TIPP: Taxon Identification using Phylogeny-Aware Profiles

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

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.)

3. What are the organisms in this metagenomic sample doing together?

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:

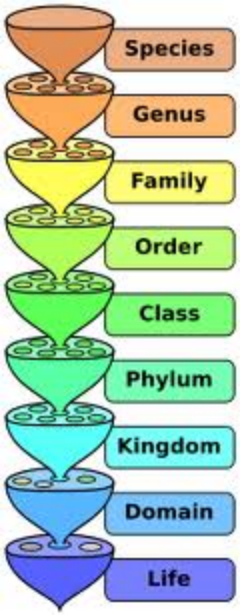
50%	species A
20%	species B
15%	species C
14%	species D
1%	species E

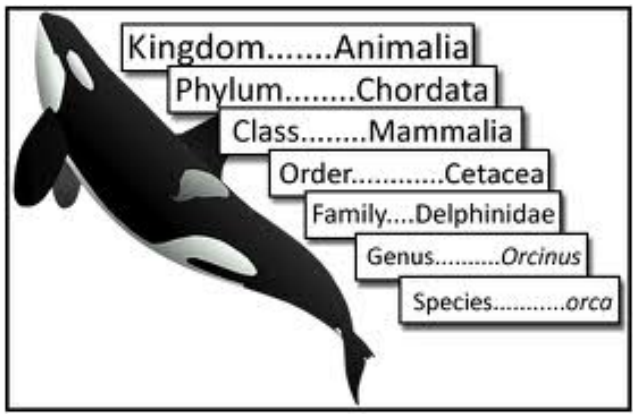
This talk

- SEPP (Mirarab et al., PSB 2012): SATé-enabled Phylogenetic Placement, and Ensembles of HMMs (eHMMs)
- TIPP (Nguyen et al., Bioinformatics 2014): Applications of the eHMM technique to metagenomic abundance classification
- TIPPI: Our planned extension to TIPP using HIPPI (Nguyen et al., to appear, BMC Genomics)

Metagenomic Taxon Identification

Objective: classify short reads in a metagenomic sample

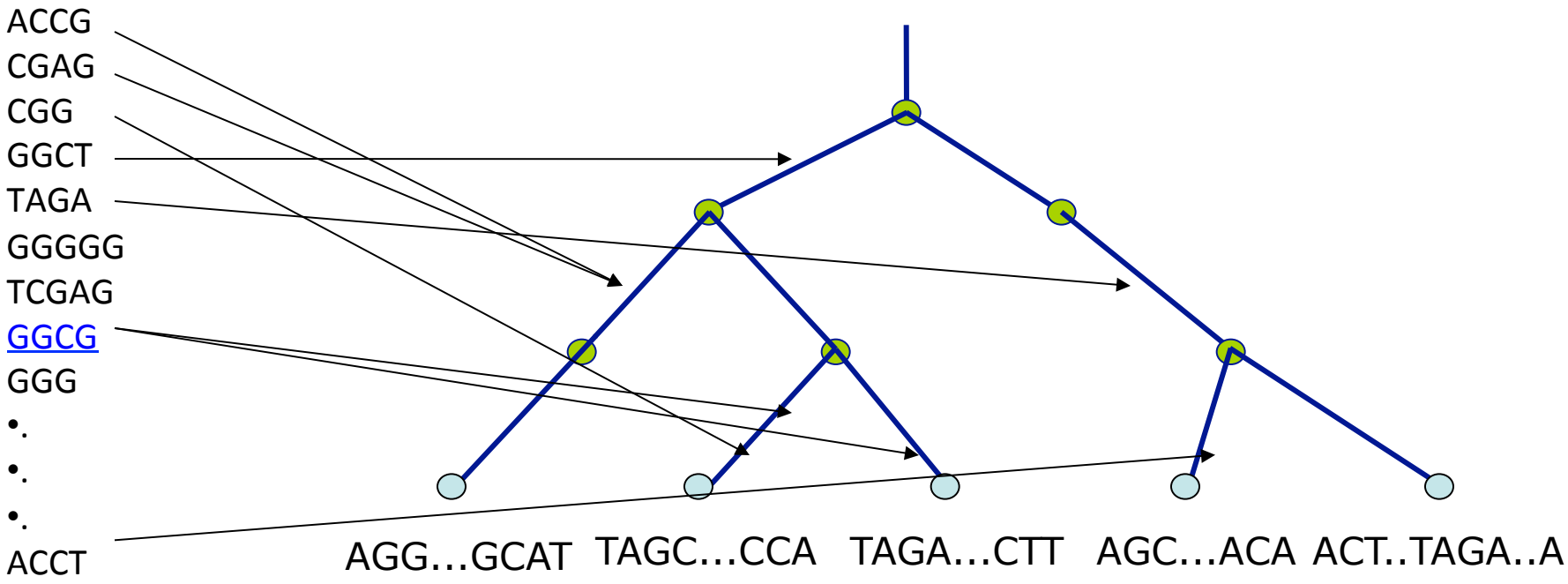




Marker-based Taxon Identification via Phylogenetic Placement

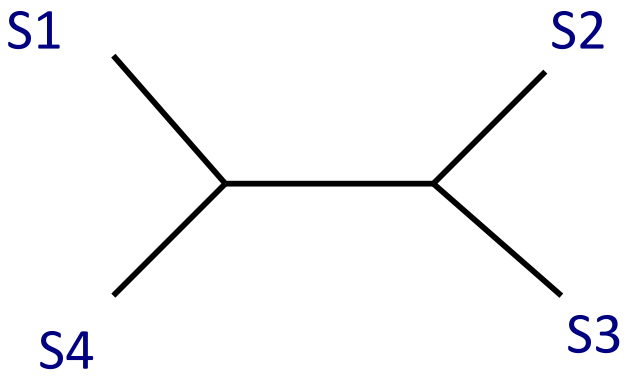
Fragmentary sequences from some gene

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



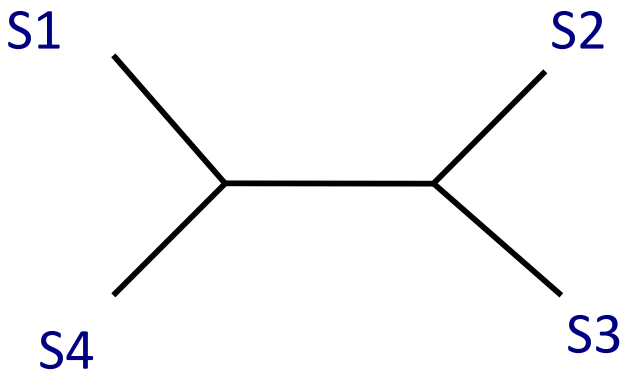
Align Sequence

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



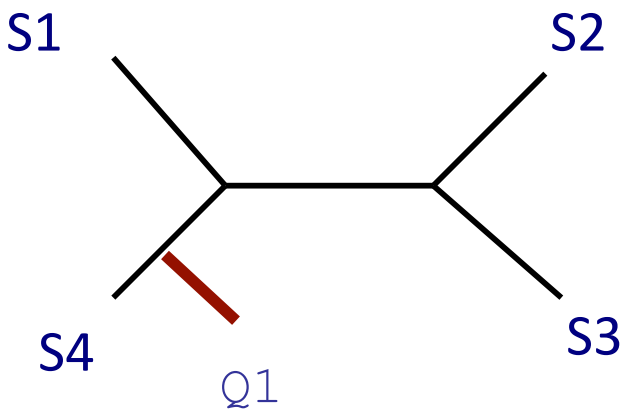
Align Sequence

```
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--------
```

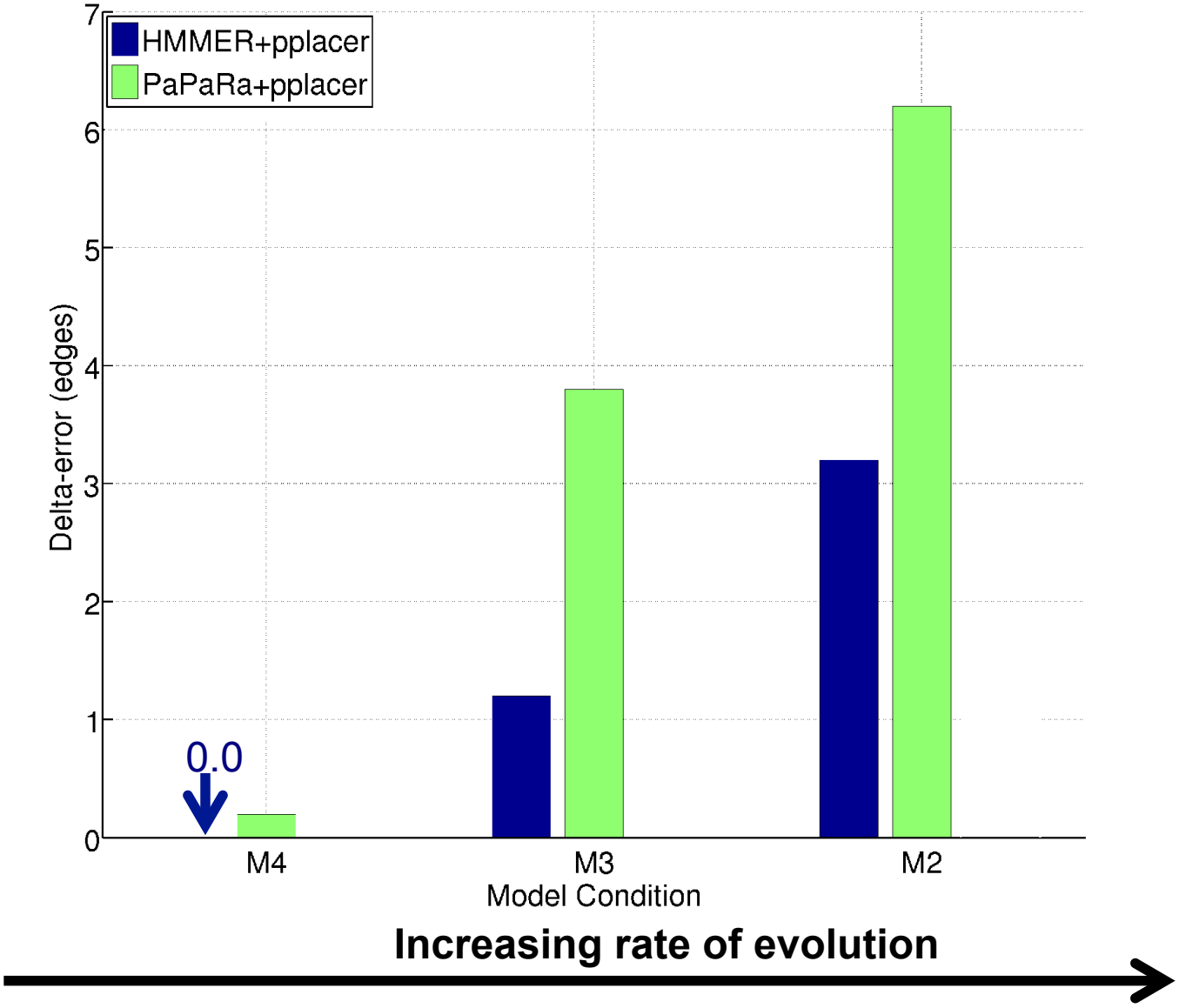


Phylogenetic Placement

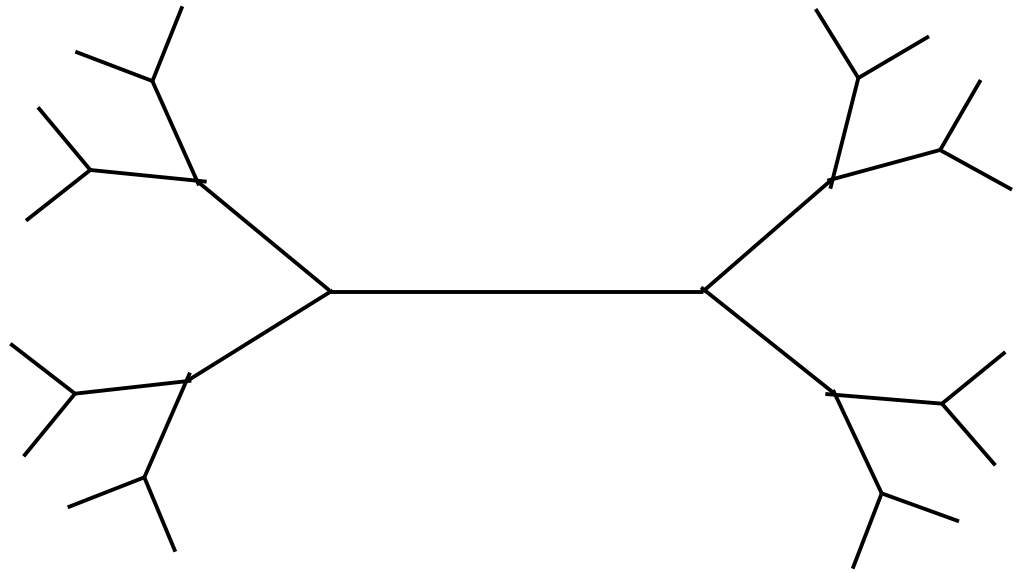
- Align each query sequence to backbone alignment
  - HMMALIGN (Eddy, Bioinformatics 1998)
  - PaPaRa (Berger and Stamatakis, Bioinformatics 2011)
- Place each query sequence into backbone tree
  - Pplacer (Matsen et al., BMC Bioinformatics, 2011)
  - EPA (Berger and Stamatakis, Systematic Biology 2011)

Note: pplacer and EPA use maximum likelihood

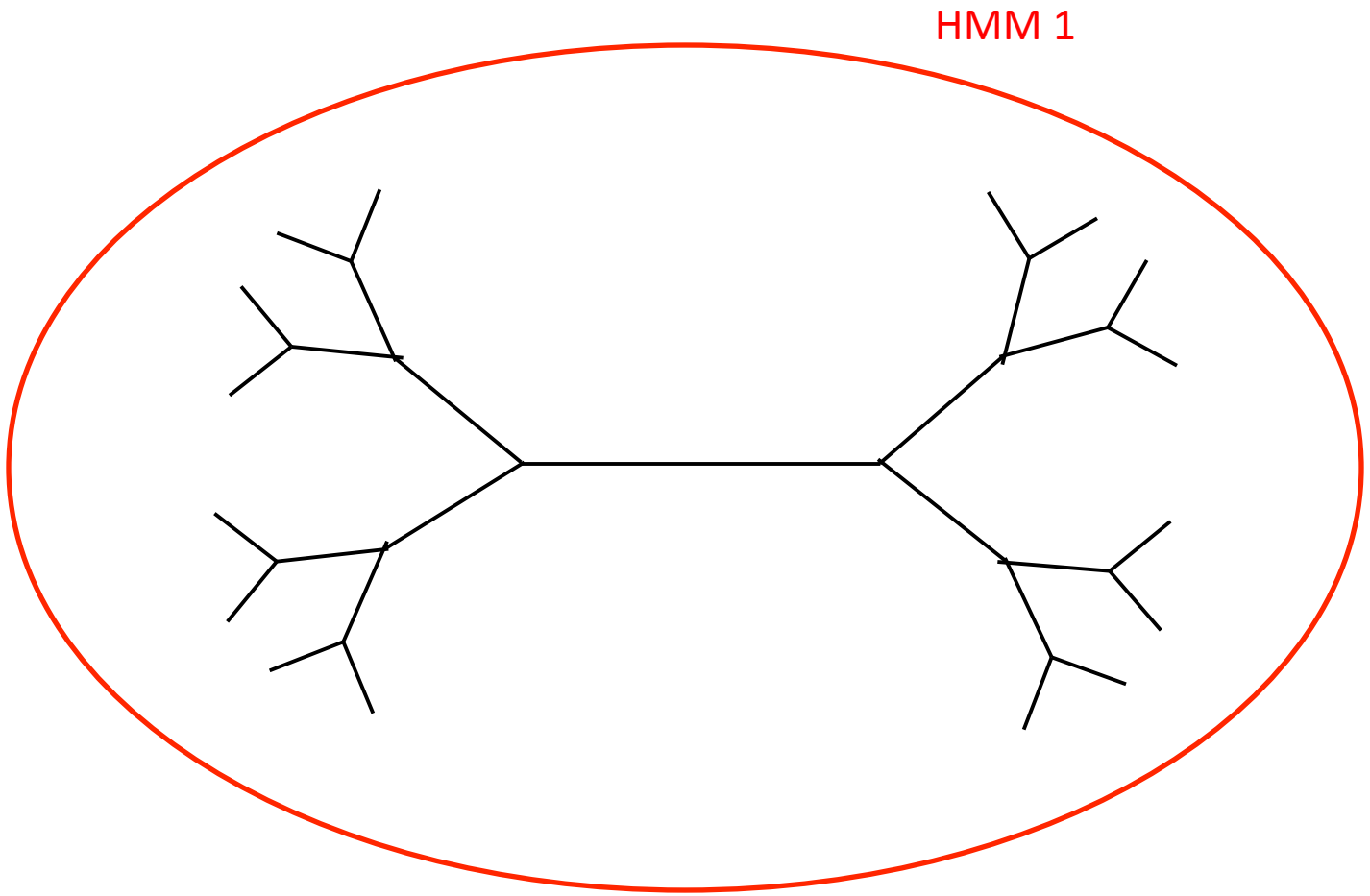
HMMER vs. PaPaRa Alignments



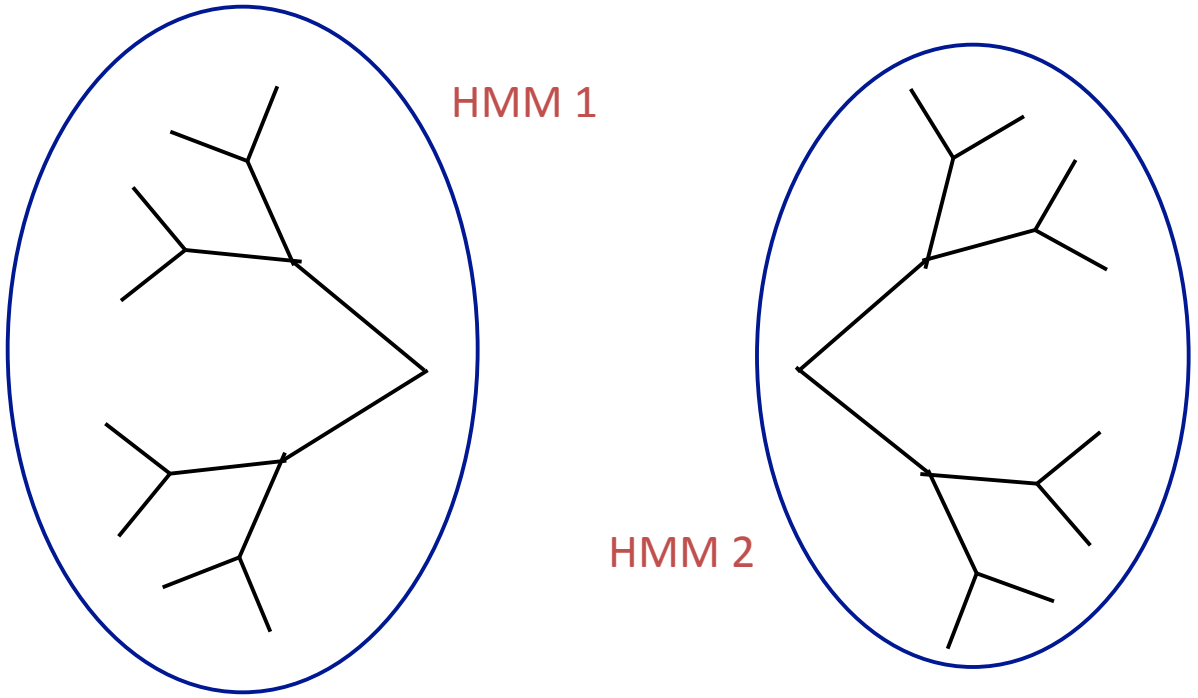
One Hidden Markov Model for the entire alignment?



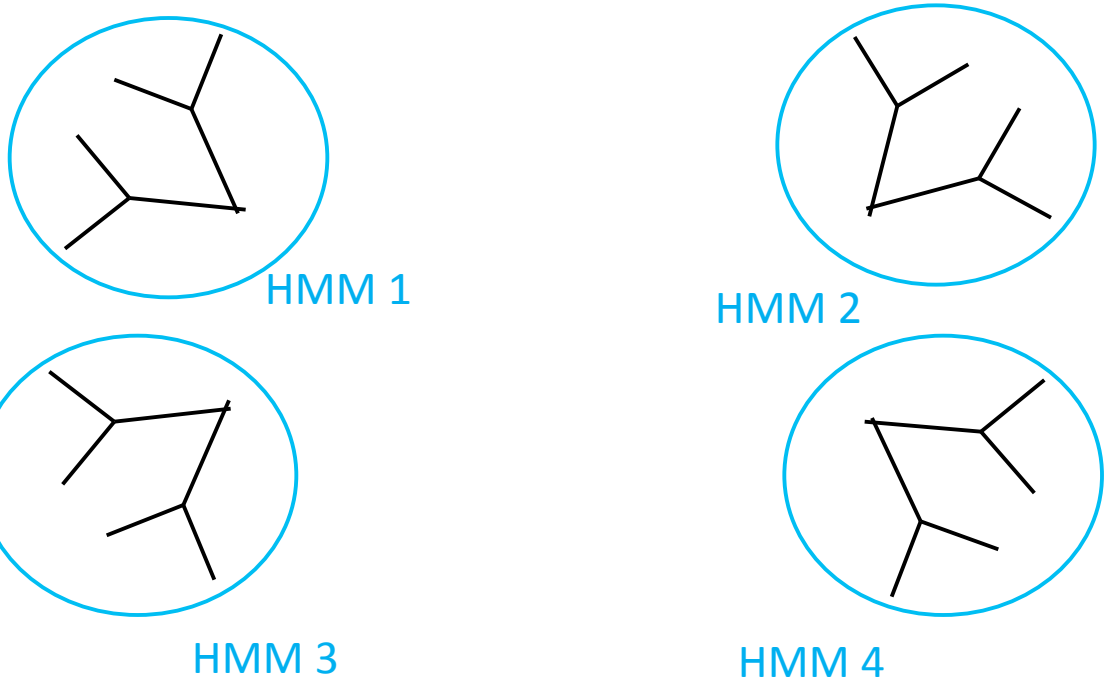
One Hidden Markov Model for the entire alignment?



Or 2 HMMs?



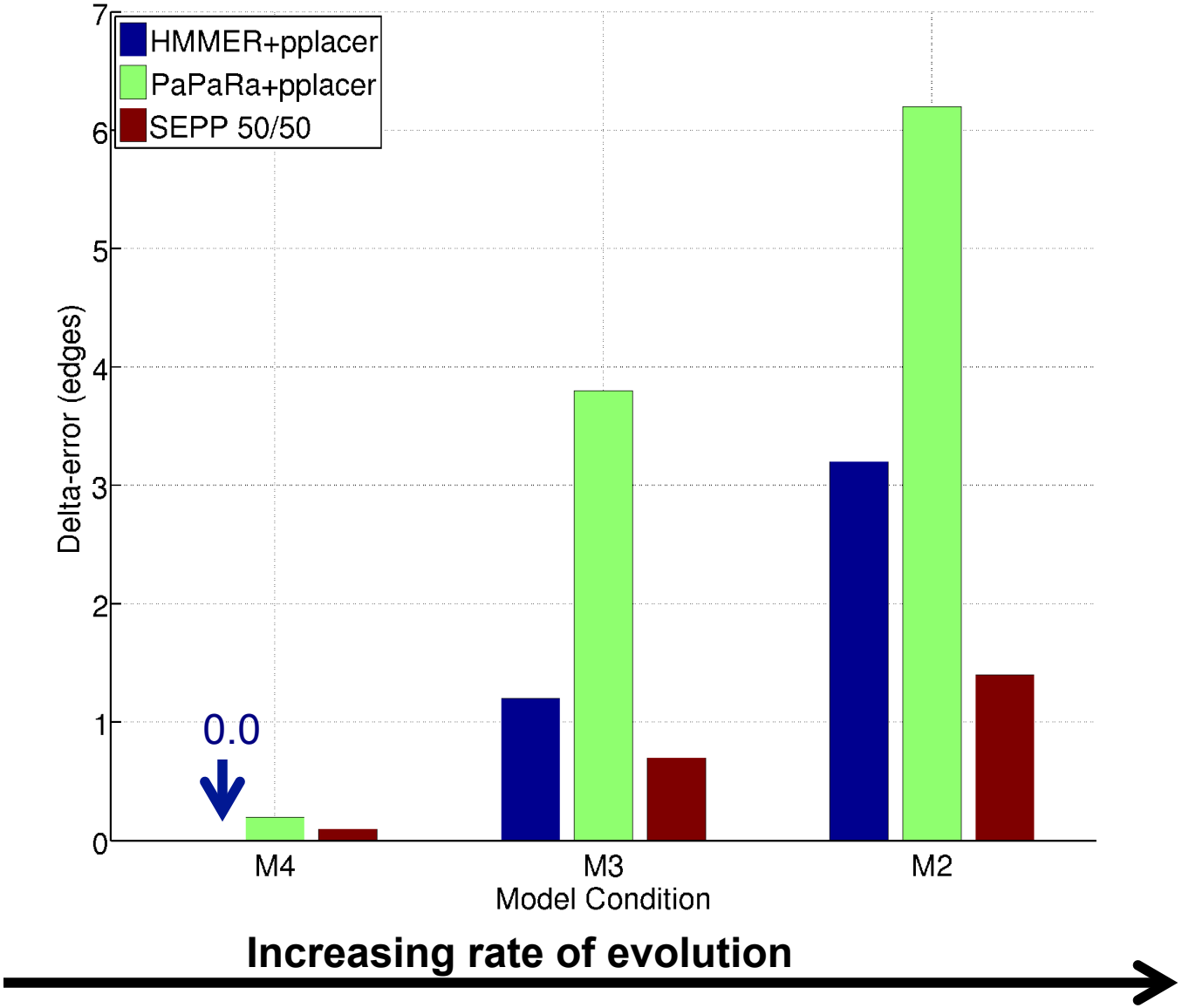
Or 4 HMMs?



SEPP Parameter Exploration

- Alignment subset size and placement subset size impact the accuracy, running time, and memory of SEPP
- 10% rule (subset sizes 10% of backbone) had best overall performance

SEPP (10%-rule) on simulated data



TIPP (https://github.com/smirarab/sepp)

TIPP (Nguyen, Mirarb, Liu, Pop, and Warnow, Bioinformatics 2014), marker-based method that only characterizes those reads that map to the Metaphyler's marker genes

TIPP pipeline

- Uses BLAST to assign reads to marker genes
- Computes UPP/PASTA reference alignments
- Uses reference taxonomies, refined to binary trees using reference alignment
- Modifies SEPP by considering statistical uncertainty in the extended alignment and placement within the tree

Abundance Profiling

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

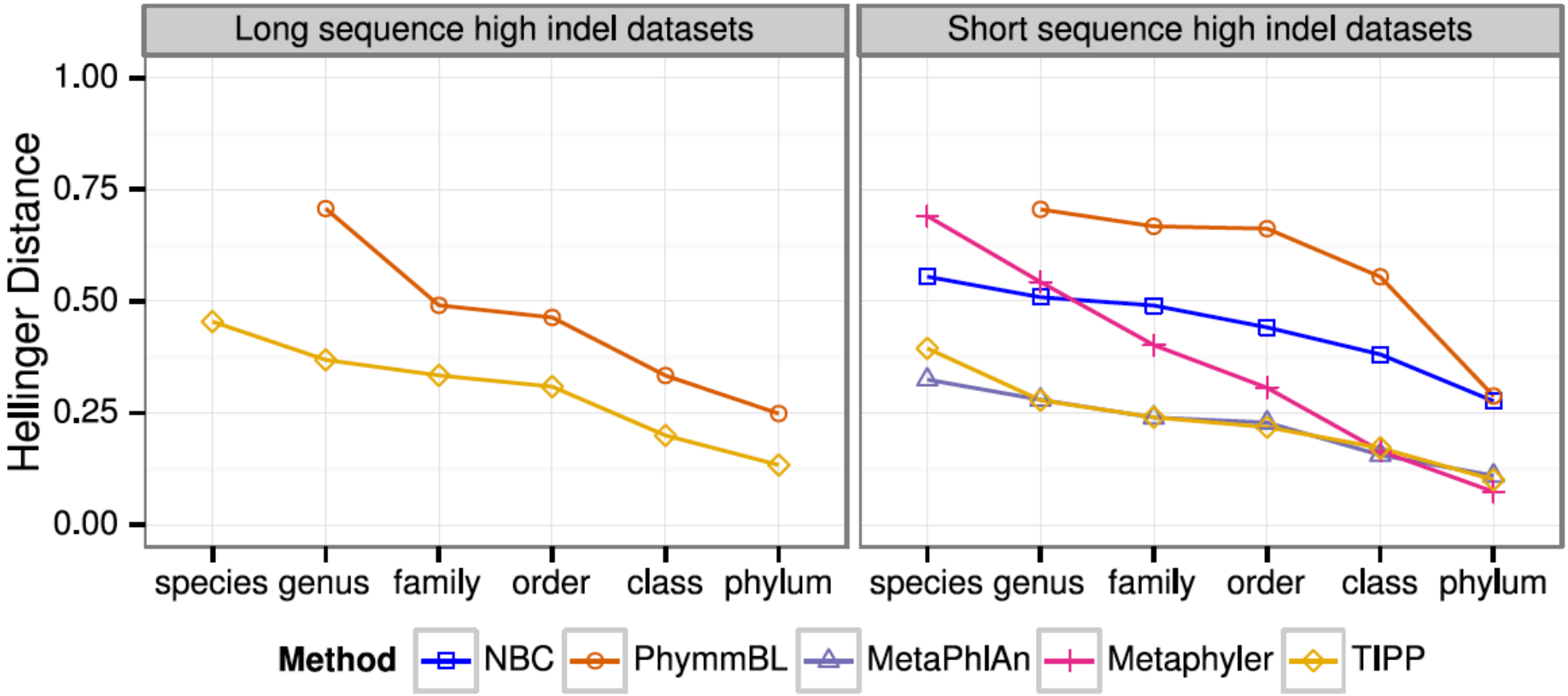
We compared TIPP to

- 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
- mOTU (Bork et al., Nature Methods 2013)

MetaPhyler, MetaPhlAn, and mOTU are marker-based techniques (but use different marker genes).

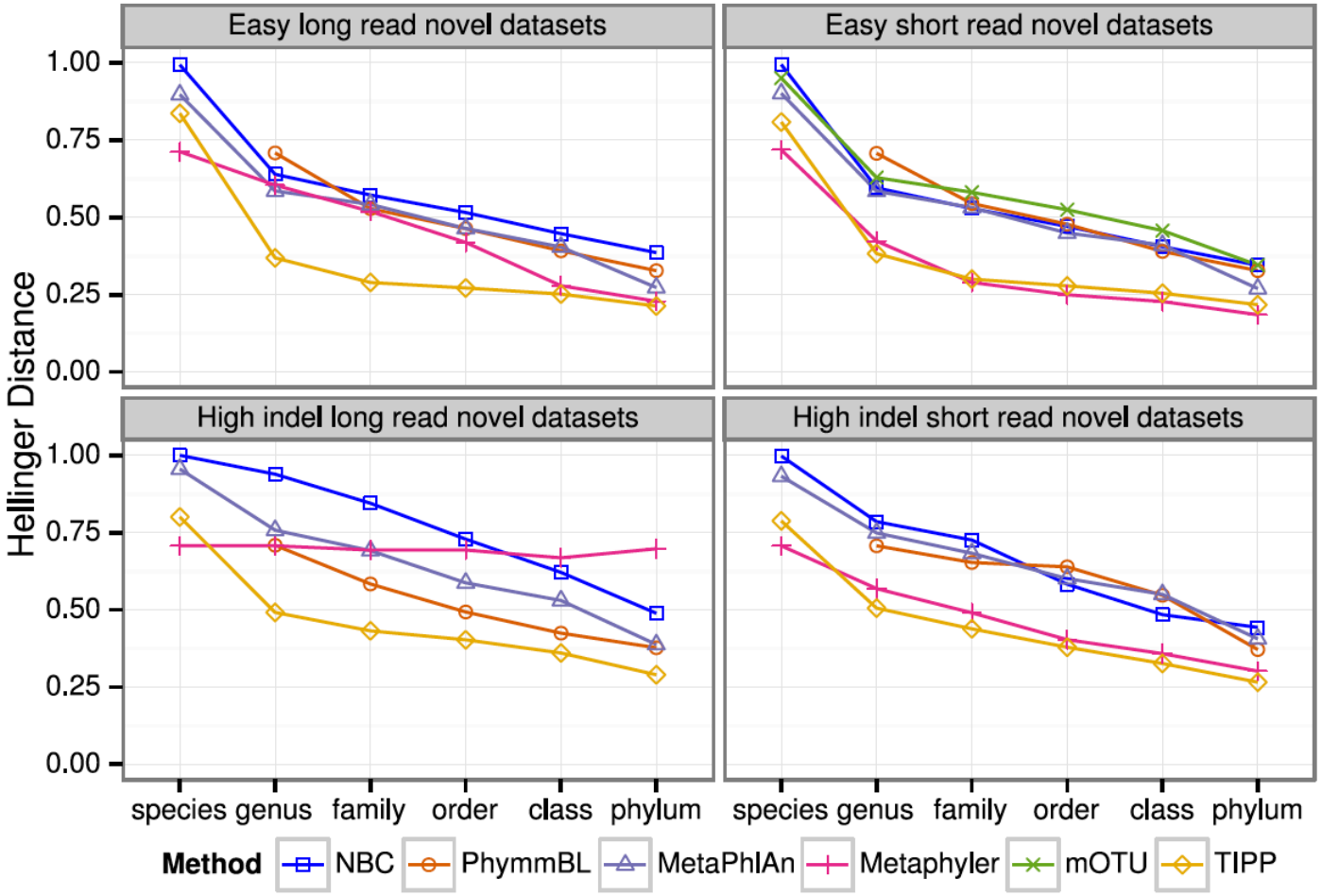
Marker gene are single-copy, universal, and resistant to horizontal transmission.

High indel datasets containing known genomes



Note: NBC, MetaPhlAn, and MetaPhyler cannot classify any sequences from at least one high indel long sequence dataset, and mOTU terminates with an error message on all the high indel datasets.

“Novel” genome datasets



Note: mOTU terminates with an error message on the long fragment datasets and high indel datasets.

TIPP vs. other abundance profilers

- TIPP is highly accurate, even in the presence of high indel rates and novel genomes, and for both short and long reads.

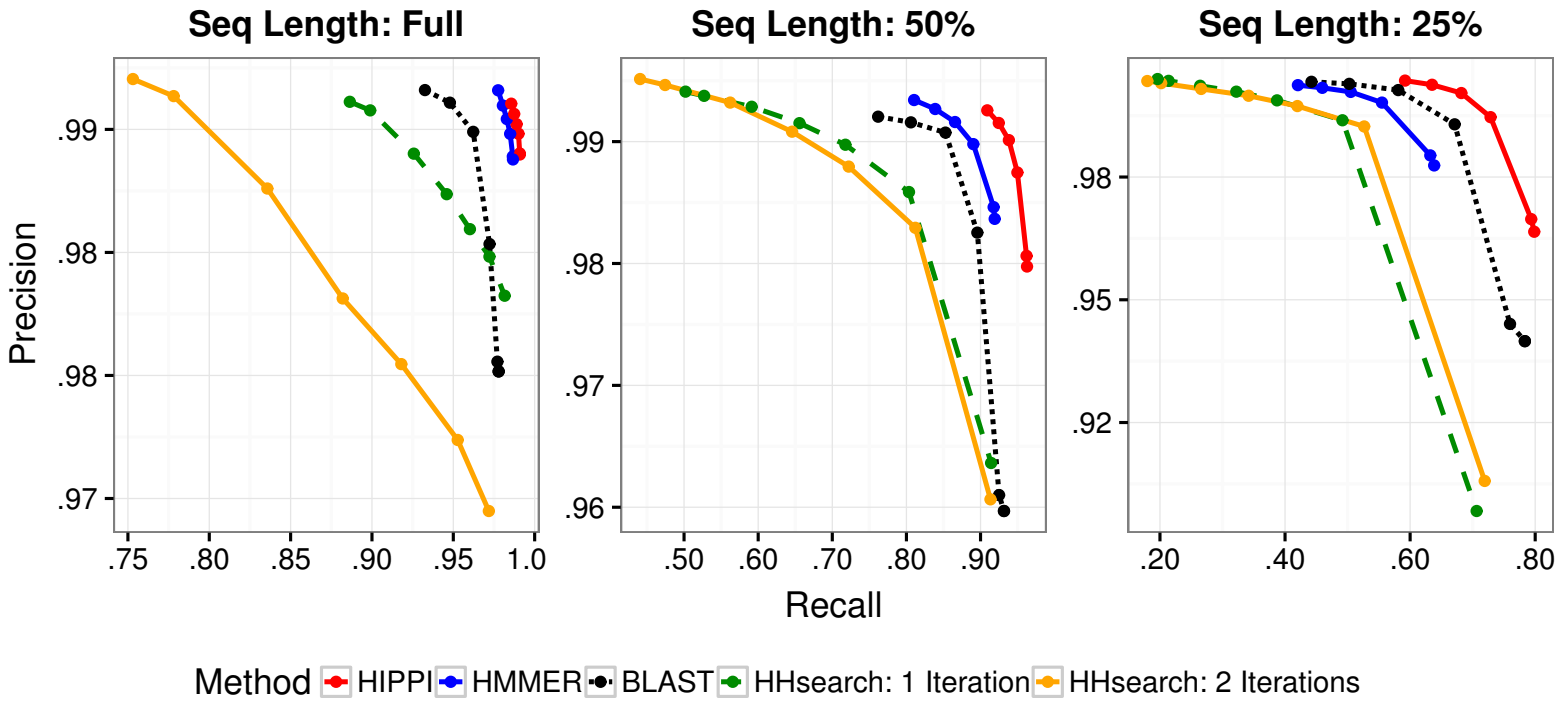
- All other methods have some vulnerability (e.g., mOTU is only accurate for short reads and is impacted by high indel rates).

- Improved accuracy is due to the use of eHMMs; single HMMs do not provide the same advantages, especially in the presence of high indel rates.

Still to do

- Evaluate TIPP in comparison to newer methods (e.g., Kraken)
- Evaluating TIPP with respect to taxonomic identification and identification of novel taxa.
- Update TIPP's design!

TIPPI: Replacing BLAST by HIPPI within TIPP



To appear, Nguyen et al., BMC Genomics

Acknowledgments




PhD students: Nam Nguyen (now postdoc at UCSD), Siavash Mirarab (now faculty at UCSD), Mike Nute, Bo Liu (now at Square)

Mihai Pop, University of Maryland

NSF grants to TW: DBI:1062335, DEB 0733029, III:AF:1513629

NIH grant to MP: R01-A1-100947

Also: Guggenheim Foundation Fellowship (to TW), Microsoft Research New England (to TW), David Bruton Jr. Centennial Professorship (to TW), Grainger Foundation (to TW), HHMI Predoctoral Fellowship (to SM)

TACC, UTCS, and UIUC computational resources

Publications using eHMMs

- "SEPP: SATé-Enabled Phylogenetic Placement.” S. Mirarab, N. Nguyen, and T. Warnow. Proceedings of the 2012 Pacific Symposium on Biocomputing
- "TIPP:Taxonomic Identification and Phylogenetic Profiling." N. Nguyen, S. Mirarab, B. Liu, M. Pop, and T. Warnow Bioinformatics, 2014;
- "Ultra-large alignments using phylogeny aware profiles". N. Nguyen, S. Mirarab, K. Kumar, and T. Warnow, Proceedings RECOMB 2015 and Genome Biology (2015)
- “HIPPI: Highly accurate protein family classification with ensembles of HMMs.” N. Nguyen, M. Nute, S. Mirarab, and T. Warnow. To appear, BMC Genomics, special issue for RECOMB Comparative Genomics 2016.

SEPP, TIPP, UPP, and HIPPI are all available in open source form on github (smirarab/sepp)

Comments

- Marker-based abundance profiling methods can be more accurate than analyses that classify all the reads.
- Improved results possible using:
  - Better sets of marker genes
  - Better ways of assigning reads to marker genes
  - Better taxonomies, or use of estimated gene trees and species trees
  - Better ways of combining classifications from multiple markers

Comments

- Marker-based abundance profiling methods can be more accurate than analyses that classify all the reads.
- Improved results possible using:
  - Better sets of marker genes
  - Better ways of assigning reads to marker genes
  - Better taxonomies, or use of estimated gene trees and species trees
  - Better ways of combining classifications from multiple markers

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.