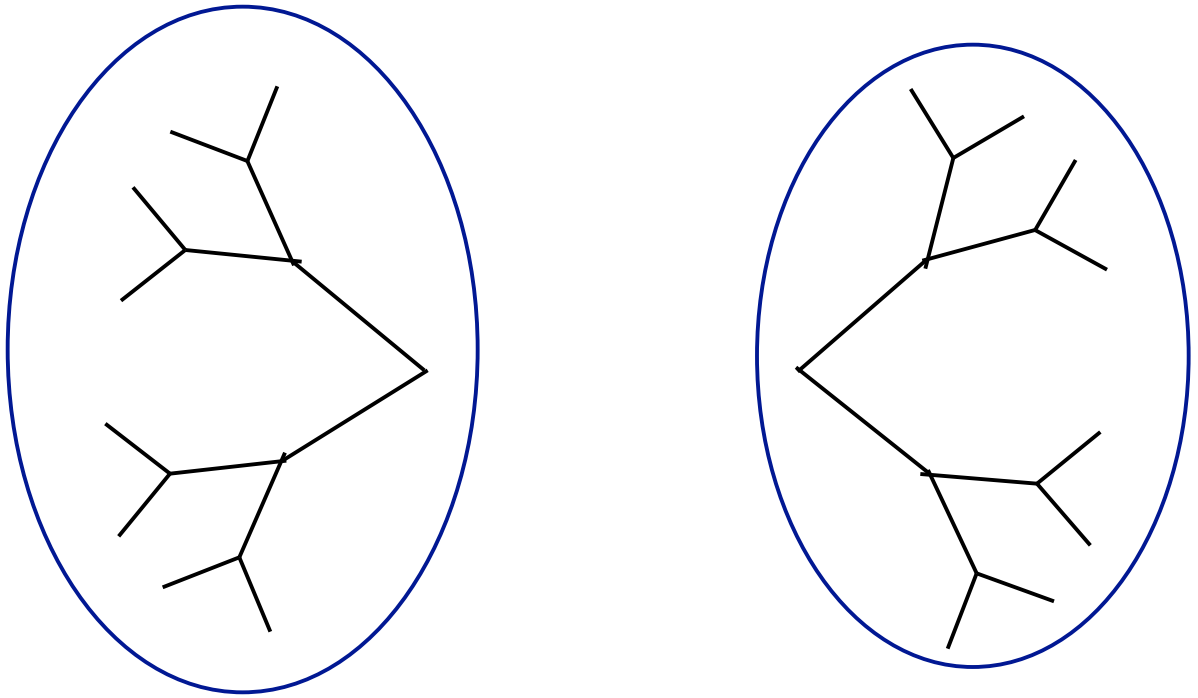
1) Build one HMM for the entire alignment
2) Align fragment to the HMM, and insert into alignment
3) Insert fragment into tree to optimize likelihood



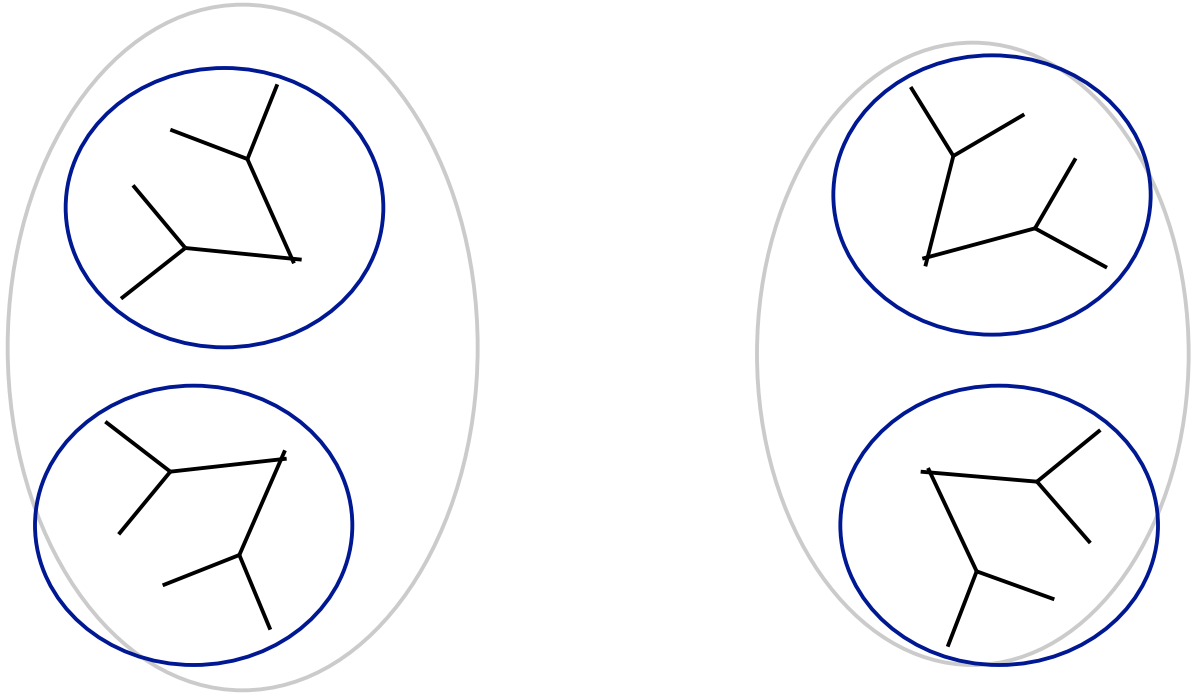
One Hidden Markov Model for the entire alignment?



Or 2 HMMs?



Or 4 HMMs?

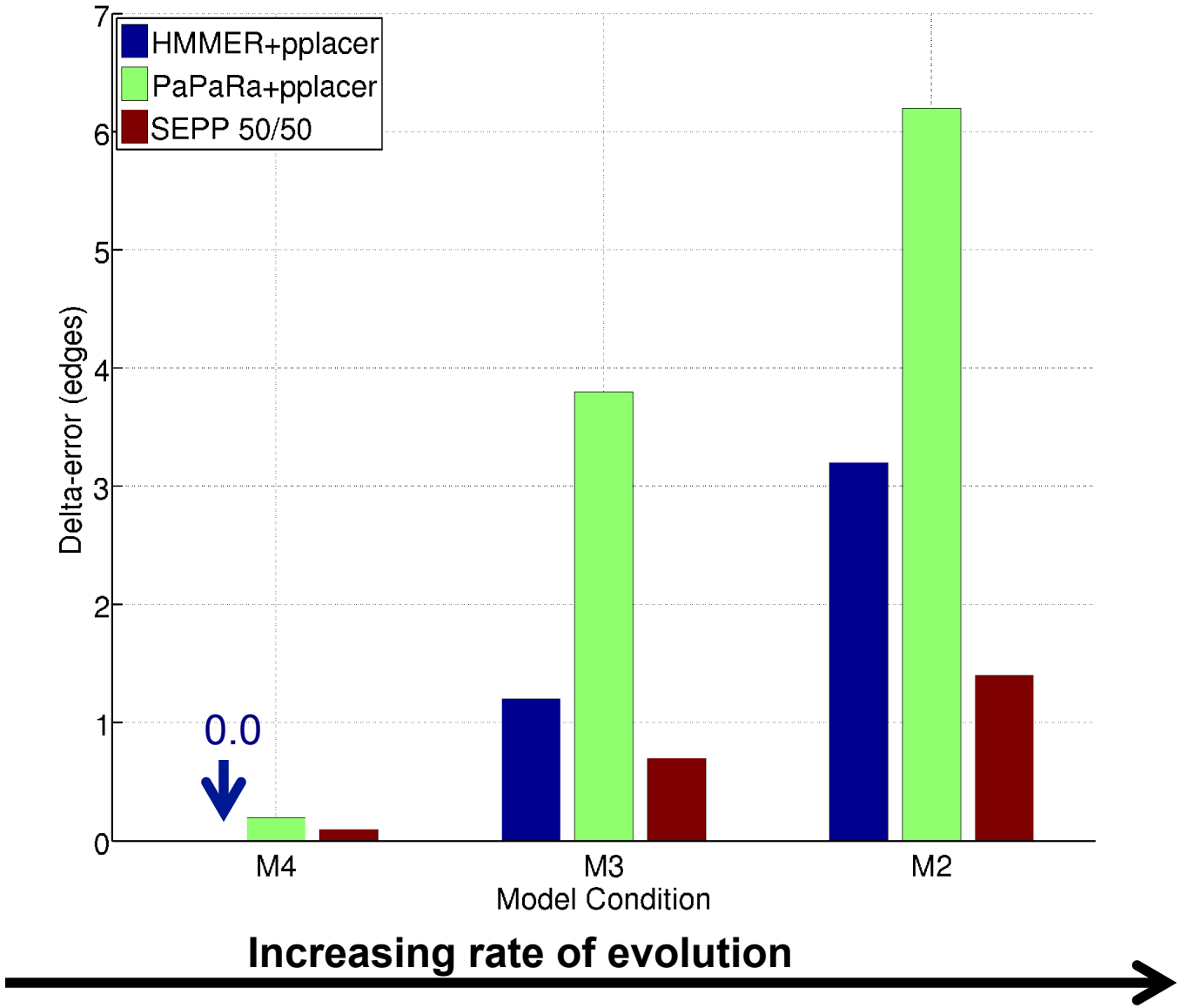


SEPP

- **SEPP: SATé-enabled Phylogenetic Placement**, by Mirarab, Nguyen, and Warnow

- Pacific Symposium on Biocomputing, 2012 (special session on the Human Microbiome)

SEPP(10%), based on ~10 HMMs



TIPP: SEPP + statistics

Using SEPP as a taxon identification technique has high recall but low precision (classifies almost everything)

TIPP: dramatically reduces false positive rate with small reduction in true positive rate, by considering uncertainty in alignment (HMMER) and placement (pplacer)

Taxonomic Identification

Objective: Identify species/genus/family (etc.) for each fragment within the sample.

Methods:

PhymmBL (Brady & Salzberg, Nature Methods 2009)

NBC (Rosen, Reichenberger, and Rosenfeld, Bioinformatics 2011)

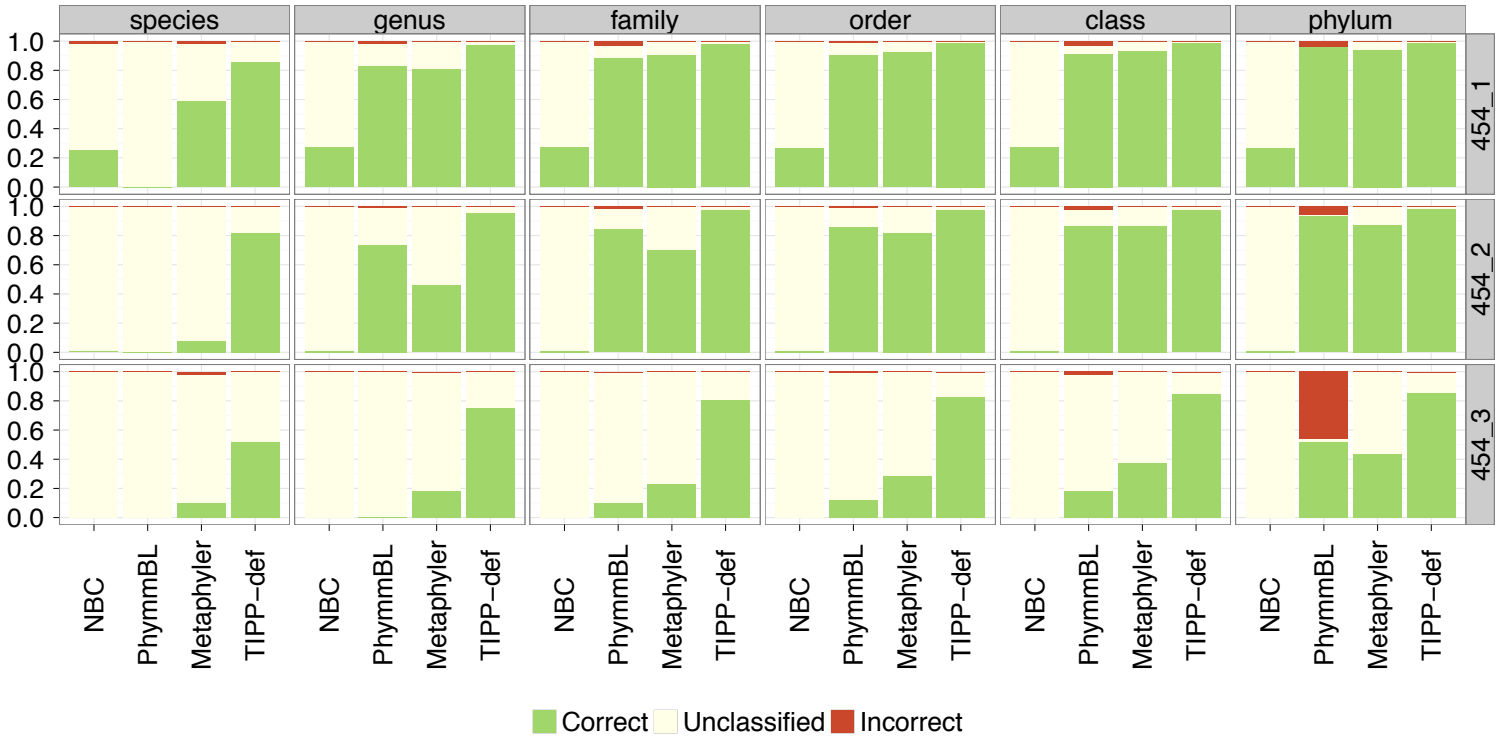
Metaphyler (Liu et al., BMC Genomics 2011), from the Pop lab at the University of Maryland

Metaphyler is a marker-based method.

Marker gene are single-copy, universal, and resistant to horizontal transmission.

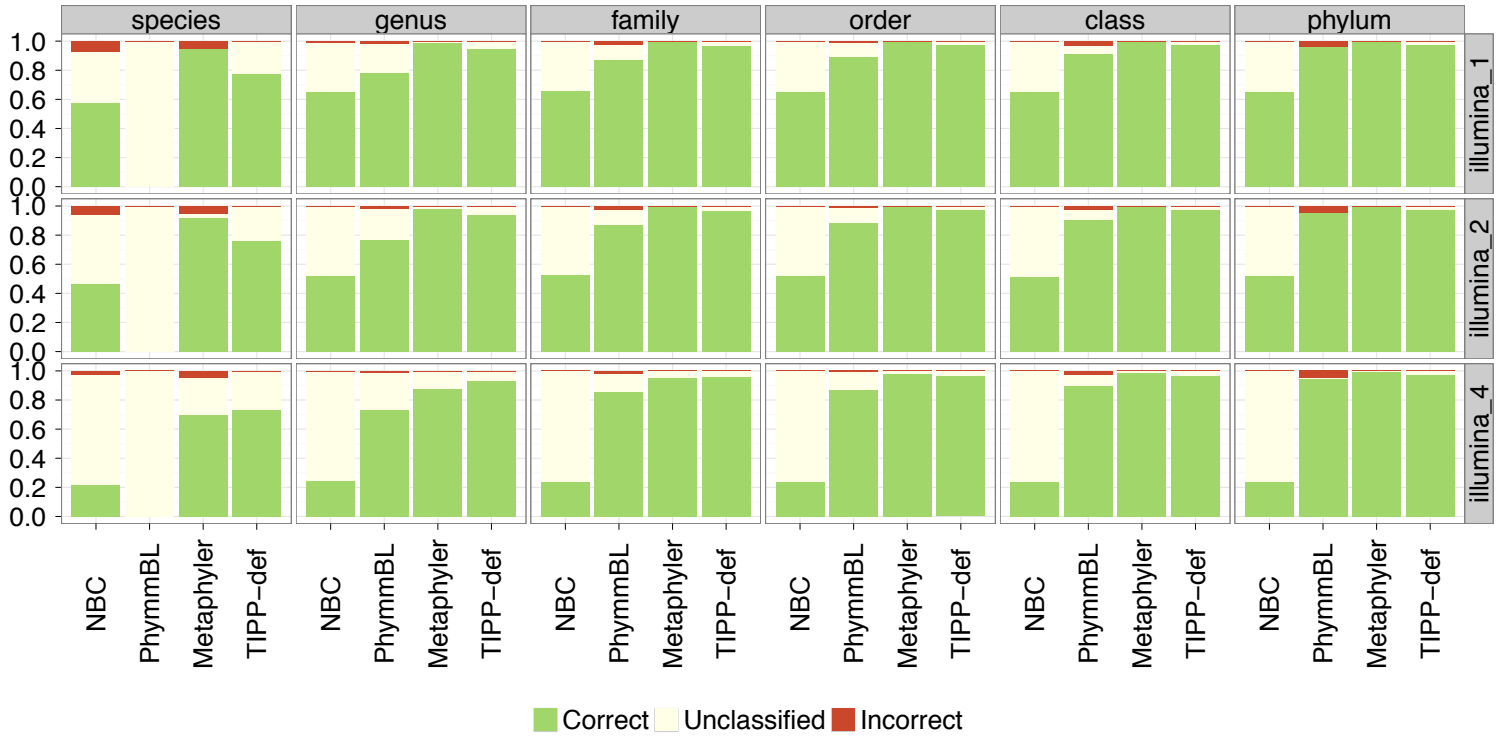
We test this for the 30 marker genes used in Metaphyler.

Criteria: Correct classification, incorrect classification, or no classification, at each level.



(a) 454 error model

Figure: Non-leave-one-out experiments comparing the classification accuracy for NBC, PhymmBL, MetaPhyler and TIPP-default (i.e., TIPP-default refers to TIPP(95%,95%,100)) for fragments simulated from the 30 marker genes under 454-like errors.



(a) Illumina error model

Figure: Non-leave-one-out experiments comparing the classification accuracy for NBC, PhymmBL, MetaPhyler and TIPP-default (i.e., TIPP-default refers to TIPP(95%,95%,100)) for fragments simulated from the 30 marker genes under Illumina-like errors.

Abundance Profiling

Objective: Distribution of the species (or genera, or families, etc.) within the sample.

For example: The distribution of the sample at the species-level is:

50%	species A
20%	species B
15%	species C
14%	species D
1%	species E

Abundance Profiling

Objective: Distribution of the species (or genera, or families, etc.) within the sample.

Leading techniques:

PhymmBL (Brady & Salzberg, Nature Methods 2009)

NBC (Rosen, Reichenberger, and Rosenfeld, Bioinformatics 2011)

Metaphyler (Liu et al., BMC Genomics 2011), from the Pop lab at the University of Maryland

MetaphlAn (Segata et al., Nature Methods 2012), from the Huttenhower Lab at Harvard

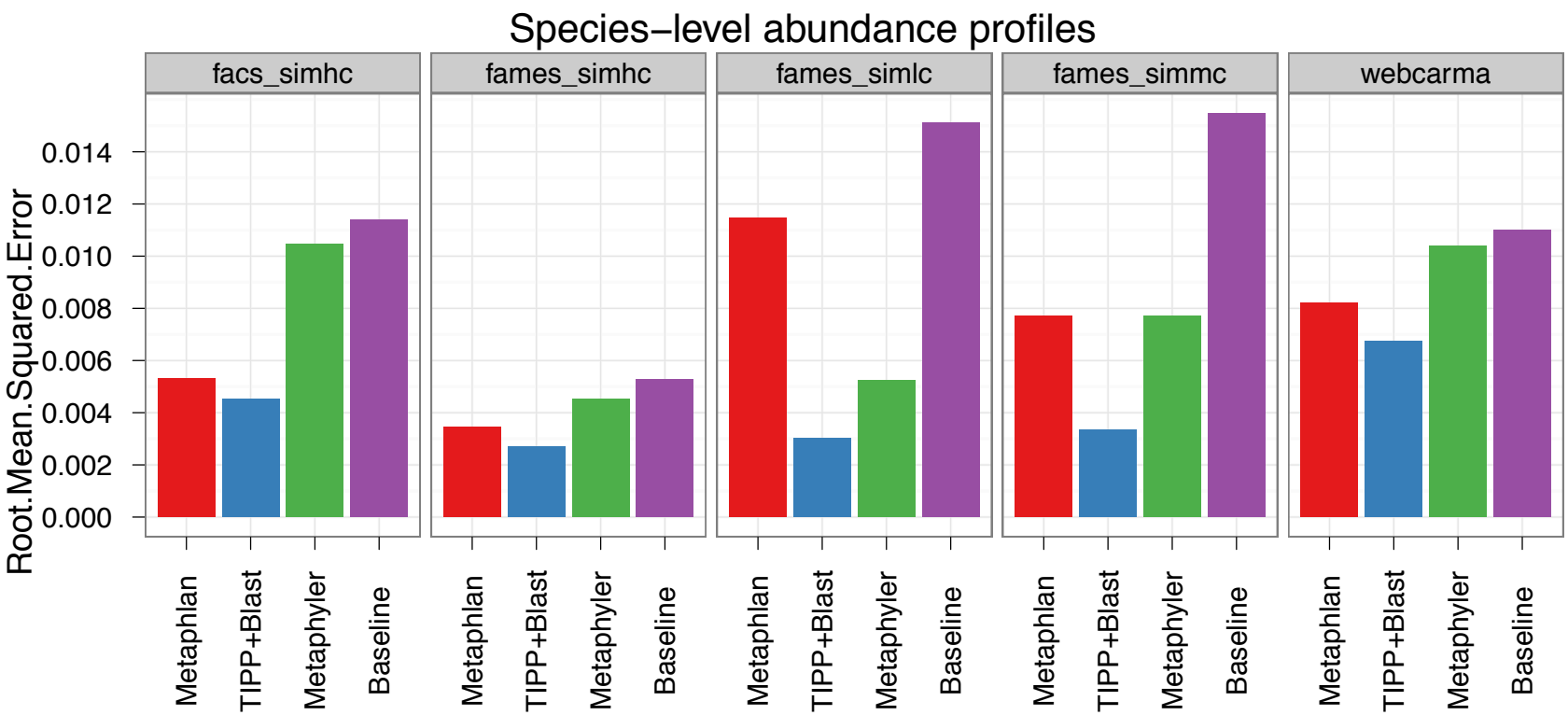
Metaphyler and MetaphlAn are marker-based techniques (but use different marker genes).

Marker gene are single-copy, universal, and resistant to horizontal transmission.

Table: Summary of all simulated abundance datasets. Complexity refers to the distribution of species in the profile: high complexity datasets have an even distribution of species, low complexity datasets have a staggered distribution of species, and medium complexity datasets fall in between.

Dataset	Genomes	Complexity	Seq. Model	Reads	Avg. length
MetaPhlAn HC	100	High	NA	1000000	88
MetaPhlAn LC	25	Low	NA	240000	88
FAMeS HC	113	High	DOE-JGI	116771	949
FAMeS MC	113	Medium	DOE-JGI	114457	969
FAMeS LC	113	Low	DOE-JGI	97495	951
FACS HC	19	High	454	26984	268
FACS HC Illumina	19	High	Illumina	300000	100
WebCarma	25	High	454	25000	265
WebCarma Illumina	25	High	Illumina	300000	100

Short fragment datasets: average length at most 100
Long fragment datasets: average length 265 to 969



- FACs HC: Fragments simulated from 19 bacterial genomes, all in equal abundance (Stranneheim et al. 2010)

- FAMeS: Fragments simulated from 113 bacterial and archael genomes, under 3 different abundance complexity profiles. (Mavromatis et al. 2007)

- WebCarma: Fragments simulated from 25 bacterial genomes, all in equal abundance (Gerlach and Stoye 2011).

Table: The average *RMSE* on the short and long fragment datasets. Note that PhymmBL does not output any species level classifications. We use TIPP(0%,0%,100) for abundance profiling (see SOM for results using other variants). The best results for each level and fragment length are in boldface.

Short Fragments Dataset	Species	Genus	Family	Order	Class	Phylum
NBC	0.022	0.026	0.028	0.029	0.030	0.038
PhymmBL	NA	0.026	0.028	0.029	0.028	0.035
MetaPhlAn	**0.012**	**0.012**	0.012	0.014	0.017	0.020
MetaPhyler	0.082	0.046	0.027	0.019	0.025	0.017
TIPP	0.013	**0.012**	**0.011**	**0.012**	**0.016**	**0.014**
Long Fragments Dataset	Species	Genus	Family	Order	Class	Phylum
NBC	0.016	0.019	0.023	0.025	0.031	0.033
PhymmBL	NA	0.018	0.020	0.021	**0.023**	0.024
MetaPhlAn	0.023	0.020	0.019	0.023	0.031	0.025
MetaPhyler	0.061	0.026	0.024	0.024	0.040	0.026
TIPP	**0.013**	**0.014**	**0.018**	**0.020**	0.032	**0.017**

Observations

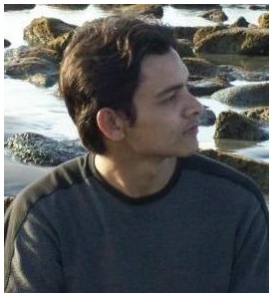
- Classification of fragments:
  - TIPP and Metaphyler are methods that use marker genes for taxon identification and phylogenetic profiling. These methods only classify fragments that are assigned to their marker genes. They will fail to classify some fragments.
  - TIPP and Metaphlyler are both more accurate than PhymmBL at classifying fragments from the 30 marker genes (perhaps not surprisingly).
  - Most methods are affected by sequencing errors, and especially by indels (454 errors). TIPP is fairly robust to 454 error (indels).
- Taxonomic profiling:
  - Marker-based profiling can produce more accurate taxonomic profiles (distributions) than techniques that attempt to classify all fragments.
  - Using marker genes from Metaphyler, TIPP produces more accurate taxonomic distributions (profiles) than Metaphyler..
- TIPP uses multiple sequence alignment and phylogenetic placement to improve accuracy. This is probably why TIPP has better robustness to indel errors, and high sensitivity.

Future Work

- Extending TIPP to non-marker genes. TIPP easily extends as a fragment identification method (as long as the gene is represented in sufficient quantity in existing databases), and preliminary results on 16S genes show very good fragment identification. However, using non-marker genes for abundance profiling requires normalization for multi-copy and missing data.
- Implementation for HPC (big data problems).

Warnow Laboratory





PhD students: Siavash Mirarab*, Nam Nguyen, and Md. S. Bayzid**
Undergrad: Keerthana Kumar
Lab Website: http://www.cs.utexas.edu/users/phylo

Funding: Guggenheim Foundation, Packard, NSF, Microsoft Research New England, David Bruton Jr. Centennial Professorship, TACC (Texas Advanced Computing Center), and the University of Alberta (Canada)

TACC and UTCS computational resources

* Supported by HHMI Predoctoral Fellowship

** Supported by Fulbright Foundation Predoctoral Fellowship

UPP: Ultra-large alignment using SEPP[1]

Objective: highly accurate multiple sequence alignments and trees on ultra-large datasets

Authors: Nam Nguyen, Siavash Mirarab, and Tandy Warnow

In preparation – expected submission February 2014

[1] SEPP: SATe-enabled phylogenetic placement, Nguyen, Mirarab, and Warnow, PSB 2012

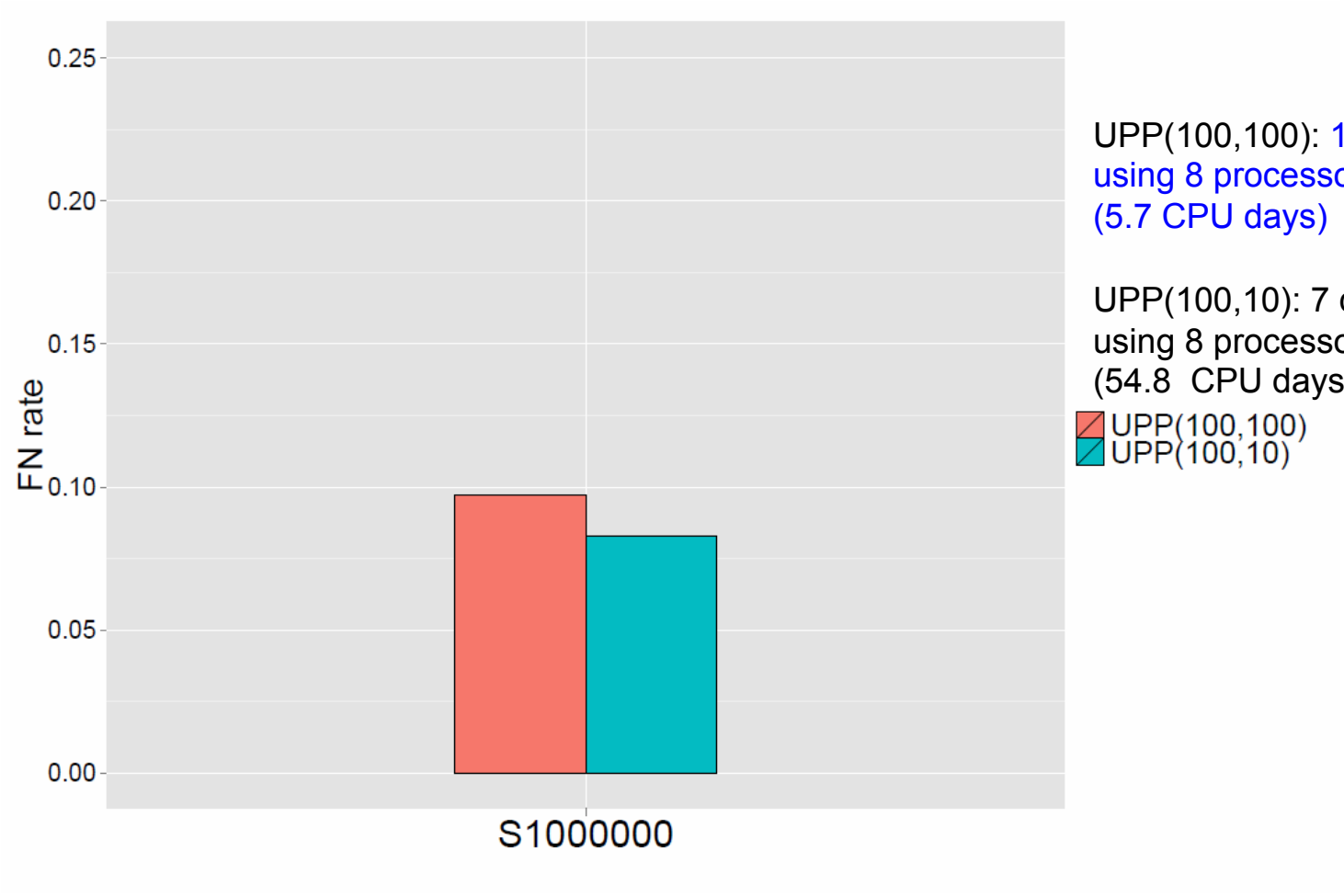
UPP: basic idea

Input: set S of unaligned sequences

Output: alignment on S

- Select random subset X of S
- Estimate “backbone” alignment A and tree T on X
- Independently align each sequence in S-X to A
- Use transitivity to produce multiple sequence alignment A* for entire set S

One Million Sequences: Tree Error



Note: UPP Decomposition improves accuracy