TIPP and SEPP: Metagenomic Analysis using Phylogeny-Aware Profiles

Tandy Warnow and Erin Molloy
Department of Computer Science
The University of Illinois at Urbana-Champaign
Metagenomic taxonomic identification and phylogenetic profiling

Metagenomics, Venter et al., Exploring the Sargasso Sea: Scientists Discover One Million New Genes in Ocean Microbes
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?
This talk

• **SEPP** (PSB 2012): SATé-enabled Phylogenetic Placement, and **Ensembles of HMMs (eHMMs)**

• **TIPP** (Bioinformatics 2014): Applications of the eHMM technique to metagenomic abundance classification

Both available at [https://github.com/smirarab/sepp](https://github.com/smirarab/sepp)
Phylogenetic Placement

Input: **Backbone** alignment and 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

Phylogenetic placement can be used inside a pipeline, after determining the genes for each of the reads in the metagenomic sample.
Marker-based Taxon Identification

Fragmentary sequences from some gene

Full-length sequences for same gene, and an alignment and a tree
\textbf{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--------
S1  = -AGGCTATCACCTGACCTCCA-AA
S2  = TAG-CTATCAC--GACCGC--GCA
S3  = TAG-CT-------GACCGC--GCT
S4  = TAC-----TCAC--GACCGACAGCT
Q1  = ---------T-A--AAAC---------
Phylogenetic Placement in 2011

• Align each query sequence to backbone alignment
  – HMMER (Finn et al., NAR 2011)
  – 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 solve same problem (maximum likelihood placement under standard sequence evolution models)
HMMER vs. PaPaRa Alignments

Increasing rate of evolution
Input

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

\[
\begin{align*}
S1 &= -AGGCTATCACCTGACCTCCA-AA \\
S2 &= TAG-CTATCAC--GACCGC--GCA \\
S3 &= TAG-CT---------GACCGC--GCT \\
S4 &= TAC----TCAC--GACCGACAGCT \\
Q1 &= -------T-A--AAAC--------
\end{align*}
\]
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----------
What is HMMER+pplacer?

- **HMMER** (Finn et al., NAR 2011) is used to add the reads into the backbone alignment, thus producing an “extended alignment”. HMMER is based on **profile Hidden Markov Models** (profile HMMs).

- **pplacer** (Matsen et al. BMC Bioinformatics 2010) is used to add reads into the best location in the tree $T$. pplacer is based on **phylogenetic sequence evolution models** (e.g., GTR), and uses **maximum likelihood**.
HMMER: biosequence analysis using profile hidden Markov models

HMMER is used for searching sequence databases for sequence homologs, and for making sequence alignments. It implements methods using probabilistic models called profile hidden Markov models (profile HMMs).

HMMER is often used together with a profile database, such as Pfam or many of the databases that participate in Interpro. But HMMER can also work with query sequences, not just profiles, just like BLAST. For example, you can search a protein query sequence against a database with phmmer, or do an iterative search with jackhmmer.

HMMER is designed to detect remote homologs as sensitively as possible, relying on the strength of its underlying probability models. In the past, this strength came at significant computational expense, but as of the new HMMER3 project, HMMER is now essentially as fast as BLAST.

HMMER can be downloaded and installed as a command line tool on your own hardware, and now it is also more widely accessible to the scientific community via new search servers at the European Bioinformatics Institute.

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 Cryptogenomic for more information and discussion about HMMER3.
A general topology for a profile HMM

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

In other words, *profile HMMs can be used to compute extended alignments, and score them!*
Input

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

\[ S_1 = -\text{AGGCTATCACCTGACCTCCA-} \AA \]
\[ S_2 = \text{TAG-CTATCAC-} \text{-GACCGC-} \text{-GCA} \]
\[ S_3 = \text{TAG-CT-} \text{-GACCGC-} \text{-GCT} \]
\[ S_4 = \text{TAC---TCAC-} \text{-GACCGACAGCT} \]
\[ Q_1 = \text{-T-A--AAAC-} \text{-AAAC-} \]

1. Build a profile HMM for the backbone alignment
2. Compute a maximum likelihood path through the profile HMM for Q1 and use it to compute the extended alignment.
3. Note the maximum likelihood score for the alignment!
What is pplacer?

- **pplacer**: software developed by Erick Matsen and colleagues. See [http://matsen.fhcrc.org/pplacer/](http://matsen.fhcrc.org/pplacer/)

- **Input**: read s, alignment A (on S and s), tree on S

- **Output**:  
  - “Best” location to add s in T (under maximum likelihood).
  - For every edge e in T, the value \( p(e) \) for the probability for s being placed on e (these probabilities add up to 1)
Place Sequence using pplacer

1. For every edge 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.

\[
\begin{align*}
S1 &= -AGGCTATCACCCTGACCTCCA-AA \\
S2 &= TAG-CTATCAC--GACCGC--GCA \\
S3 &= TAG-CT--------GACCGC--GCT \\
S4 &= TAC------TCAC--GACCGACAGCT \\
Q1 &= --------T-A--AAAC----------
\end{align*}
\]
Place Sequence using pplacer

S1 = -AGGCTATCAGCTGACCTCCA- AA
S2 = TAG-CTATCAG--GACCGC--GCA
S3 = TAG-CT--------GACCGC--GCT
S4 = TAC----TCAC--GACCGAGACT
Q1 = --------T-A--AAAC--------

1. For every edge 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

1. For every edge 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

\[
\begin{align*}
S1 &= -AGGCTATCACCTGACCTCCA-AA \\
S2 &= TAG-CTATCAGACCGC--GCA \\
S3 &= TAG-CT-------GACCGC--GCT \\
S4 &= TAC-----TCAC--GACCGACAGCT \\
Q1 &= --------T-A--AAAC--------
\end{align*}
\]

1. For every edge 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.
HMMER vs. PaPaRa Alignments

Increasing rate of evolution
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?
SEPP Design

To insert query sequence Q1 into backbone tree T

- Represent the backbone MSA with a collection of profile HMMs, based on maximum ”alignment subset size”
- Score Q1 against every profile HMM in the collection
- The best scoring HMM is used to compute the extended alignment
- Use pplacer to add Q1 into tree T (restricted to ”subtree” based on maximum ”placement subset size”)


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:
  - Small alignment subset sizes best
  - Large placement subset size best
- **10% rule** (both subset sizes 10% of backbone) had best overall performance
SEPP (10%-rule) on simulated data

Increasing rate of evolution
SEPP and 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.
Metagenomic Taxon Identification

Objective: classify short reads in a metagenomic sample
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 (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
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.
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
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
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
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 optimal placement in the tree for each extended alignment
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
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 of all selected placements for all selected extended alignments
TIPP Design (Step 4)

- Input: marker gene reference alignment (computed using PASTA, RECOMB 2014), species taxonomy, alignment support threshold (default 95%) and placement support threshold (default 95%)
TIPP Design (Step 4)

• Input: marker gene reference alignment (computed using PASTA, RECOMB 2014), species taxonomy, alignment support threshold (default 95%) and placement support threshold (default 95%)
• For each marker gene, and its associated bin of reads:
TIPP Design (Step 4)

- Input: marker gene reference alignment (computed using PASTA, RECOMB 2014), species taxonomy, alignment support threshold (default 95%) and placement support threshold (default 95%)
- For each marker gene, and its associated bin of reads:
  - Builds eHMM to represent the MSA
TIPP Design (Step 4)

- Input: marker gene reference alignment (computed using PASTA, RECOMB 2014), species taxonomy, alignment support threshold (default 95%) and placement support threshold (default 95%)

- For each marker gene, and its associated bin of reads:
  - Builds eHMM to represent the MSA
  - For each read:
TIPP Design (Step 4)

• Input: marker gene reference alignment (computed using PASTA, RECOMB 2014), species taxonomy, alignment support threshold (default 95%) and placement support threshold (default 95%)

• For each marker gene, and its associated bin of reads:
  – Builds eHMM to represent the MSA
  – For each read:
    • Use the eHMM to produce a set of extended MSAs that include the read, sufficient to reach the specified alignment support threshold.
TIPP Design (Step 4)

• Input: marker gene reference alignment (computed using PASTA, RECOMB 2014), species taxonomy, alignment support threshold (default 95%) and placement support threshold (default 95%)

• For each marker gene, and its associated bin of reads:
  – Builds eHMM to represent the MSA
  – For each read:
    • Use the eHMM to produce a set of extended MSAs that include the read, sufficient to reach the specified alignment support threshold.
    • For each extended MSA, use pplacer to place the read into the taxonomy optimizing maximum likelihood and identify all the clades in the tree with sufficiently high likelihood to meet the specified placement support threshold. (Note – this will be a single clade if the support threshold is at strictly greater than 50%.)
TIPP Design (Step 4)

• Input: marker gene reference alignment (computed using PASTA, RECOMB 2014), species taxonomy, alignment support threshold (default 95%) and placement support threshold (default 95%)

• For each marker gene, and its associated bin of reads:
  – Builds eHMM to represent the MSA
  – For each read:
    • Use the eHMM to produce a set of extended MSAs that include the read, sufficient to reach the specified alignment support threshold.
    • For each extended MSA, use **pplacer** to place the read into the taxonomy optimizing maximum likelihood and identify all the clades in the tree with sufficiently high likelihood to meet the specified placement support threshold. (Note – this will be a single clade if the support threshold is at strictly greater than 50%.)
    • Taxonomically characterize each read at the **MRCA** of these clades.
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 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.
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!
HIPPI

• Hierarchical Profile HMMs for Protein family Identification
• Nguyen, Nute, Mirarab, and Warnow, RECOMB-CG and BMC-Genomics 2016
• Uses an ensemble of HMMs to classify protein sequences
• Tested on HMMER
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
TIPPI: Replacing BLAST by HIPPI within TIPP

To appear, Nguyen et al., BMC Genomics
Our Publications using eHMMs


All codes are available in open source form at https://github.com/smirarab/sepp
Summary

• Using an ensemble of HMMs tends to improve accuracy, for a cost of running time. Applications so far to taxonomic placement (SEPP), multiple sequence alignment (UPP), protein family classification (HIPPI). Improvements are mostly noticeable for large diverse datasets.

• Phylogenetically-based construction of the ensemble helps accuracy (note: the decompositions we produce are not clade-based), but the design and use of these ensembles is still in its infancy. (Many relatively similar approaches have been used by others, including FlowerPower by Sjolander)

• The basic idea can be used with any kind of probabilistic model, doesn’t have to be restricted to profile HMMs.

• Basic question: why does it help?
Using 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 by increasing placement subset size (and maybe decreasing alignment subset size), for a running time cost.
Using TIPP

TIPP algorithmic parameters (other than SEPP parameters)
  – Reference markers
  – Alignment threshold (default 95%)
  – Placement threshold (default 95%)

The alignment threshold and placement threshold were optimized for abundance profiling, not for Taxon ID.
Acknowledgments

PhD students: Nam Nguyen (now postdoc at UCSD) and Siavash Mirarab (now faculty at UCSD), and 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**