

Phylogenetic methods

Nidhi Shah
M³ Workshop
January 10, 2019

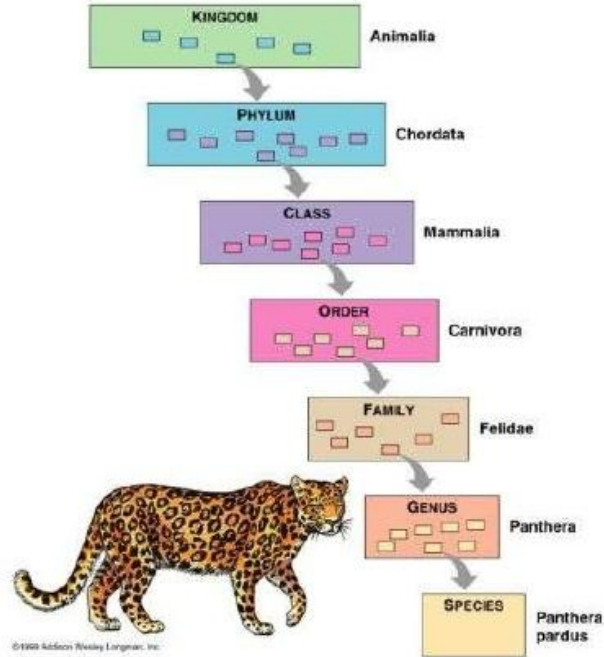
Basic Questions

1. What is this sequence? (Taxonomic classification)
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?

This Talk

- Basics
 - What are taxonomies and phylogenies
 - Phylogeny estimation problem
 - Phylogenetic placement and taxon ID
 - Ensembles of Hidden Markov models (eHMMs)
- TIPP (Bioinformatics 2014): Application of eHMMs to a) Taxon Identification and b) abundance profiling

What Are Taxonomies?



<https://www.slideshare.net/EastBayWPMeetup/custom-post-types-and-custom-taxonomies>

Phylogenies and Taxonomies

Taxonomies:

- Rooted, labels at every node for each taxonomic level

- More or less based on phylogenies

Phylogenies:

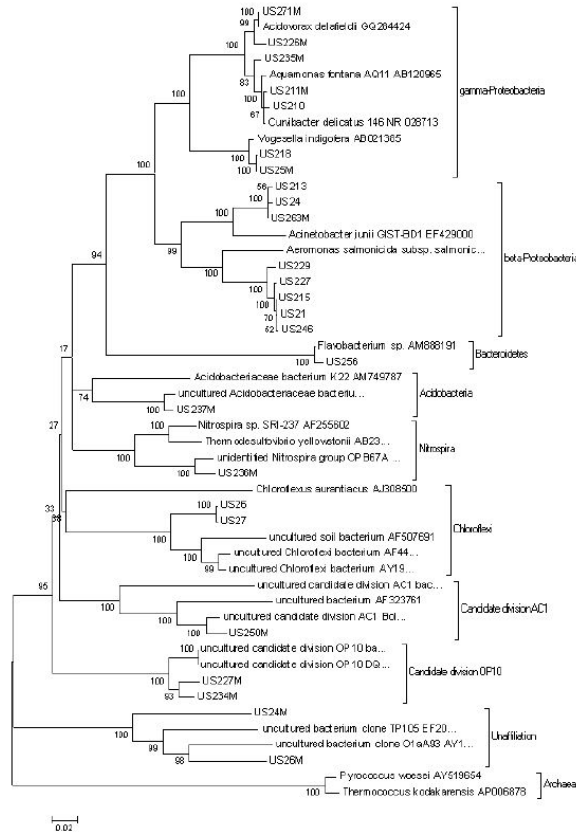
- Estimated from sequences (usually)

- Branch lengths reflect amount of change

- Edges/nodes sometimes given with support (typically bootstrap)

- Phylogenies are usually unrooted (because root can be difficult to infer)

Rooted neighbor-joining 16S rRNA gene-based phylogenetic tree of uncultured bacteria.



https://www.researchgate.net/Rooted-neighbor-joining-16S-rRNA-gene-based-phylogenetic-tree-of-uncultured-bacteria-The_fig1_279565247 [accessed 26 Jul, 2018]

This Talk

- Basics
 - What are taxonomies and phylogenies
 - Phylogeny estimation problem
 - Phylogenetic placement and taxon ID
 - Ensembles of Hidden Markov models (eHMMs)
- TIPP (Bioinformatics 2014): Application of eHMMs to a) Taxon Identification and b) abundance profiling

How are Phylogenies Estimated?

Input: Sequences (DNA, RNA, or Aminoacid), unaligned

Output: Tree on the sequences

Notes:

Two steps: first align, then compute a tree on the alignment

Many different techniques for each step

Input: unaligned sequences

S1 = AGGCTATCACCTGACCTCCA

S2 = TAGCTATCACGACCGC

S3 = TAGCTGACCGC

S4 = TCACGACCGACA

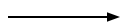
Phase 1: Alignment

S1 = AGGCTATCACCTGACCTCCA

S2 = TAGCTATCACGACCGC

S3 = TAGCTGACCGC

S4 = TCACGACCGACA



S1 = -AGGCTATCACCTGACCTCCA

S2 = TAG-CTATCAC--GACCGC--

S3 = TAG-CT-----GACCGC--

S4 = -----TCAC--GACCGACA

Phase 2: Construct tree

S1 = AGGCTATCACCTGACCTCCA

S2 = TAGCTATCACGACCGC

S3 = TAGCTGACCGC

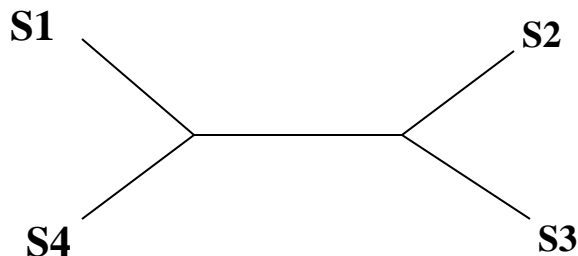
S4 = TCACGACCGACA

S1 = -AGGCTATCACCTGACCTCCA

S2 = TAG-CTATCAC--GACCGC--

S3 = TAG-CT-----GACCGC--

S4 = -----TCAC--GACCGACA



Phylogeny estimation

First compute a multiple sequence alignment

Computationally challenging, accuracy on large divergent datasets can be low

Best current methods for moderate-sized DNA datasets: MAFFT

Best current methods for large datasets: PASTA and UPP

Then compute a tree

Typically using maximum likelihood heuristics (e.g., RAxML, IQTree, PhyML, and FastTree)

Computationally challenging, accuracy on large divergent datasets can be low

Produces trees with branch lengths and other numeric model parameters

This Talk

- Basics
 - What are taxonomies and phylogenies
 - Phylogeny estimation problem
 - Phylogenetic placement and taxon ID
 - Ensembles of Hidden Markov models (eHMMs)
- TIPP (Bioinformatics 2014): Application of eHMMs to a) Taxon Identification and b) abundance profiling

Phylogenetic Placement

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

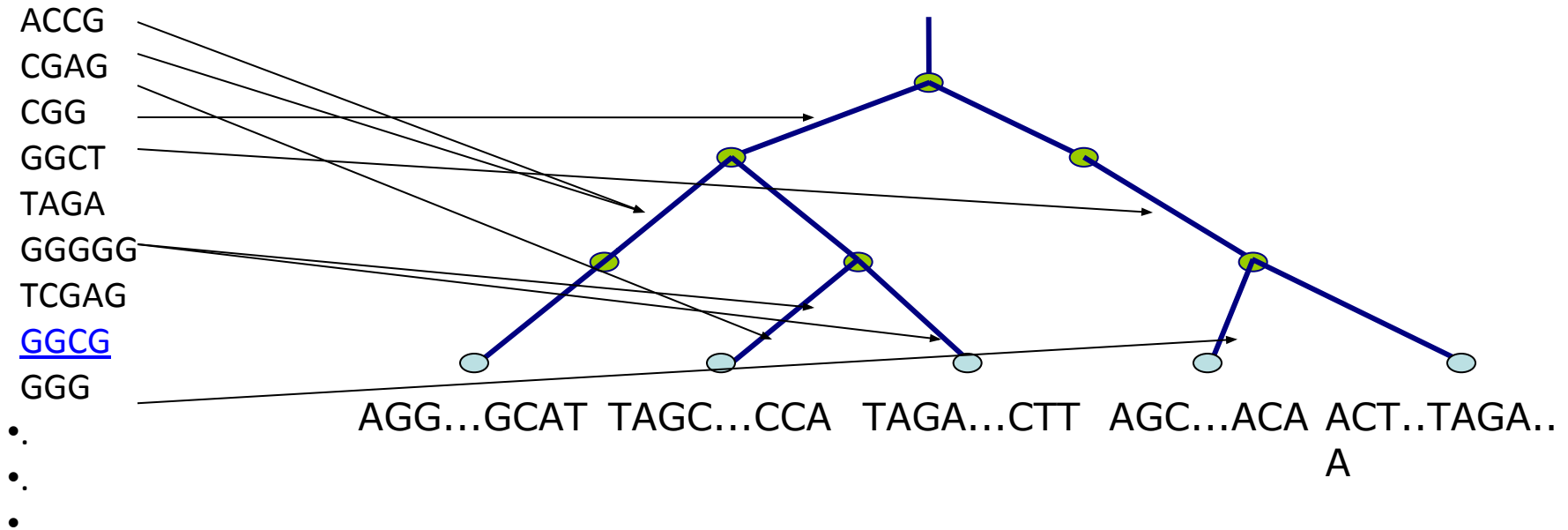
Output: Placement of query sequences on backbone tree

Note: if the backbone tree is a Taxonomy, then the placement gives taxonomic information about the query sequences (i.e., reads)!

Marker-based Taxon Identification

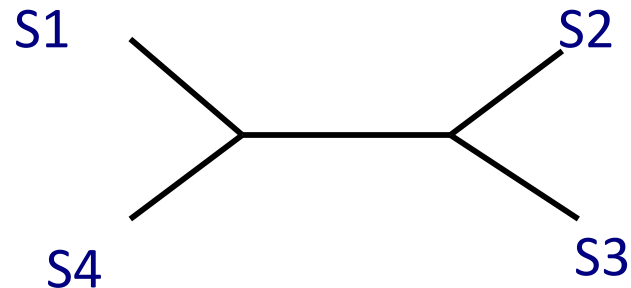
Fragmentary sequences
from some gene

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



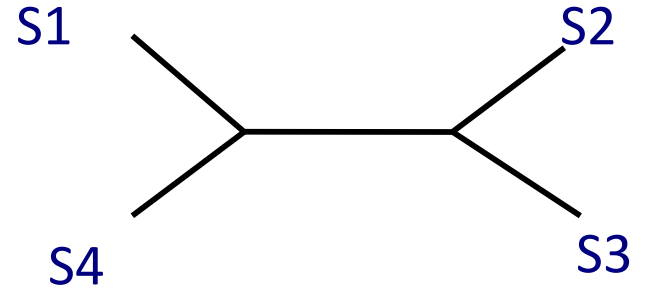
Input

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



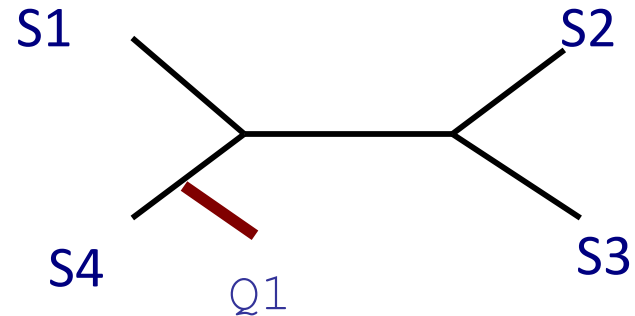
Align Sequence

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



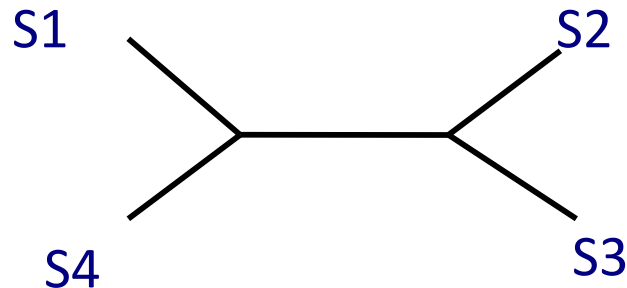
Place Sequence

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



Align Sequence using HMMER

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



SafariFileEditViewHistoryBookmarksWindowHelp

warlow-msa-ucsd-v3.ppt

Search in Presentation

hmmerr.org

Warnow, Tandy - ...PEEC seminar 20...American Airline...Gene Filtering M...Erin -- Literature...Inbox (2,341) - ta...The New York TI...Sean R. Eddy - G...HMMERHow to take a sc...

warlow-msa-ucsd-v3.ppt

Search in Presentation

hmmerr.org


Warnow, Tandy - ...PEEC seminar 20...American Airline...Gene Filtering M...Erin -- Literature...Inbox (2,341) - ta...The New York TI...Sean R. Eddy - G...HMMERHow to take a sc...

warlow-msa-ucsd-v3.ppt

Search in Presentation

hmmerr.org

Warnow, Tandy - ...PEEC seminar 20...American Airline...Gene Filtering M...Erin -- Literature...Inbox (2,341) - ta...The New York TI...Sean R. Eddy - G...HMMERHow to take a sc...



DOWNLOADDOCUMENTATIONSEARCHPUBLICATIONSBLOG

HMMER: biosequence analysis using profile hidden Markov models

Get the latest version

v3.1b2

Download (MacOSX / Intel)

Alternative Download Options

PERFORM A SEARCH

An online interactive [search](#) service is available at the European Bioinformatics Institute. Go there to [search](#) against the latest Uniprot databases.

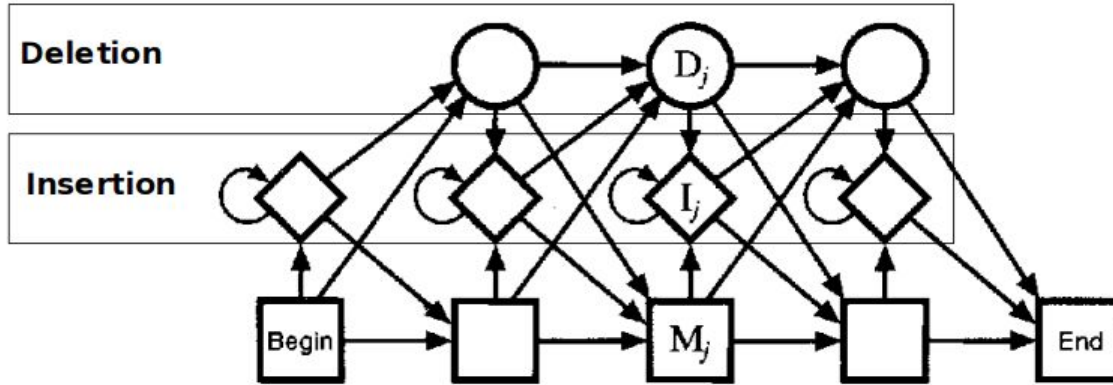
DOCUMENTATION

The HMMER User's Guide: [\[PDF, 119 pages\]](#).
[Release notes](#) for the current release.

NEWS

See the blog [Cryptogenomicon](#) for more information and discussion about HMMER3.

A general topology for a profile HMM



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

Profile Hidden Markov Models

Profile HMMs are probabilistic generative models to represent multiple sequence alignments.

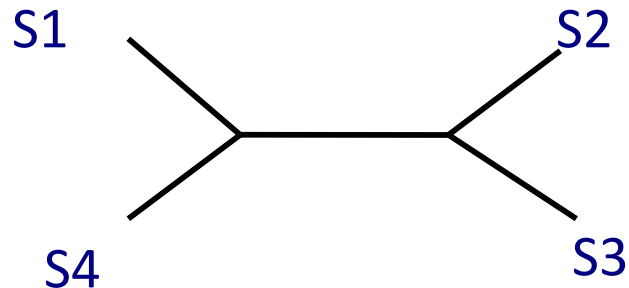
HMMER software suite can

Build a profile HMM given a multiple sequence alignment
A

Use the profile HMM to add a sequence s into A, and return the “probability” that the HMM generated s (the “score”)

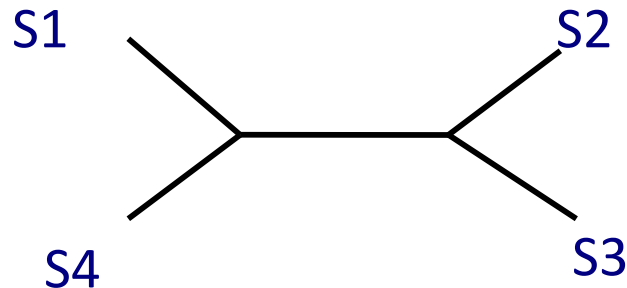
Align Sequence using HMMER

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



Align Sequence using HMMER

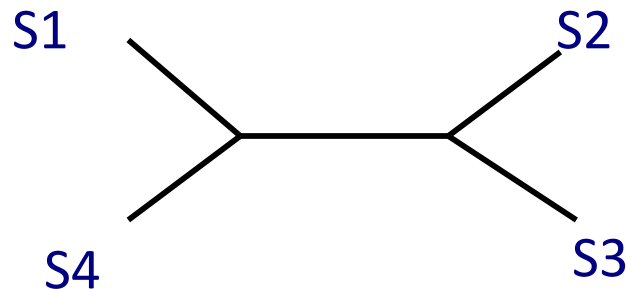
S1 = -AGGCTATCACCTGACCTCCA-AA
S2 = TAG-CTATCAC--GACCGC--GCA
S3 = TAG-CT-----GACCGC--GCT
S4 = TAC----TCAC--GACCGACAGCT
Q1 = -----T-A--AAAC-----



1. Build a profile HMM for the backbone alignment
2. Add Q1 to backbone alignment: Compute a maximum likelihood path through the profile HMM for Q1 and use it to compute the extended alignment.
3. Record the score for the alignment!

Place Sequence using pplacer

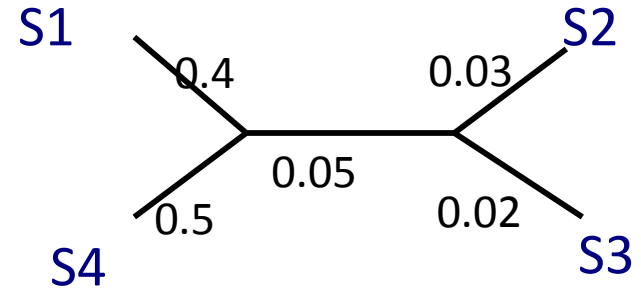
S1 = -AGGCTATCACCTGACCTCCA-AA
S2 = TAG-CTATCAC--GACCGC--GCA
S3 = TAG-CT-----GACCGC--GCT
S4 = TAC----TCAC--GACCGACAGCT
Q1 = -----T-A--AAAC-----



1. For every edge e in T , let T_e be the tree created by adding $Q1$ to that edge. Compute the maximum likelihood (ML) score of the tree T_e for the extended alignment. (Use the ML scores to assign probabilities $p(e)$ to all edges e !)
2. Return T_e that has the best ML score.

Place Sequence using pplacer

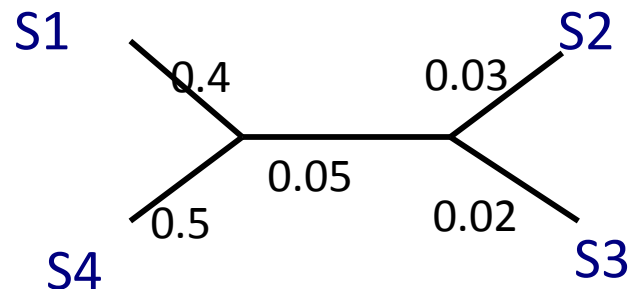
S1 = -AGGCTATCACCTGACCTCCA-AA
S2 = TAG-CTATCAC--GACCGC--GCA
S3 = TAG-CT-----GACCGC--GCT
S4 = TAC----TCAC--GACCGACAGCT
Q1 = -----T-A--AAAC-----



1. For every edge e in T , let T_e be the tree created by adding Q1 to that edge. Compute the maximum likelihood (ML) score of the tree T_e for the extended alignment. (Use the ML scores to assign probabilities $p(e)$ to all edges e !)

Place Sequence using pplacer

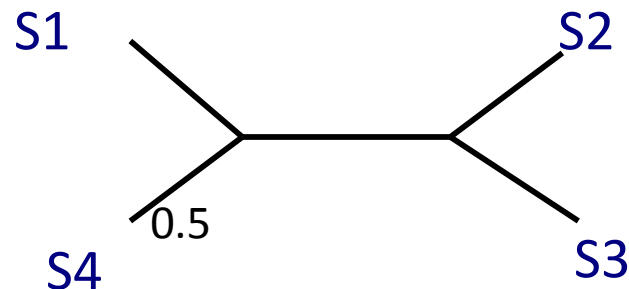
S1 = -AGGCTATCACCTGACCTCCA-AA
S2 = TAG-CTATCAC--GACCGC--GCA
S3 = TAG-CT-----GACCGC--GCT
S4 = TAC----TCAC--GACCGACAGCT
Q1 = -----T-A--AAAC-----



1. For every edge e in T , let T_e be the tree created by adding Q1 to that edge. Compute the maximum likelihood (ML) score of the tree T_e for the extended alignment. (Use the ML scores to assign probabilities $p(e)$ to all edges e !)
2. Find edge e^* producing the best ML score

Place Sequence using pplacer

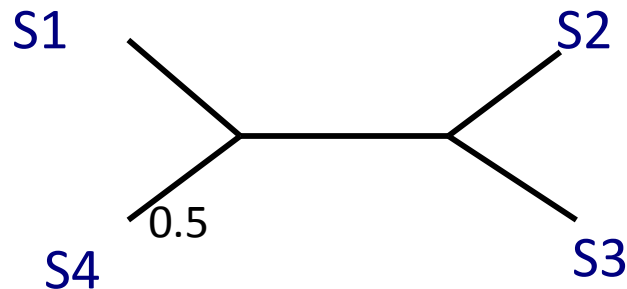
S1 = -AGGCTATCACCTGACCTCCA-AA
S2 = TAG-CTATCAC--GACCGC--GCA
S3 = TAG-CT-----GACCGC--GCT
S4 = TAC----TCAC--GACCGACAGCT
Q1 = -----T-A--AAAC-----



1. For every edge e in T , let T_e be the tree created by adding Q1 to that edge. Compute the maximum likelihood (ML) score of the tree T_e for the extended alignment. (Use the ML scores to assign probabilities $p(e)$ to all edges e !)
2. Find edge e^* producing the best ML score

Place Sequence using pplacer

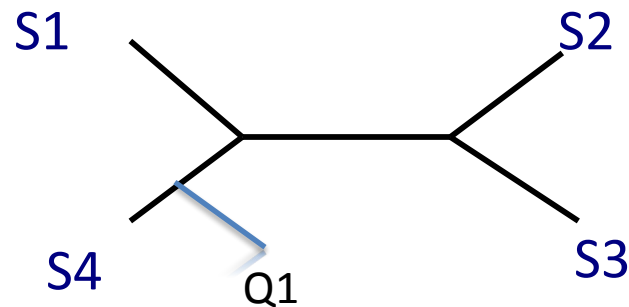
S1 = -AGGCTATCACCTGACCTCCA-AA
S2 = TAG-CTATCAC--GACCGC--GCA
S3 = TAG-CT-----GACCGC--GCT
S4 = TAC----TCAC--GACCGACAGCT
Q1 = -----T-A--AAAC-----



1. For every edge e in T , let T_e be the tree created by adding Q1 to that edge. Compute the maximum likelihood (ML) score of the tree T_e for the extended alignment. (Use the ML scores to assign probabilities $p(e)$ to all edges e !)
2. Add Q1 into edge e^*

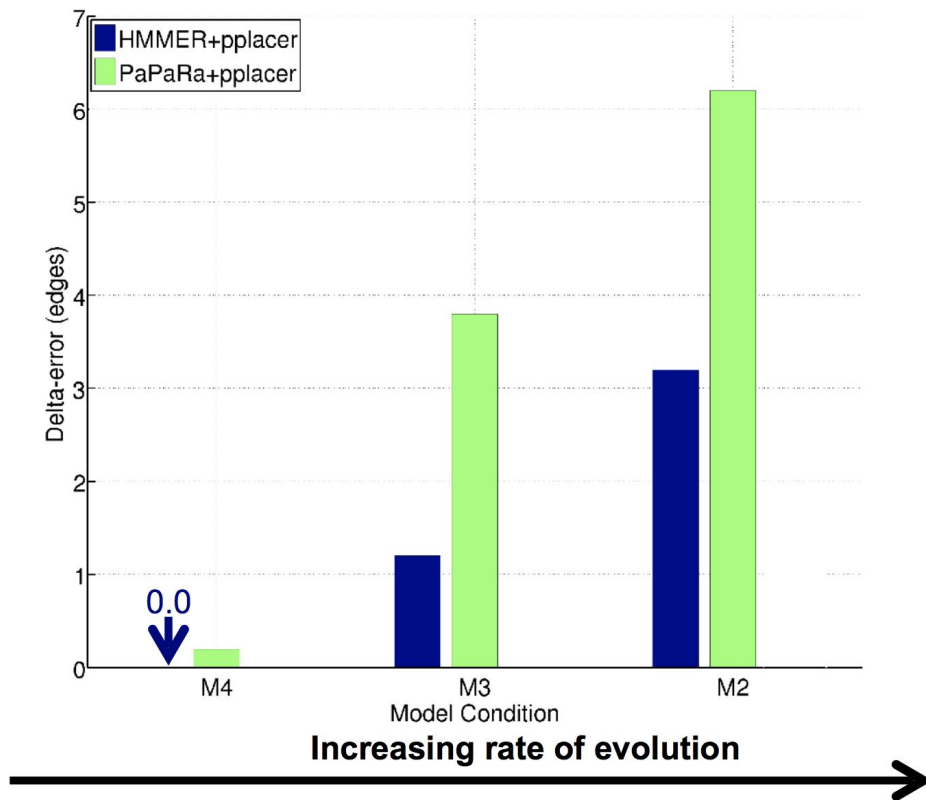
Place Sequence using pplacer

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

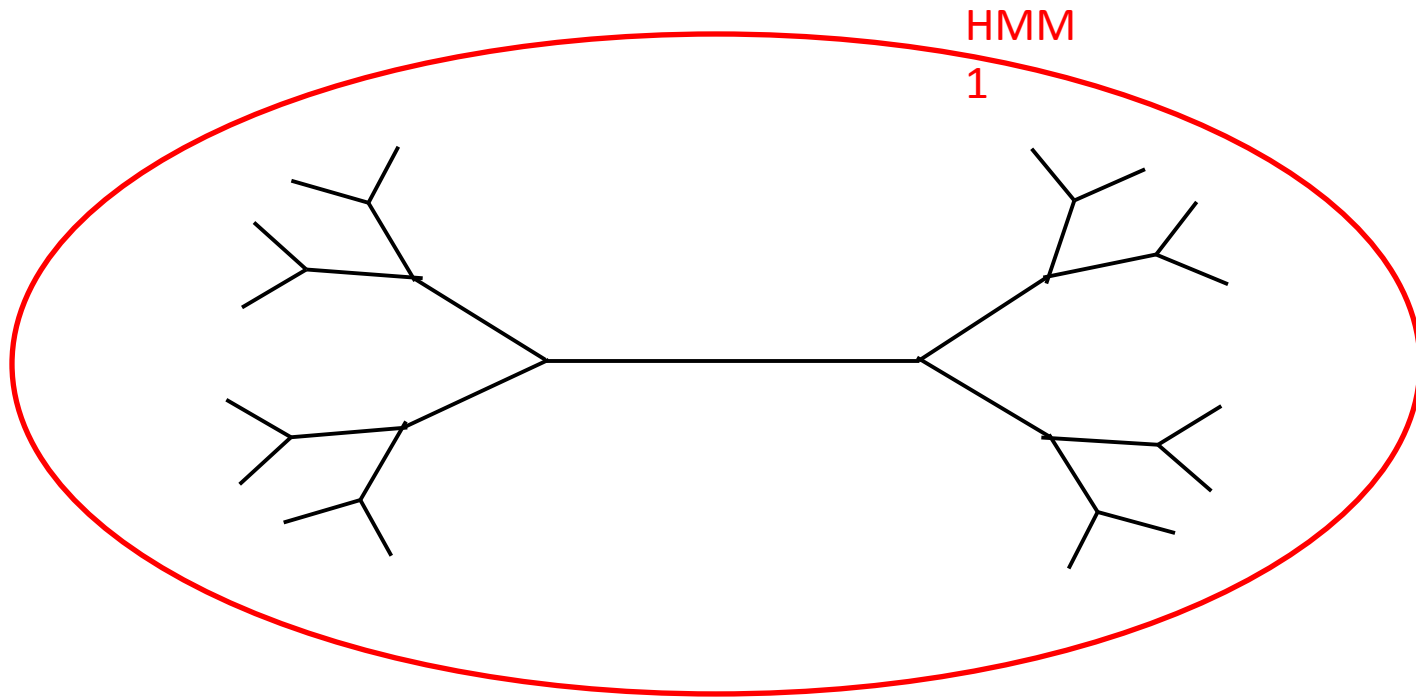


1. For every edge e in T , let T_e be the tree created by adding Q1 to that edge. Compute the maximum likelihood (ML) score of the tree T_e for the extended alignment. (Use the ML scores to assign probabilities $p(e)$ to all edges e !)
2. Add Q1 into edge e^*

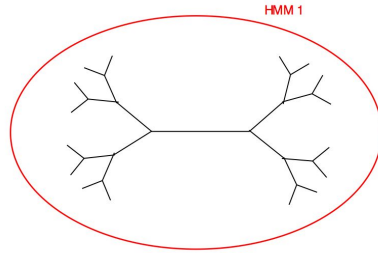
HMMER vs. PaPaRa Alignments



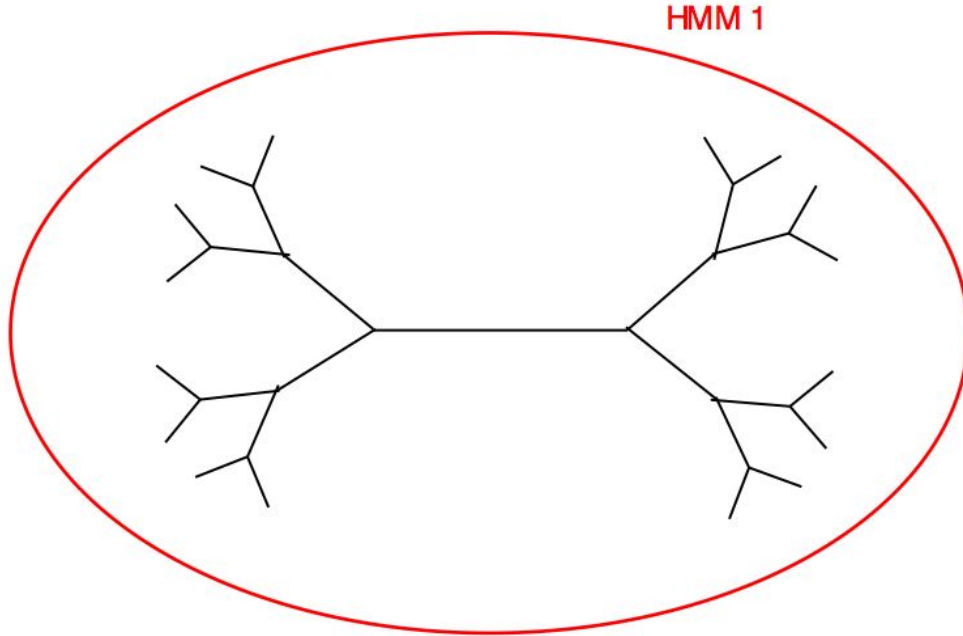
One Hidden Markov Model for the entire alignment?



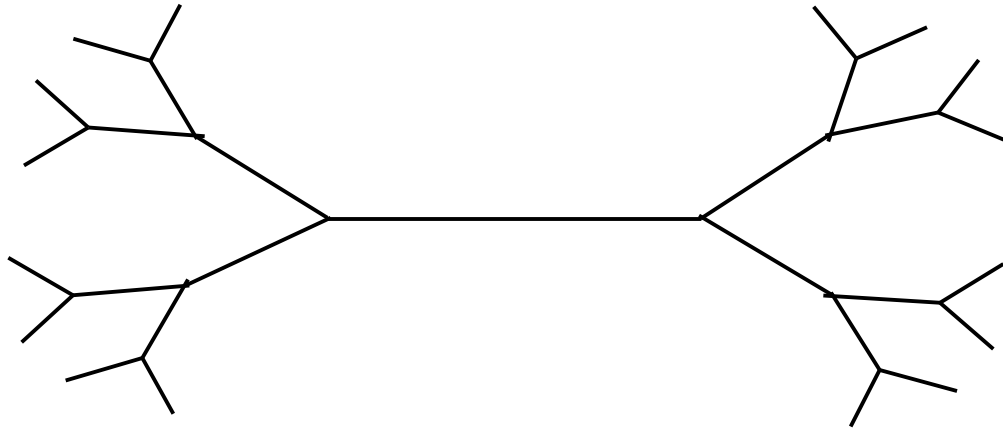
One HMM works beautifully for small-diameter trees



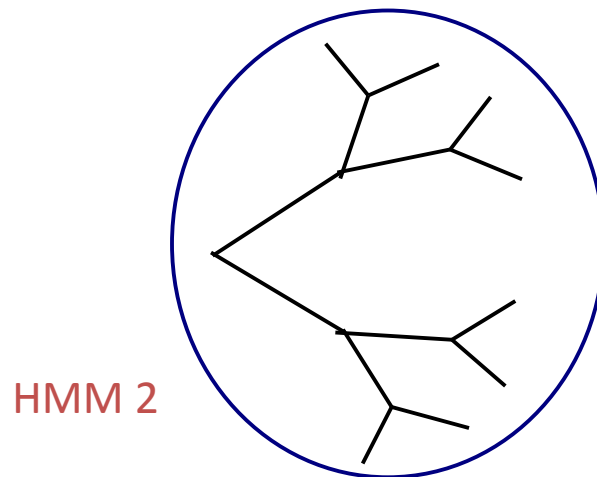
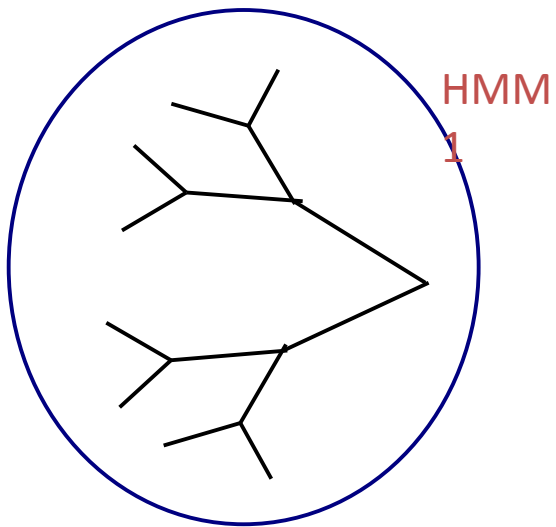
One HMM works poorly for large-diameter trees



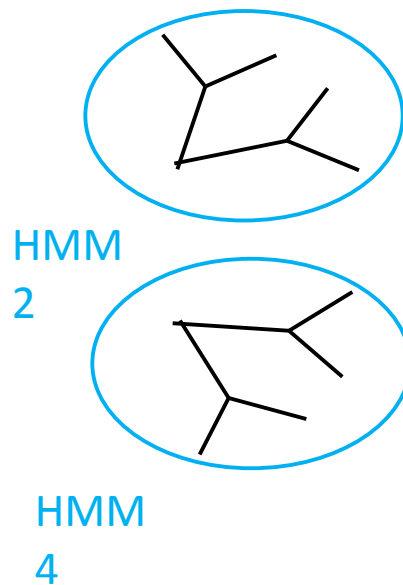
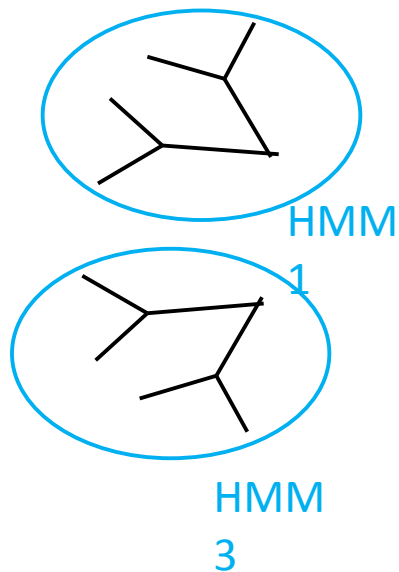
One Hidden Markov Model for the entire alignment?



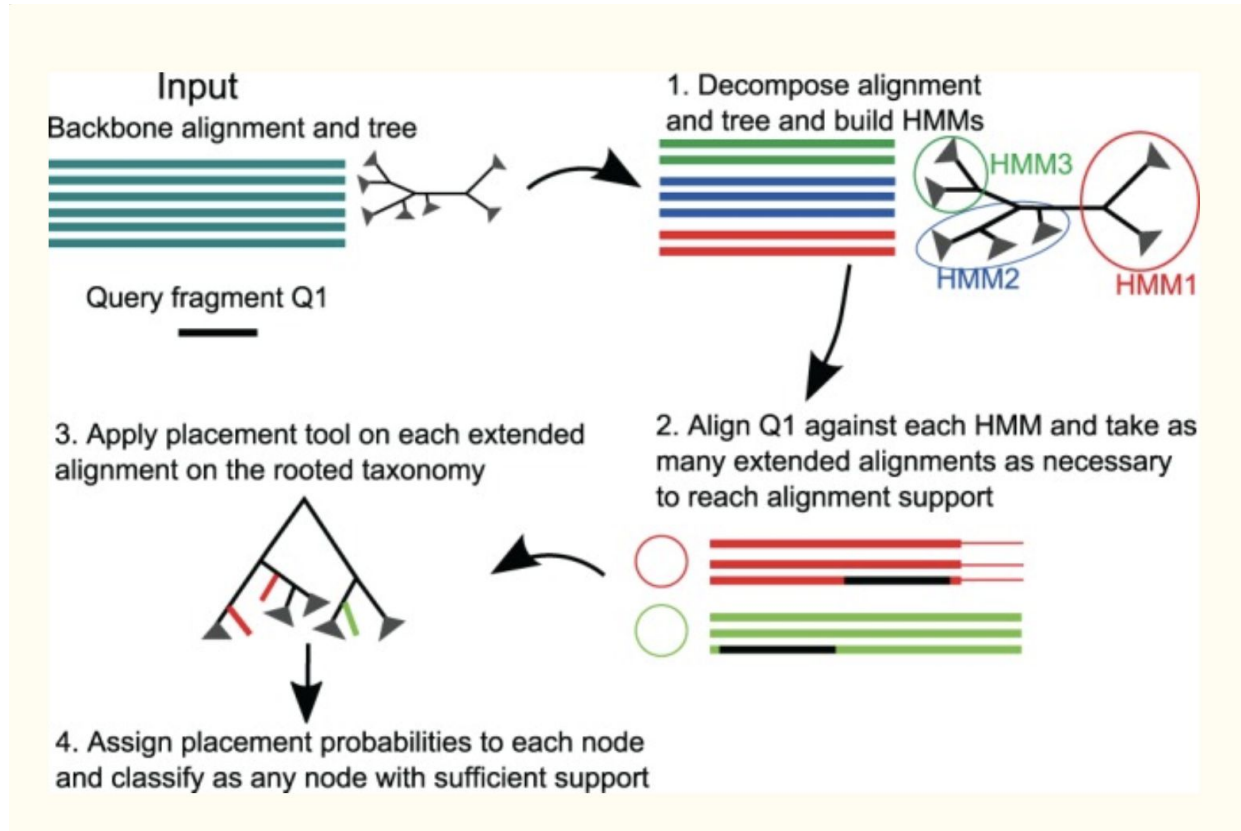
Or 2 HMMs?



Or 4 HMMs?



TIPP: Taxonomic identification and phylogenetic profiling



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

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

TIPP pipeline

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

Can consider more than one extended alignment

Can consider more than one placement in the tree for each extended alignment

Assign taxonomic label based on MRCA (most recent common ancestor) of all selected placements for all selected extended alignments

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.

Parameters used in placement (SEPP)

SEPP algorithmic parameters:

Alignment subset size (how many sequences for each profile HMM in the ensemble?)

Placement subset size (how much of the tree to search for optimal placement?)

Default settings are acceptable, but you can improve accuracy (but increase running time) by:

- increasing placement subset size
- and decreasing alignment subset size

Parameters used in TIPP

TIPP algorithmic parameters (other than SEPP parameters)

- Reference markers, alignments, and refined taxonomy

- Alignment threshold (default 95%)

- Placement threshold (default 95%)

Note:

The default alignment and placement thresholds were optimized for abundance profiling, not for Taxon ID.

Reducing the placement threshold will increase probability of taxonomic classification at the species level (but could also increase the false positive rate)

Run TIPP on command line

TIPP is under development!

We are modifying TIPP's design to improve taxonomic identification and abundance profiling on shotgun sequencing data

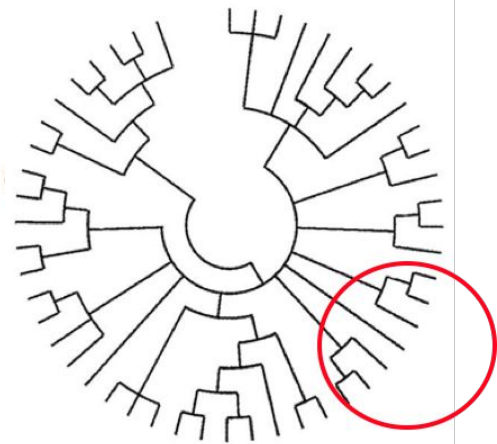
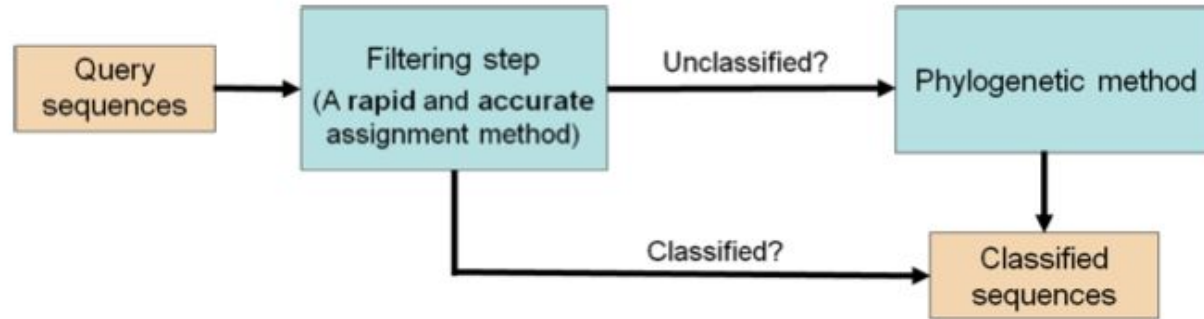
Students: Erin Molloy, Mike Nute, Nidhi Shah

Advisor: Tandy Warnow and Mihai Pop

These slides were adapted from Tandy Warnow's
talk at [STAMPS 2018](#)

Thanks!

A two-step approach



Suggestions for analysing whole metagenome sequencing data

- Get contigs from Assembly - reference guided (Metacompass) or de novo (Metaspades, Metahit)
- Taxonomic identification - contig based or read based
- Abundance profiling - marker based methods
- Binning approaches
- Looking for strain variations? Closely study metagenome assembly graphs (MetaCarvel, and MetagenomeScope)