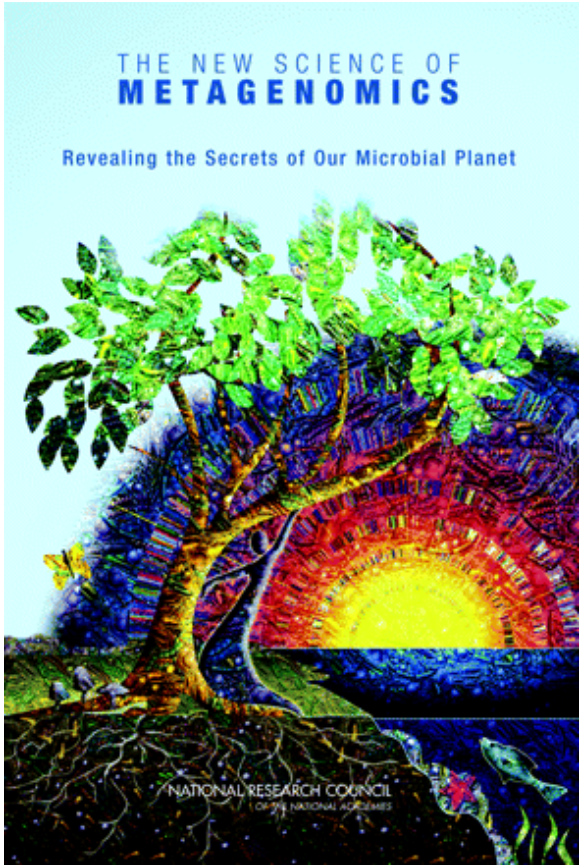
Using Ensembles of Hidden Markov Models in metagenomic analysis

Tandy Warnow, UIUC

Joint work with Siavash Mirarab, Nam-Phuong Nguyen, Mike Nute, Mihai Pop, and Bo Liu

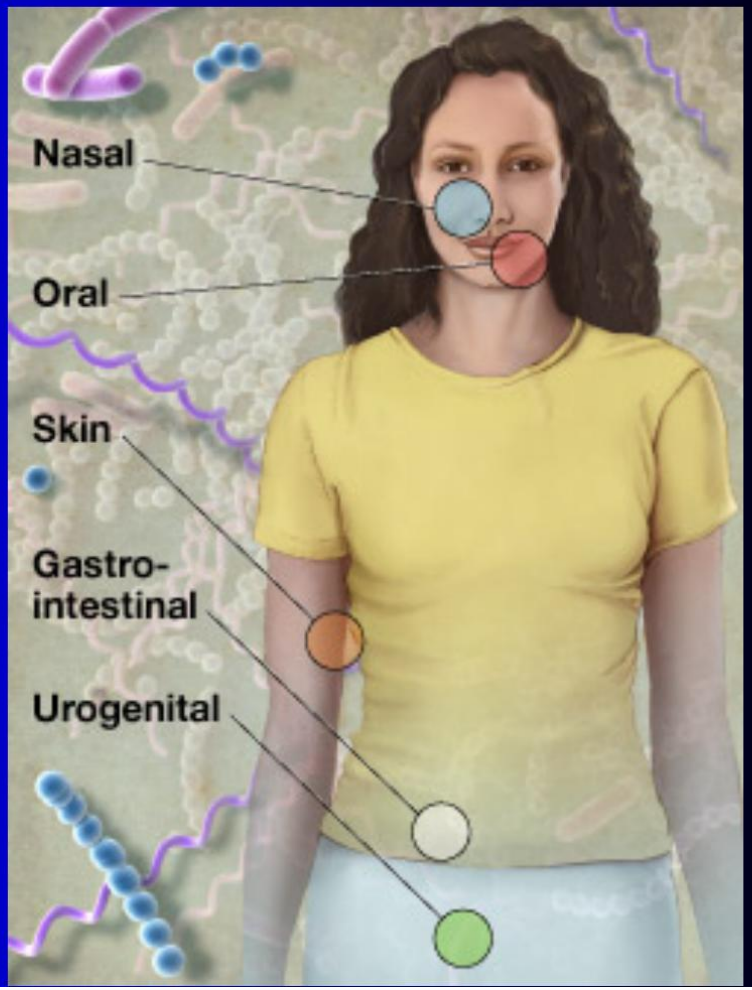
Computational Phylogenetics and Metagenomics

Courtesy of the Tree of Life project



The NIH Human Microbiome Project

25,000 human genes,
1,000,000 bacterial genes



Metagenomic data analysis

Metagenomic datasets include fragmentary sequences of unknown origin (species/gene)

Some questions:

- What taxa are present?
- In what abundances?
- What genes are present?

Four Problems

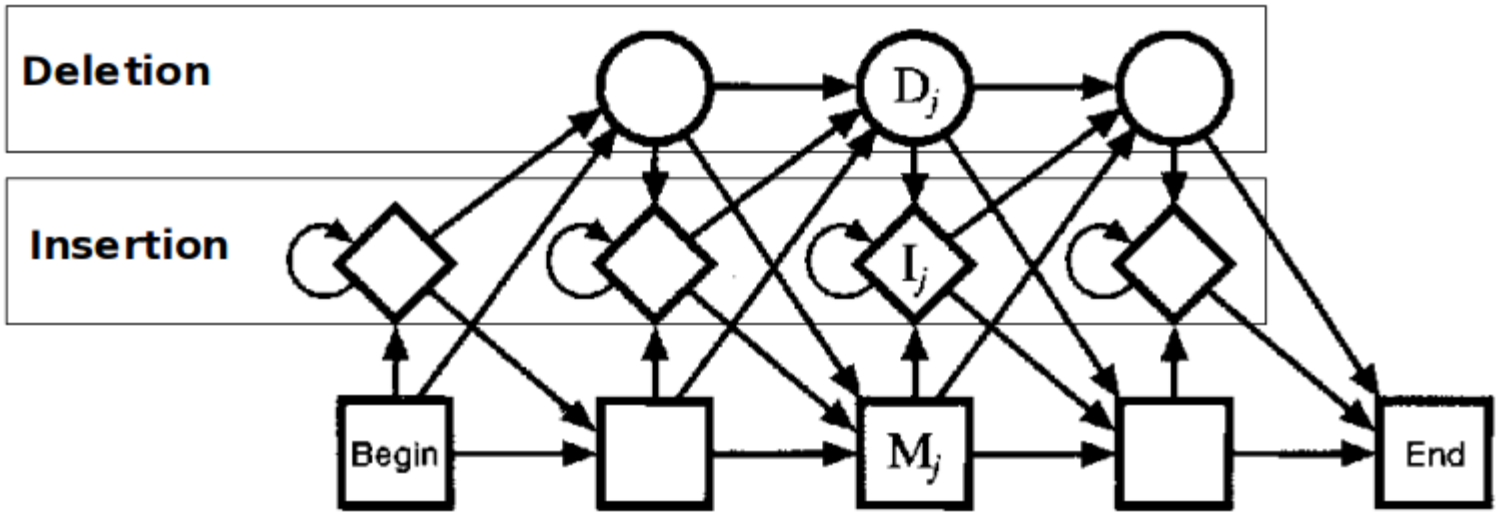
- Phylogenetic Placement (SEPP, PSB 2012)
- Multiple sequence alignment (UPP, RECOMB 2014 and Genome Biology 2014)
- Metagenomic taxon identification (TIPP, Bioinformatics 2014)
- Gene family assignment and homology detection (HIPPI, RECOMB-CG 2016 and BMC Genomics 2016)

Unifying technique: Ensemble of Hidden Markov Models (introduced in PSB 2012)

Profile HMMs

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

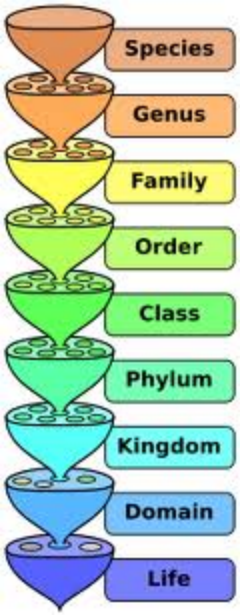
A general topology for a profile HMM

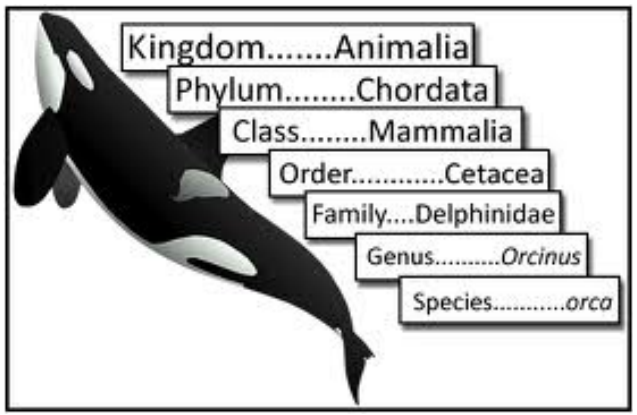


From http://codecereal.blogspot.com/2011/07/protein-profile-with-hmm.html

Taxonomic Identification

- Objective: taxonomically characterize a read





Part I: SEPP

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

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

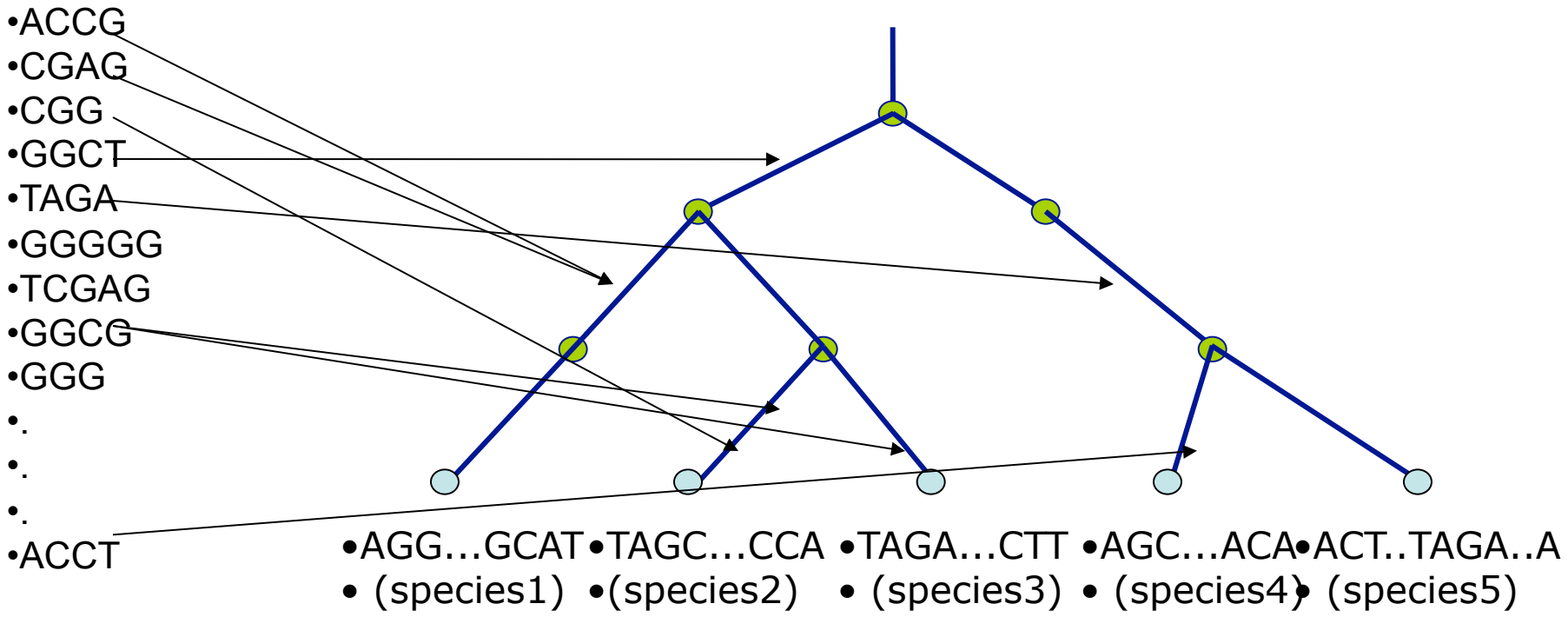
Using Phylogenetic Placement

•Fragmentary Unknown Reads:

•(60-200 bp long)

•Known Full length Sequences,
•and an alignment and a tree

•(500-10,000 bp long)



Phylogenetic Placement

Input: Backbone alignment and tree on full-length sequences, and a set of query sequences (short fragments)

Output: Placement of query sequences on backbone tree

Phylogenetic placement can be used for taxon identification, but it has general applications for phylogenetic analyses of NGS data.

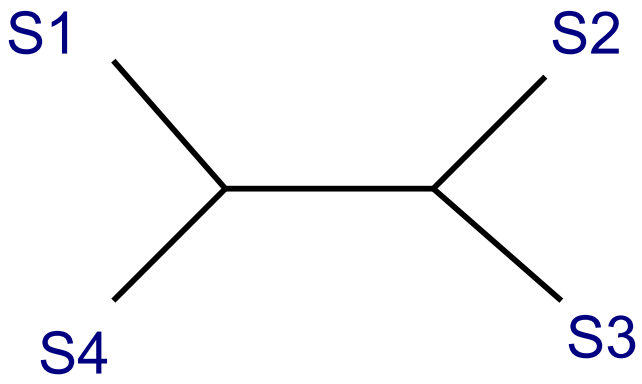
Phylogenetic Placement

- Align each query sequence to backbone alignment

- Place each query sequence into backbone tree, using extended alignment

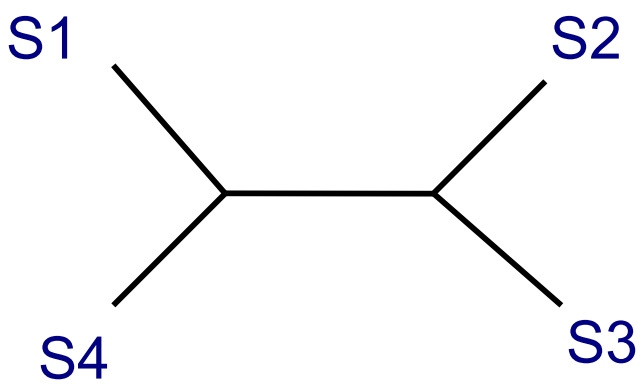
Align Sequence

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



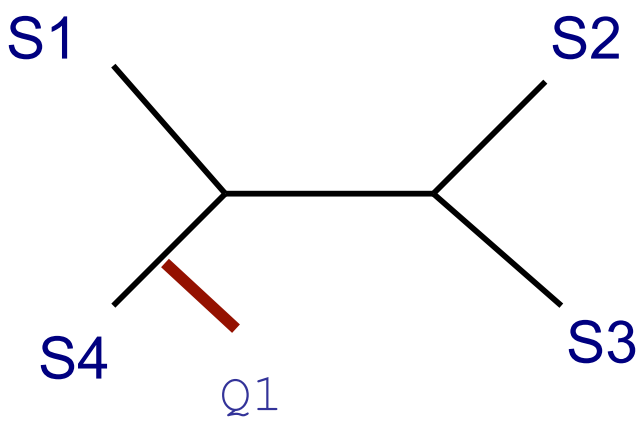
Align Sequence

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

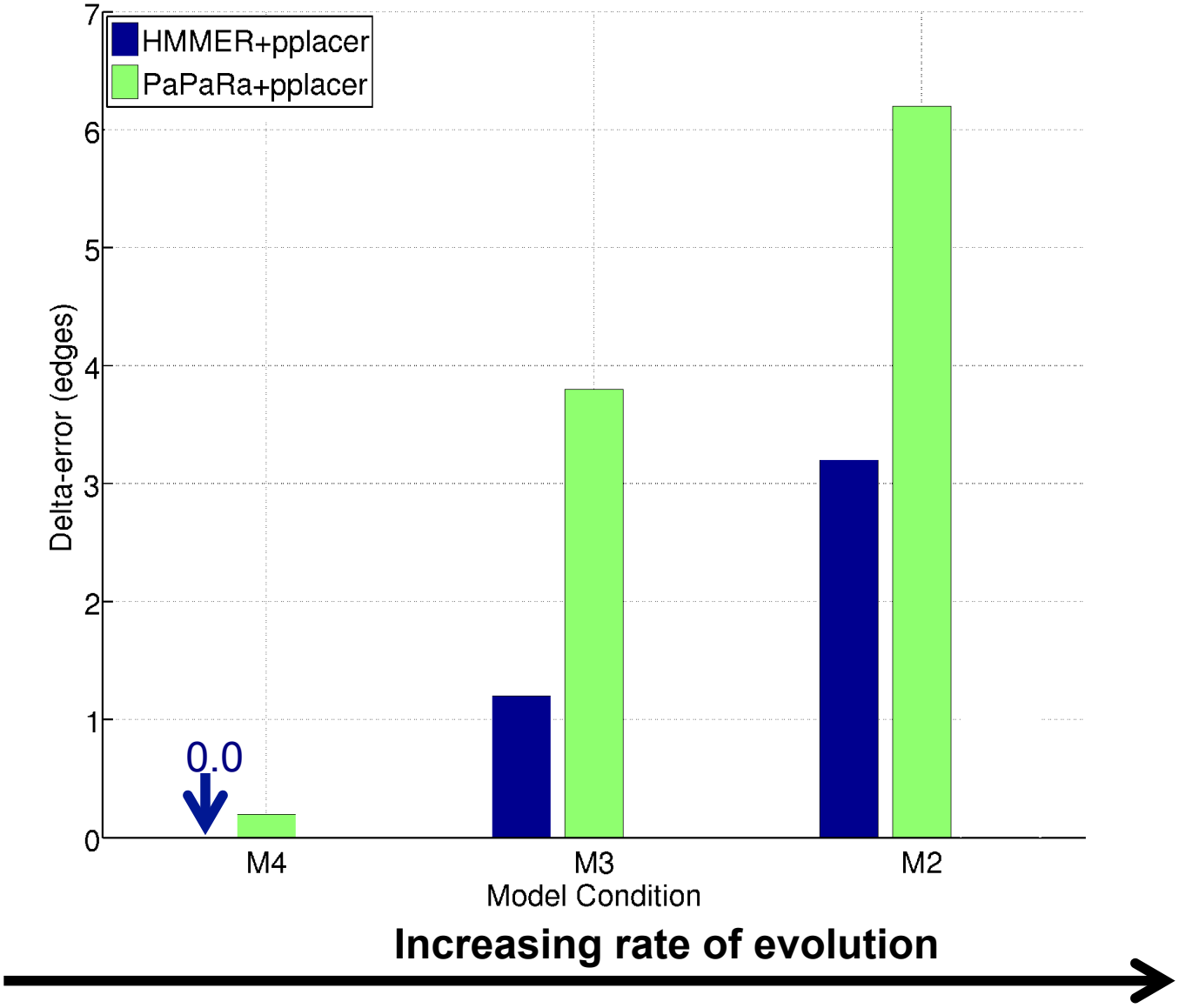


Phylogenetic Placement

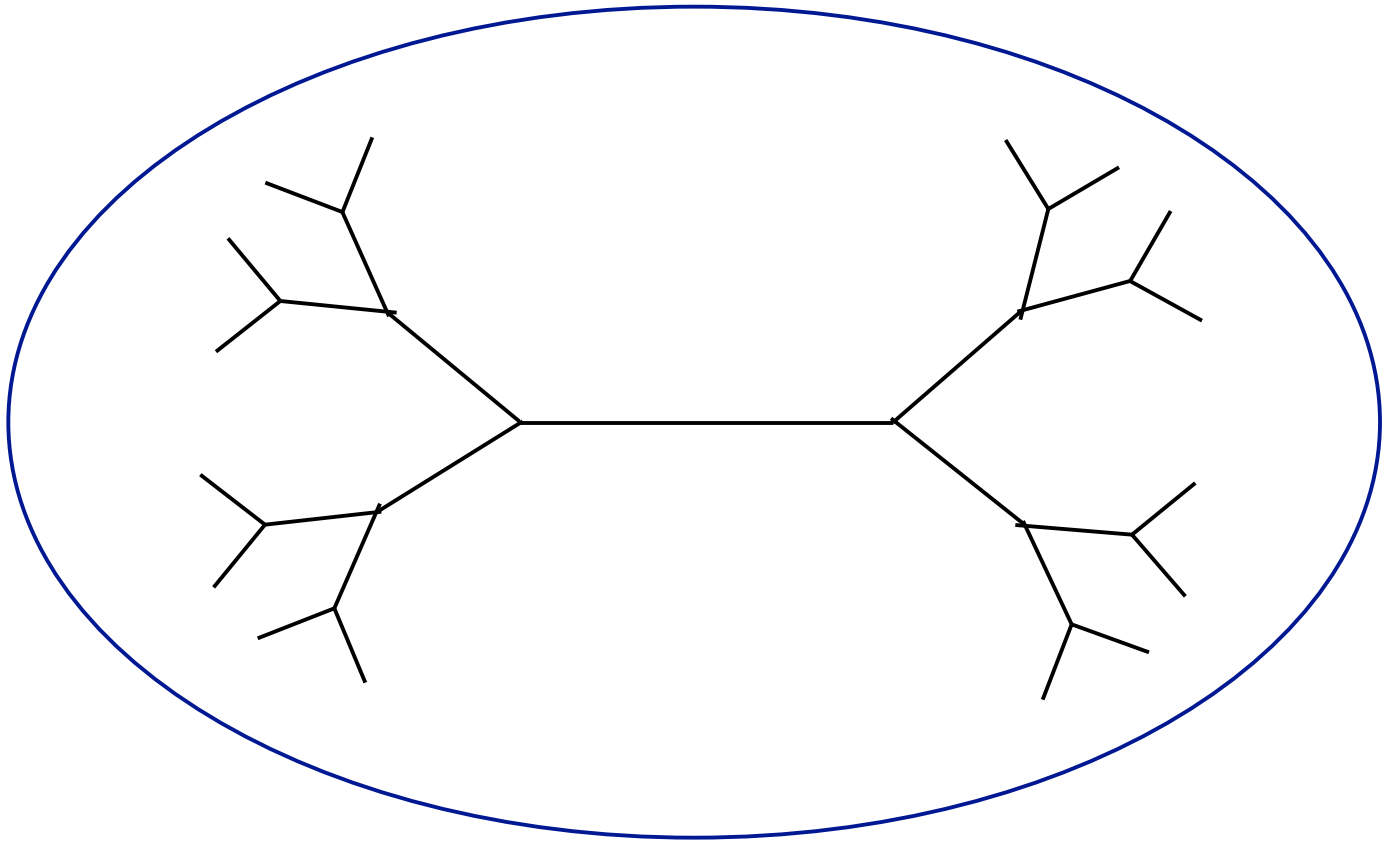
- Align each query sequence to backbone alignment
  - HMMALIGN (Eddy, Bioinformatics 1998)
  - PaPaRa (Berger and Stamatakis, Bioinformatics 2011)
- Place each query sequence into backbone tree
  - Pplacer (Matsen et al., BMC Bioinformatics, 2011)
  - EPA (Berger and Stamatakis, Systematic Biology 2011)

Note: pplacer and EPA use maximum likelihood

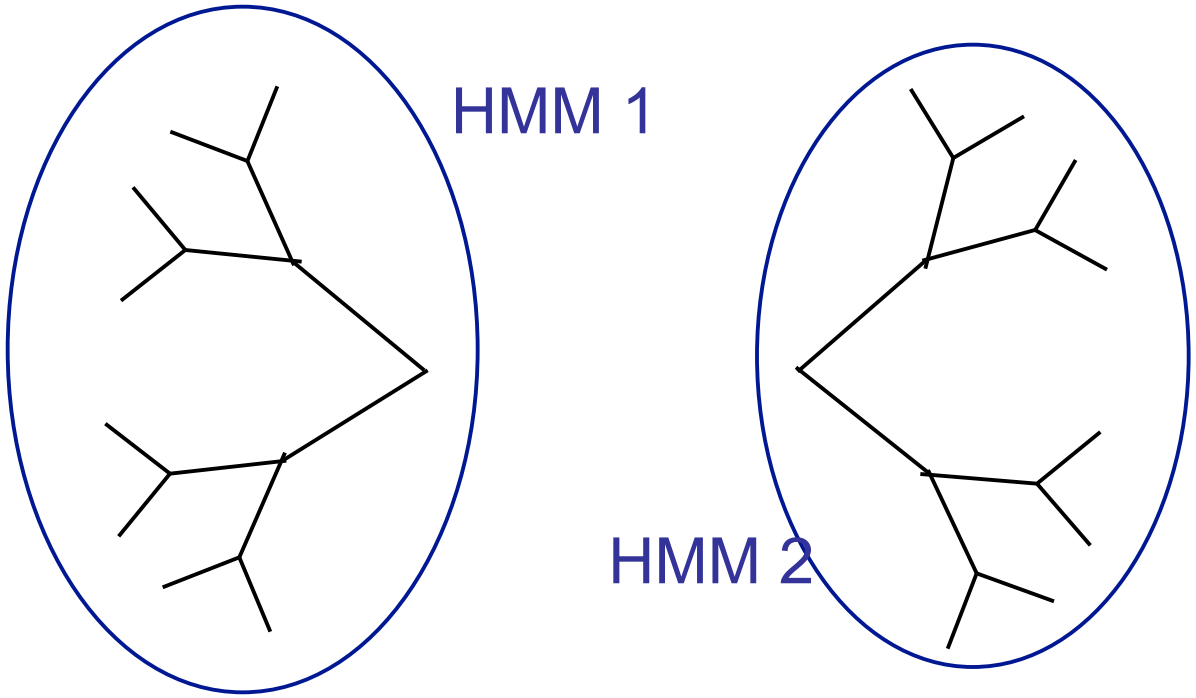
HMMER vs. PaPaRa



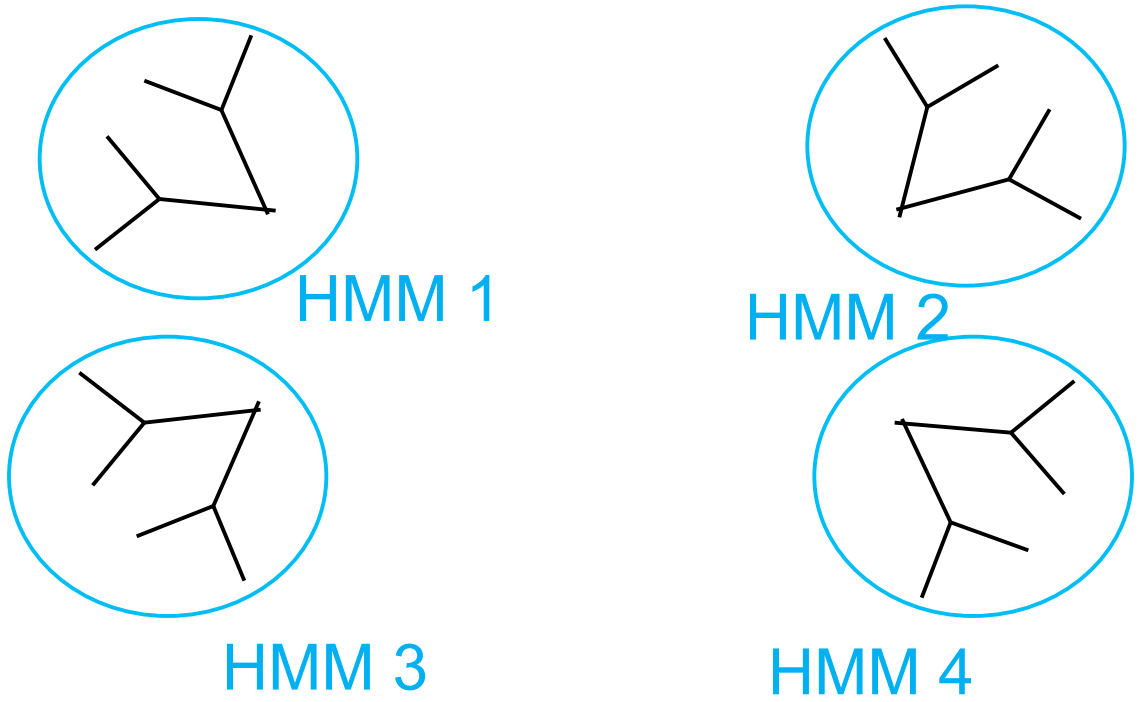
One Hidden Markov Model for the entire alignment?



Or 2 HMMs?



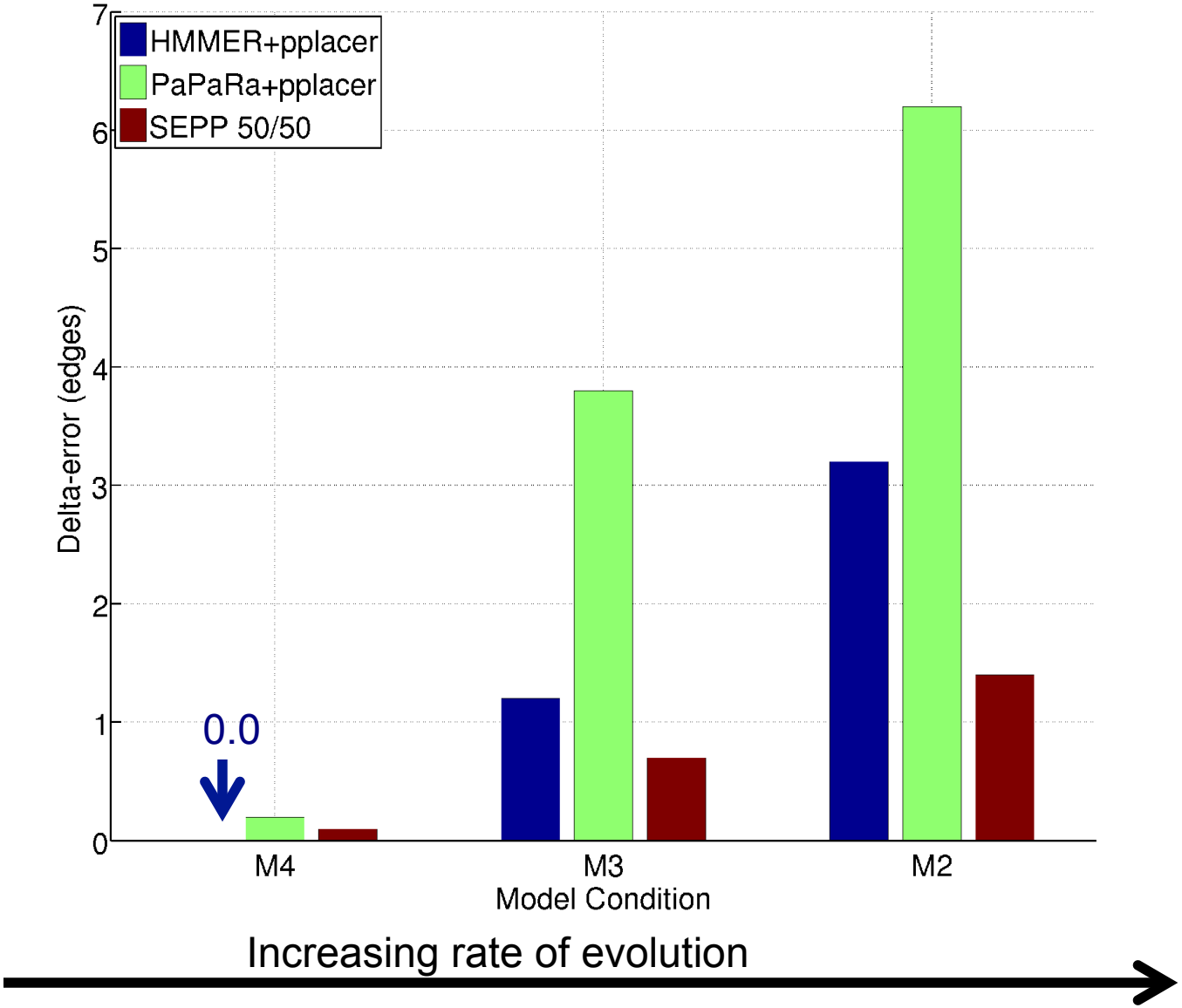
Or 4 HMMs?



SEPP Parameter Exploration

- Small subsets improved accuracy but increased running time and memory usage
- 10% rule (subset sizes 10% of backbone) had best overall performance

SEPP (10%-rule) on simulated data



Part II: 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

TIPP

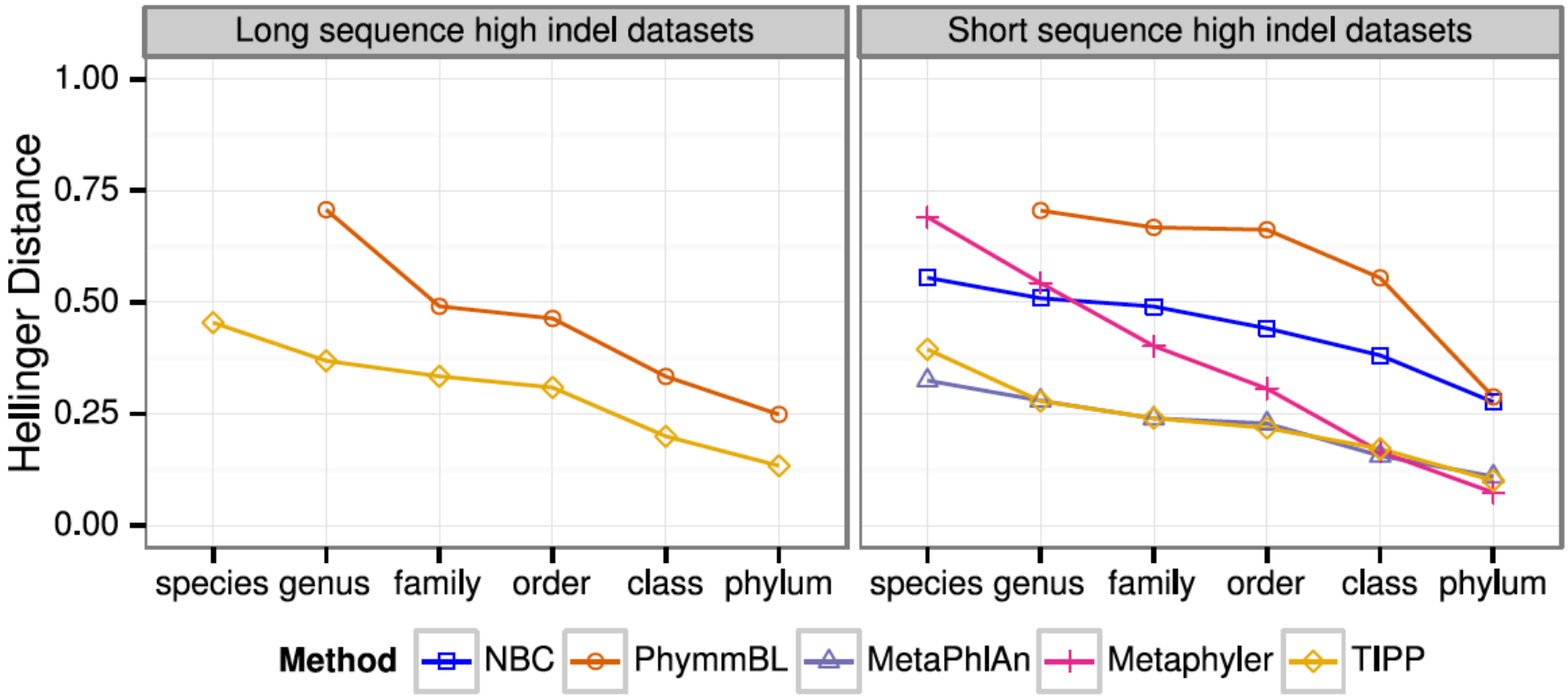
- TIPP:Taxonomic Identification and Phylogenetic Profiling.
- N. Nguyen, S. Mirarab, B. Liu, M. Pop, and T. Warnow
- Bioinformatics (2014) 30(24):3548-3555.

TIPP pipeline

Input: set of reads from a shotgun sequencing experiment of a metagenomic sample.

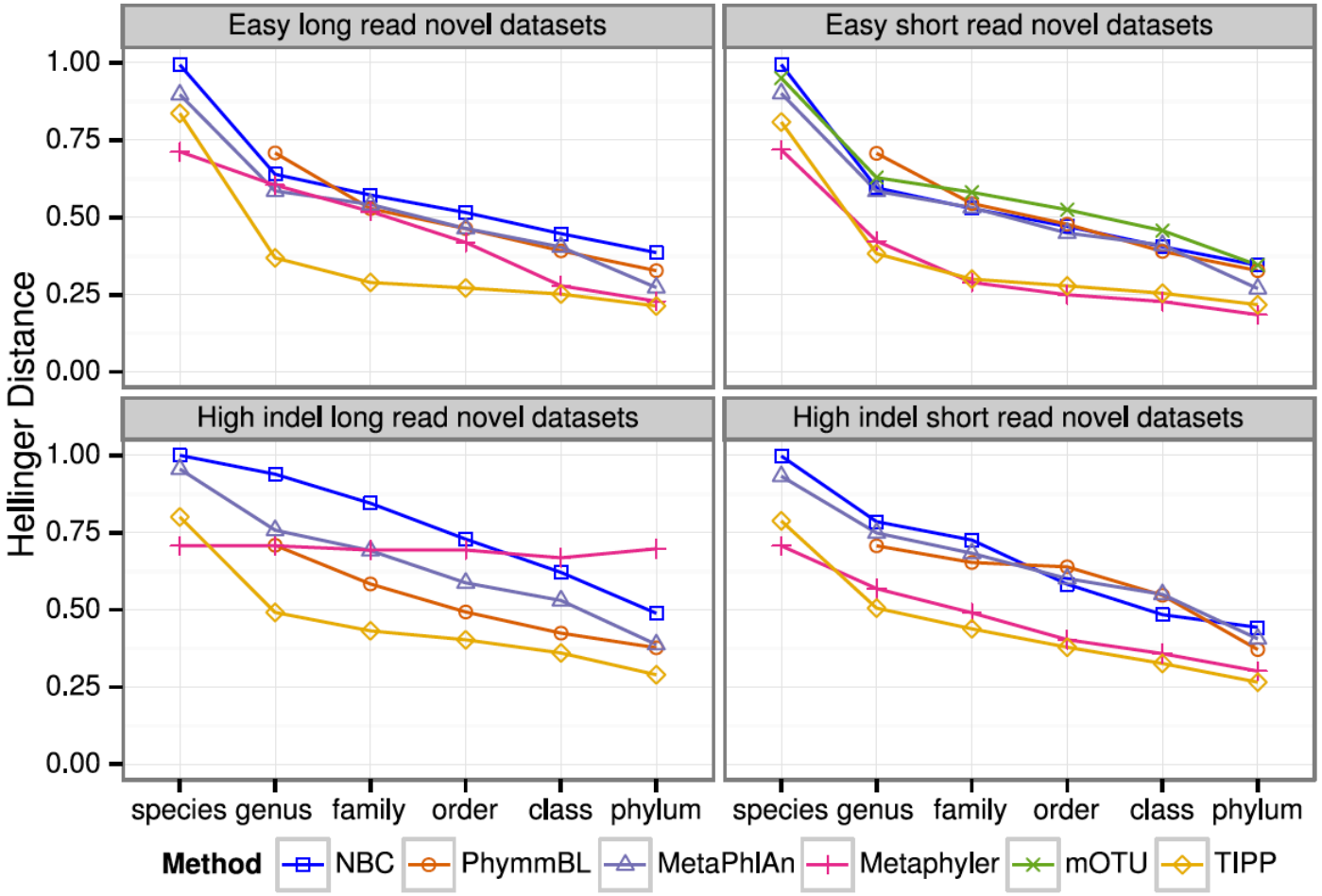
1. Assign reads to marker genes using BLAST
2. For reads assigned to marker genes, perform taxonomic analysis
3. Combine analyses from Step 2

High indel datasets containing known genomes



Note: NBC, MetaPhlAn, and MetaPhyler cannot classify any sequences from at least one of the high indel long sequence datasets, 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.

Improving TIPP

- Improve gene identification (replace BLAST?)
- Extend to other genes (not just marker genes, that are universal and single copy)
- Better reference alignments
- Better taxonomies

Improving TIPP

- Improve gene identification (replace BLAST?)
- Extend to other genes (not just marker genes, that are universal and single copy)
- Better reference alignments
- Better taxonomies

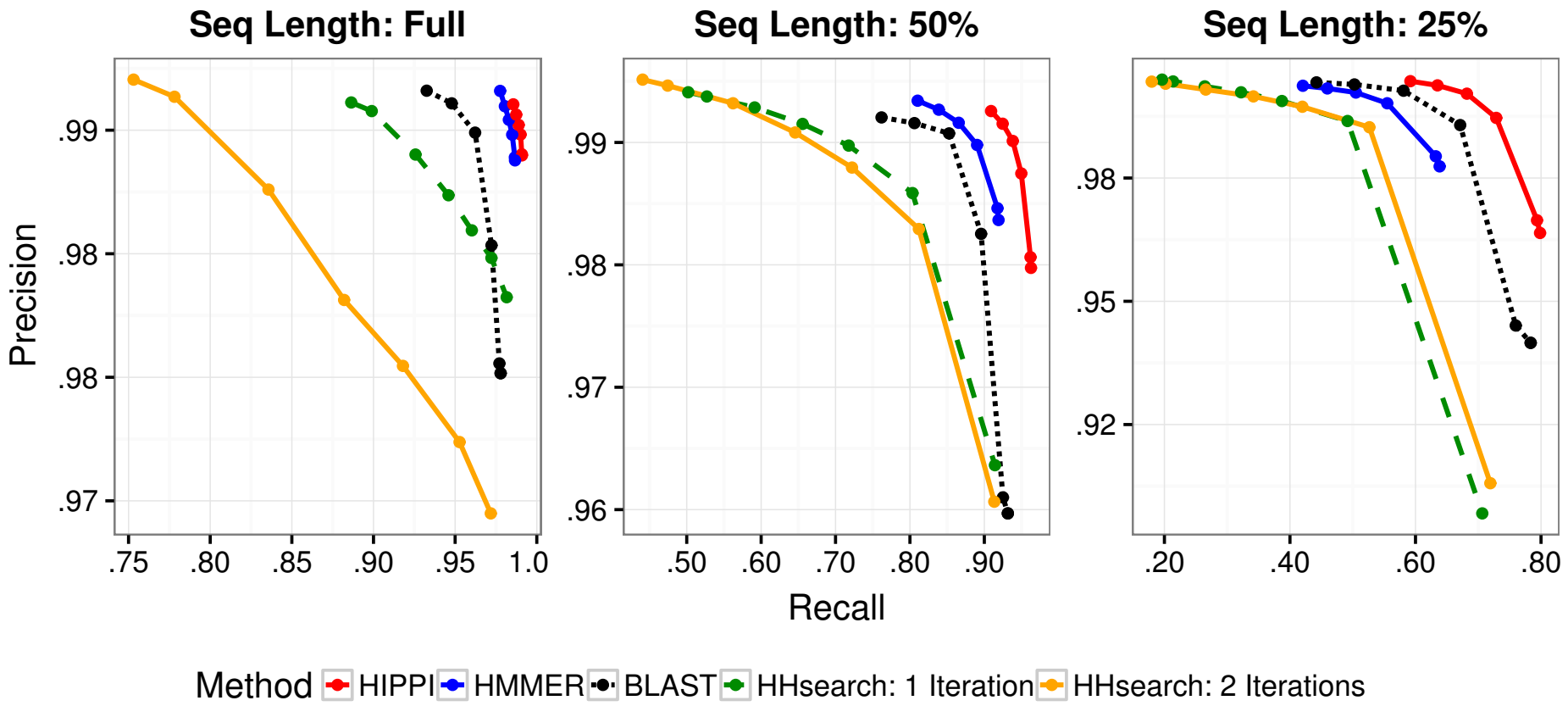
HIPPI

- HIPPI: Highly accurate protein family classification with ensembles of HMMs.
- BMC Genomics 17 (Suppl 10):765, special issue for RECOMB-CG
- Nguyen, Nute, Mirarab, and Warnow

Tested on PFAM seed sequences (i.e., proteins), but could be used on nucleotides

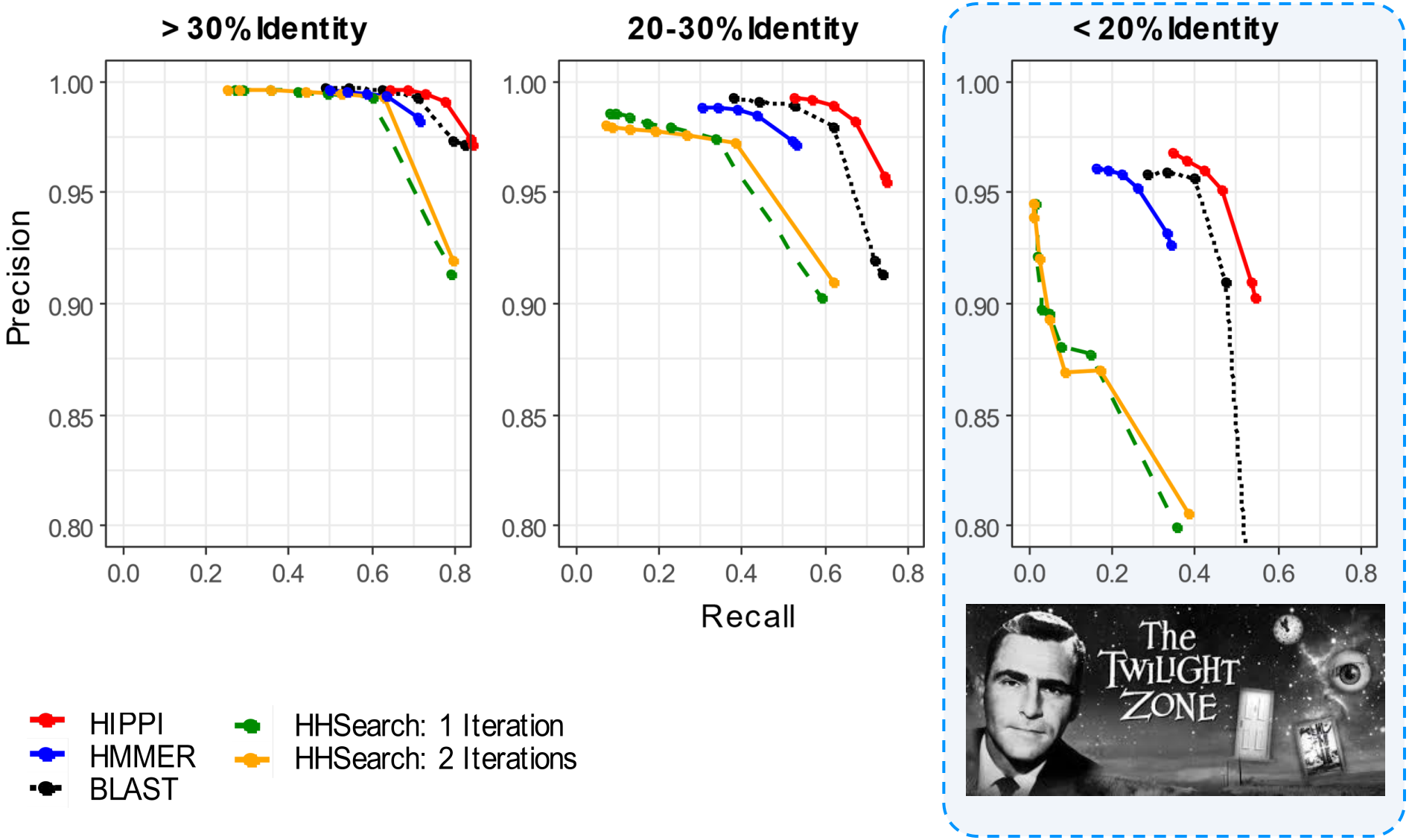
Protein Family Assignment

- 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


Seq Length: Full
Seq Length: 50%
Seq Length: 25%
Precision
.99
.98
.98
.97
.75
.80
.85
.90
.95
1.0
.99
.98
.97
.96
.50
.60
.70
.80
.90
.98
.95
.92
.20
.40
.60
.80
Recall
Method
HIPPI
HMMER
BLAST
HHsearch: 1 Iteration
HHsearch: 2 Iterations

Results in the Twilight Zone

Results on 25% Fragments



Four Problems

- Phylogenetic Placement (SEPP, PSB 2012)
- Multiple sequence alignment (UPP, RECOMB 2014 and Genome Biology 2014)
- Metagenomic taxon identification (TIPP, Bioinformatics 2014)
- Gene family assignment and homology detection (HIPPI, RECOMB-CG 2016 and BMC Genomics 2016)

A unifying technique is the "Ensemble of Hidden Markov Models" (introduced by Mirarab et al., 2012)

Github site: https://github.com/smirarab/sepp

Acknowledgments




NSF grants:

- ABI-1458652 (multiple sequence alignment)
- III:AF:1513629 (metagenomics – with Univ of Maryland)

Computational analyses performed on TACC, BlueWaters, and the University of Illinois Campus Cluster

Results on rpsB gene (60 bp)