

TIPP tutorial

Nam-phuong Nguyen

University of California, San Diego

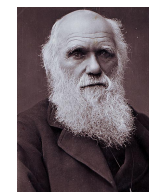
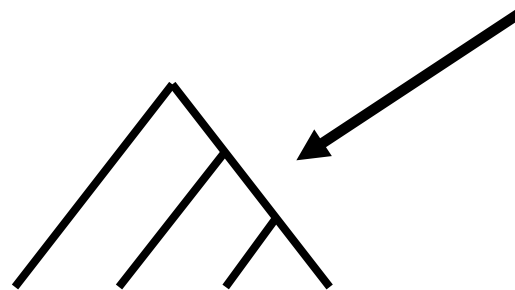
2016 Phylogenomics Symposium and Software School

Multiple Sequence Alignment (MSA)

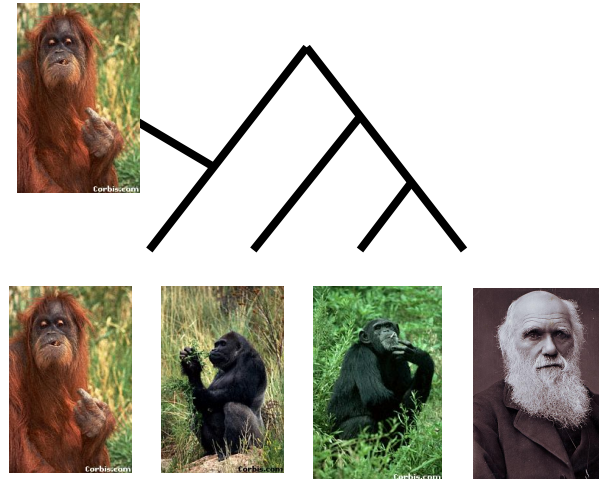
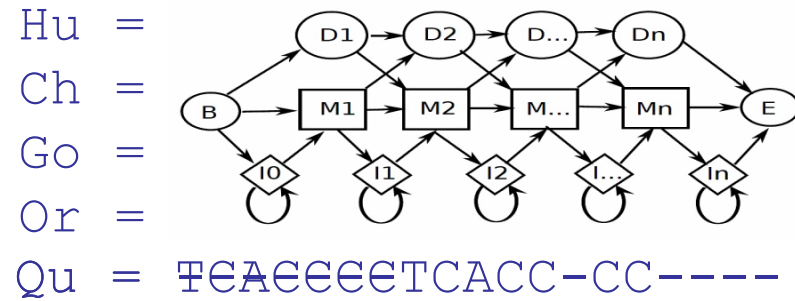
Hu = AGGCTATCACCGACTCCA
Ch = TAGCTATCACGACCGC
Go = TAGCTGACCGC
Or = TCACGACCGACA



Hu = -AGGCTATCACGACCTCCA
Ch = TAG-CTATCACGACCGC--
Go = TAG-CT-----GACCGC--
Or = -----TCACGACCGACA



Using the MSA and tree to identify reads



Represent MSA using a
profile Hidden Markov Model (HMM)

Outline

- Start a profiling job
- Inputs/outputs
- Parameter options

Before we begin....

- `cd /path/to/sepp/` (on the VM, the path is `~/tools/sepp/`)
- `chmod u+x run_tipp_tool.py`
- `export PATH=$PATH:/path/to/sepp/`

Running an initial profiling job

- First move to the test directory located within the sepp directory
 - On the VM type
 - `cd ~/tools/sepp/test/unittest/data/mock/pyrg`
 - All others
 - `cd /path/to/sepp/test/unittest/data/mock/pyrg`
- Run
 - `run_tipp.py -R pyrg -f pyrg.even.fas -o output`

Output

- Alignment files
 - output_alignment.fasta
 - output_alignment_masked.fasta
- Placement file
 - output_placement.json

Output (continued)

- Classification file

output_classification.txt

```
EAS25_26_1_15_381_1761_0_1,2157,Archaea,superkingdom,1.0000
EAS25_26_1_15_381_1761_0_1,1,root,root,1.0000
EAS25_26_1_15_381_1761_0_1,183925,Methanobacteria,class,1.0000
EAS25_26_1_15_381_1761_0_1,2172,Methanobrevibacter,genus,1.0000
EAS25_26_1_15_381_1761_0_1,28890,Euryarchaeota,phylum,1.0000
EAS25_26_1_15_381_1761_0_1,2158,Methanobacteriales,order,1.0000
EAS25_26_1_15_381_1761_0_1,2173,Methanobrevibacter smithii,species,1.0000
```


Converting classification format

- Type:
 - mkdir profile
 - `run_tipp_tool.py -g pyrg -a profile -o profile -p pyrg -i output_classification.txt -t 0.95`

Tab separated classification file

pyrg.classification

fragment	species	genus	family	order	class	phylum
EAS25...	2173	2172	2159	2158	183925	28890
EAS25...	NA	NA	2159	2158	183925	28890

Abundance profile

abundance.species.csv

taxa abundance

Methanobrevibacter smithii 0.7969

Methanococcus maripaludis 0.0156

unclassified 0.1875

Profiling a metagenomic dataset

- Change to the mixed test directory
 - `cd ../mixed`

- Run the abundance profiler
 - `run_abundance.py -f facts_simhc.short.fas -c ~/.sepp/tipp.config -d out`

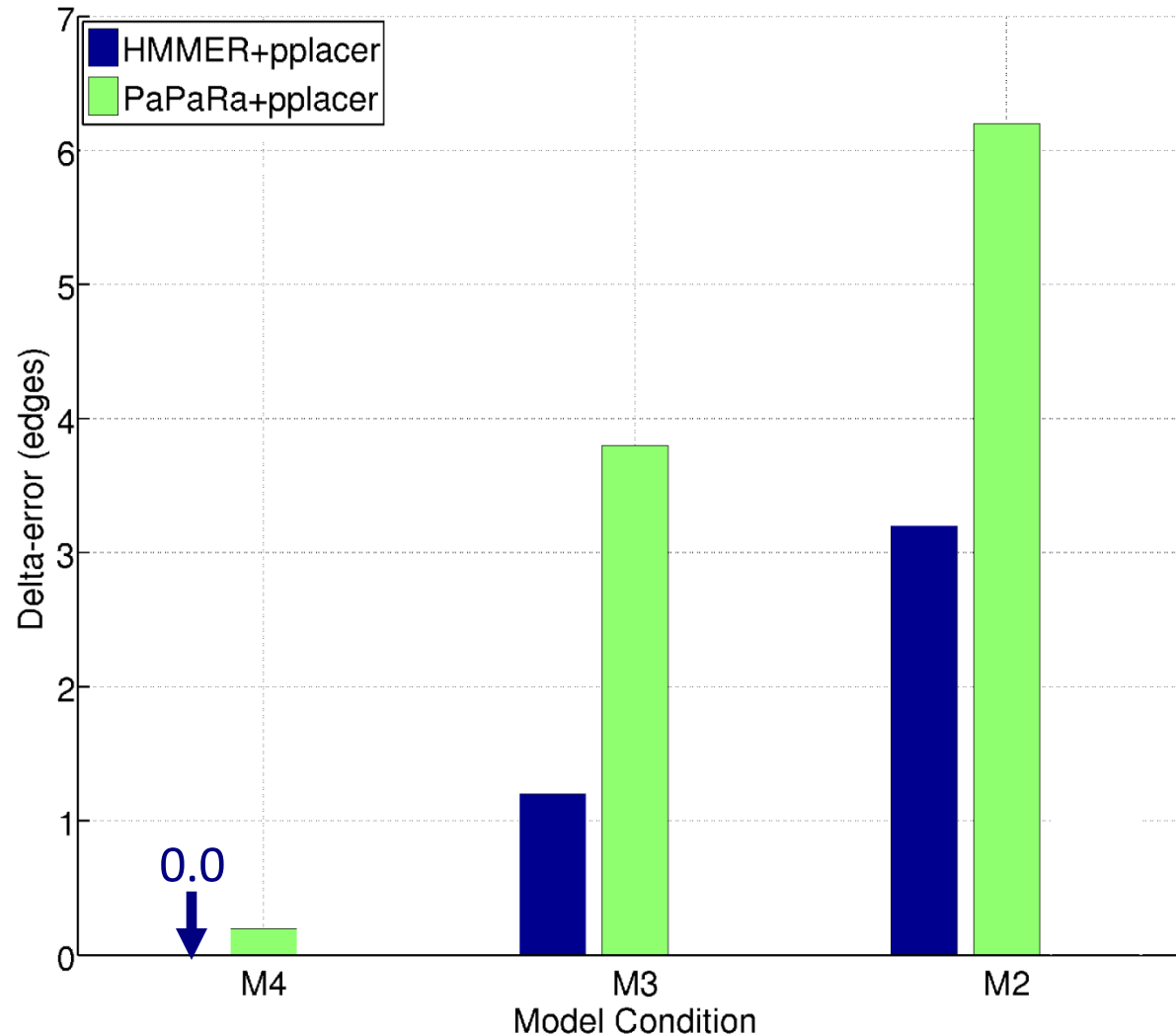
Phylogenetic Placement

- Align each query sequence to backbone alignment:
 - HMMALIGN (Eddy, Bioinformatics 1998)
 - PaPaRa (Berger and Stamatakis, Bioinformatics 2011)
- Place each query sequence into backbone tree, using extended alignment:
 - pplacer (Matsen et al., BMC Bioinformatics 2010)
 - EPA (Berger et al., Systematic Biology 2011)

Phylogenetic Placement

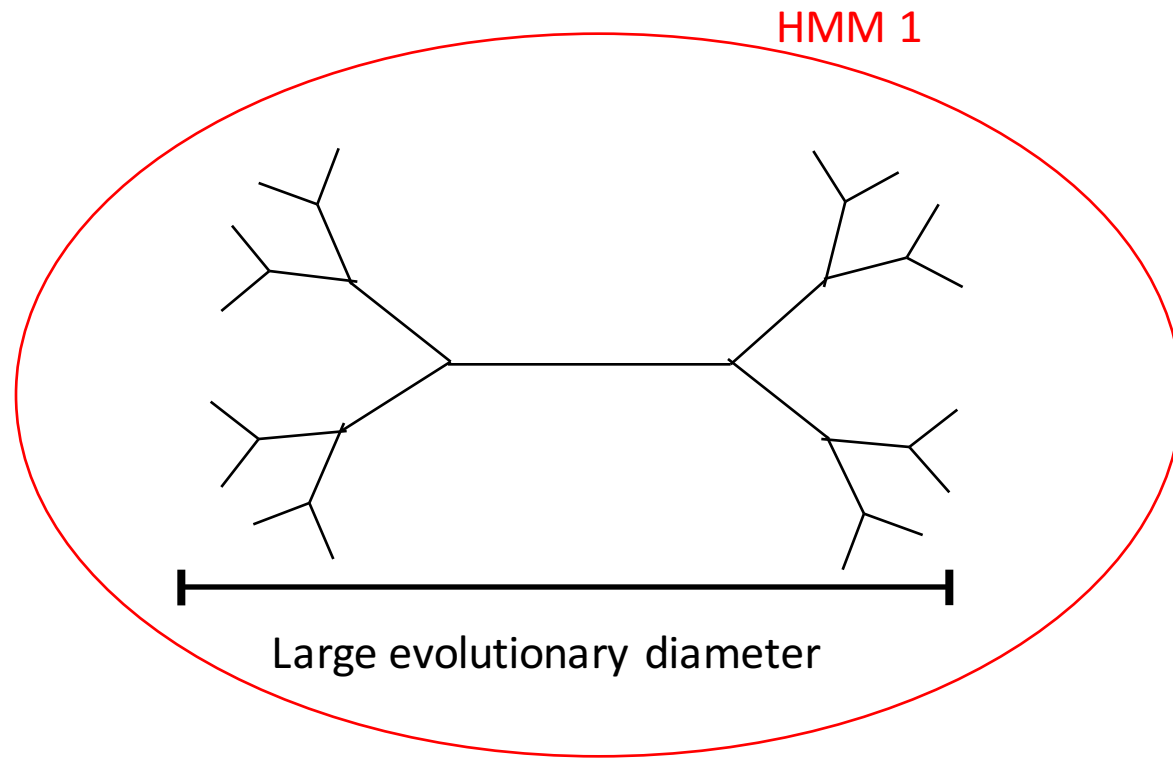
- Align each query sequence to backbone alignment:
 - **HMMALIGN** (Eddy, Bioinformatics 1998)
 - **PaPaRa** (Berger and Stamatakis, Bioinformatics 2011)
- Place each query sequence into backbone tree, using extended alignment:
 - **pplacer** (Matsen et al., BMC Bioinformatics 2010)
 - EPA (Berger et al., Systematic Biology 2011)

HMMER and PaPaRa results

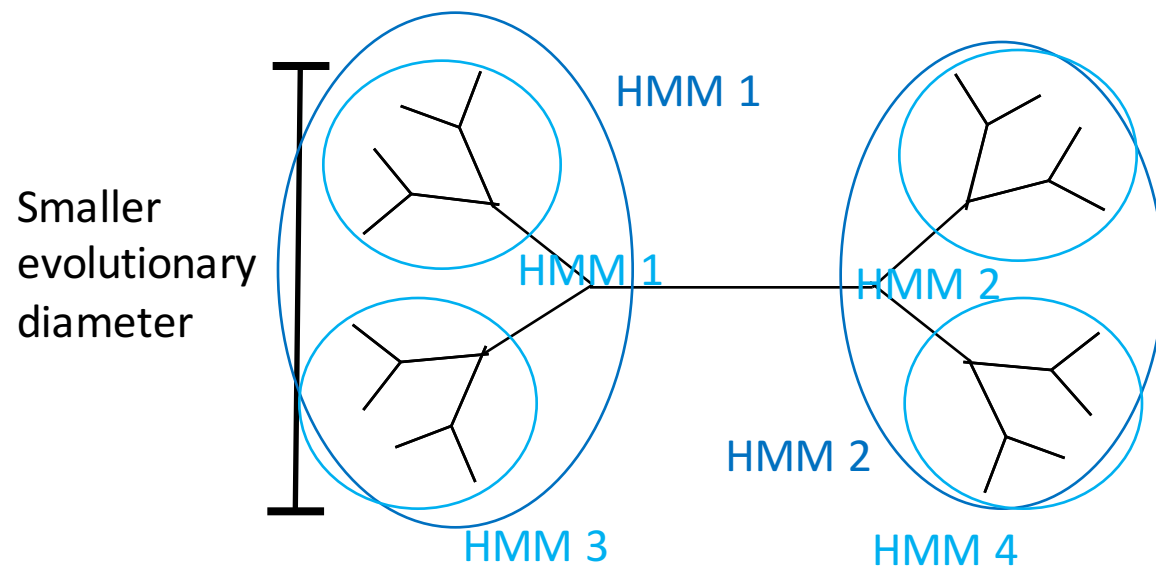


Backbone size: 500
5000 fragments
20 replicates

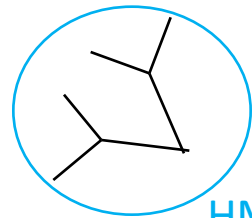
Standard approach (single HMM)



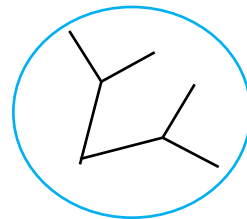
New approach



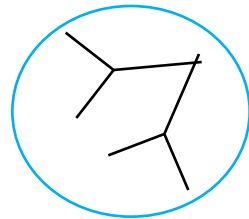
Ensemble of HHMs (eHMMs)



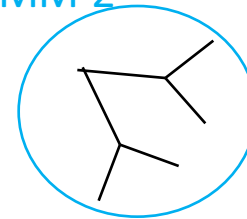
HMM 1



HMM 2

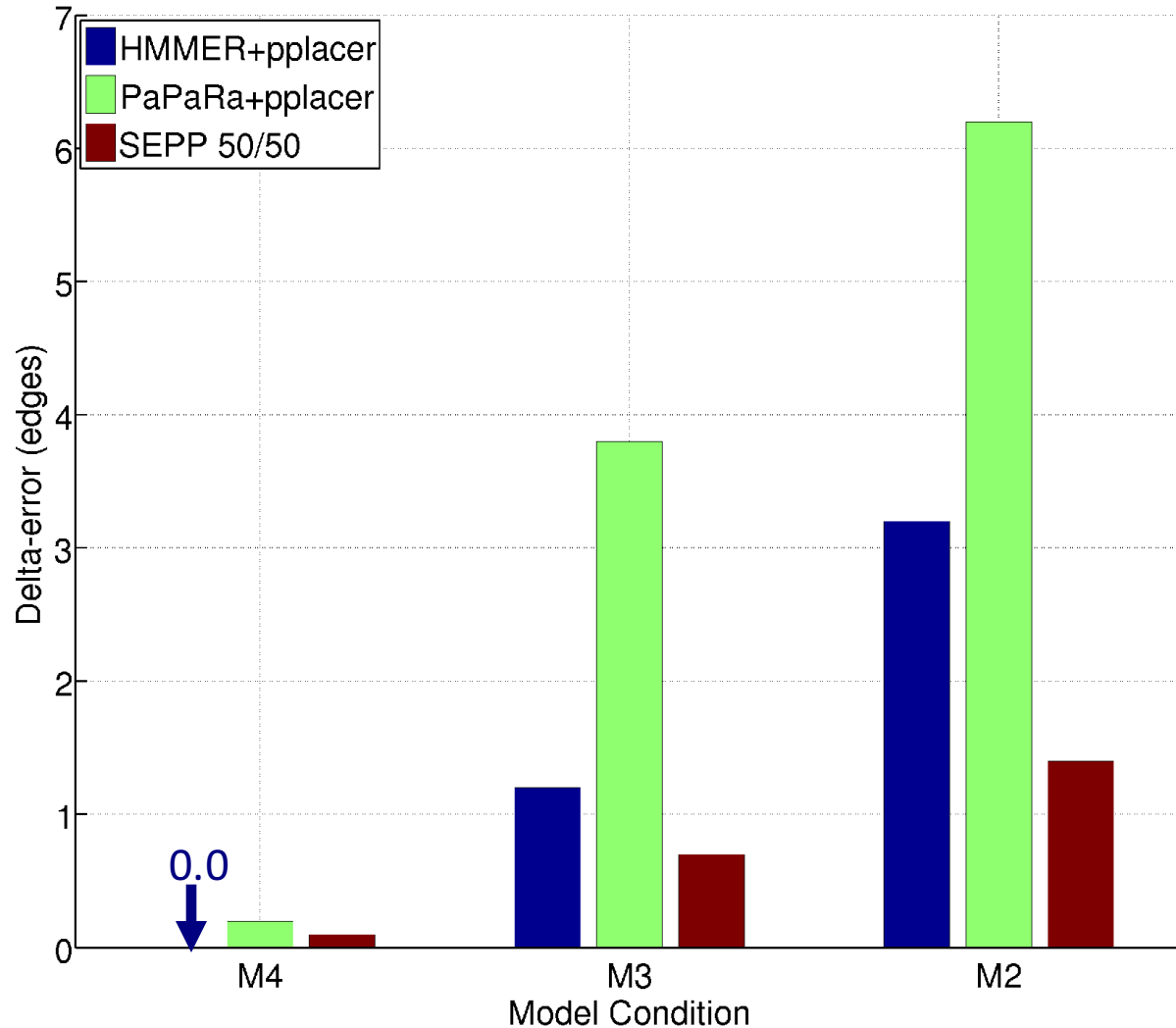


HMM 3



HMM 4

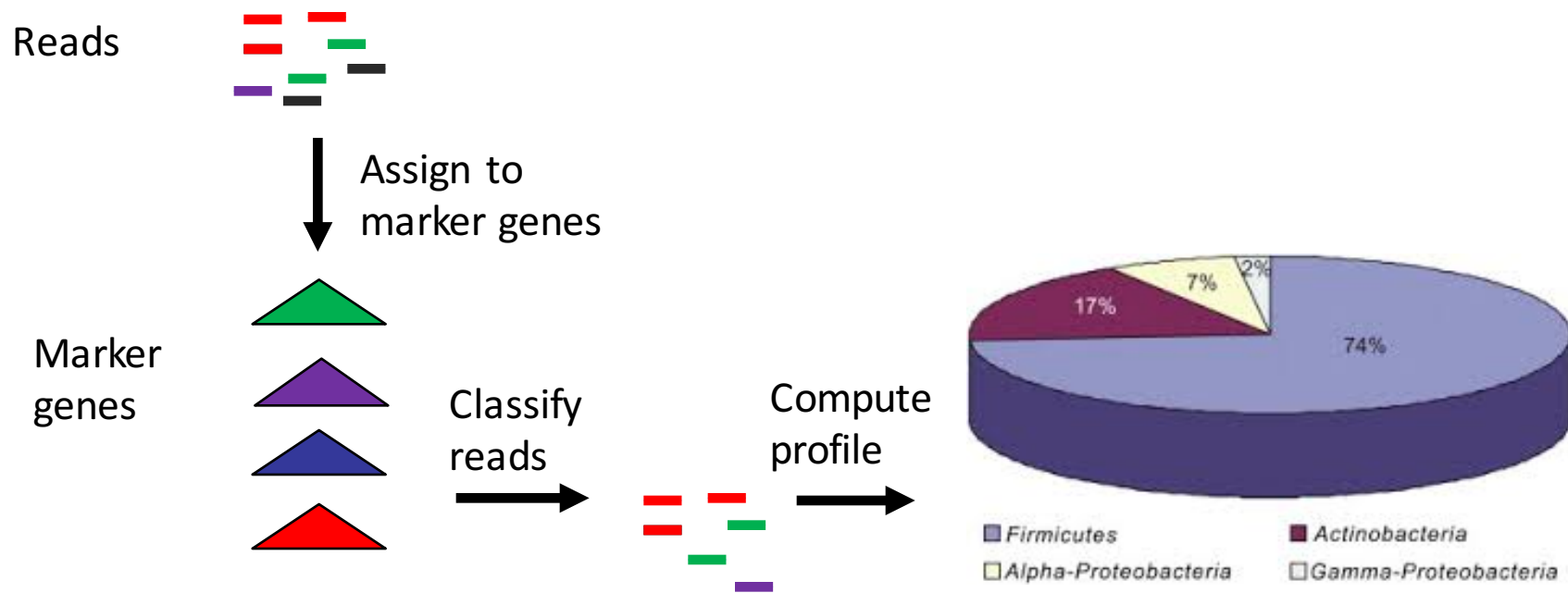
SEPP (10% rule) Simulated Results



Backbone size: 500
5000 fragments
20 replicates

TIPP: Taxonomic Identification And Phylogenetic Profiling

- Nguyen et al., Bioinformatics, 2014



Reference package

- Alignment (sate.fasta)
- Refined taxonomic tree (sate.taxonomy)
- Rate parameters (sate.taxonomy.RAxML_info)
- Taxonomy file (all_taxon.taxonomy)
- Sequence mapping (species.mapping)

Looking at results

- Pooled abundance profiles

- Results for each individual marker

16S amplicon analysis

- Change to the 16S_bacteria test directory
 - `cd ../16S_bacteria`

- Run TIPP on the gut data
 - `run_tipp.py -R 16S_bacteria -f human_gut_16S.fas -o 16s -A 1000 -P 1000`