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

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)

  - ACCG
  - CGAG
  - CGG
  - GGCT
  - TAGA
  - GGGG
  - TCGAG
  - GGCG
  - GGG
  
  - ACCT

- Known Full length Sequences, and an alignment and a tree:
  - (500-10,000 bp long)

  - TAGG...GCA
  - CCA
  - TAG...CTT
  - AGC...ACA
  - ACT...TAGA...A
  - (species1)
  - (species2)
  - (species3)
  - (species4)
  - (species5)
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

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

Increasing rate of evolution

Delta-error (edges)
One Hidden Markov Model for the entire alignment?
Or 2 HMMs?
Or 4 HMMs?

HMM 1

HMM 2

HMM 3

HMM 4
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

Increasing rate of evolution
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

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%

Method
- HIPPI
- HMMER
- BLAST
- HHsearch: 1 Iteration
- HHsearch: 2 Iterations
Results in the Twilight Zone

Results on 25% Fragments

- > 30% Identity
- 20-30% Identity
- < 20% Identity

Precision vs Recall for:
- HIPPI
- HMMER
- BLAST

HHSearch: 1 Iteration
HHSearch: 2 Iterations
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)
Acknowledgments

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

Computational analyses performed on TACC and BlueWaters
TIPP version 2- considering uncertainty

TIPP version 1 too aggressive (over-classification)
TIPP version 2 dramatically reduces false positive rate with small reduction in true positive rate, by considering uncertainty, using statistical techniques:

• Find 2-4 alignments of reference sequences and construct tree (SATé, ClustalW, MAFFT, others)

• For each reference alignment/tree pair, compute many extended alignments (using statistical support computed using HMMER to cover 95% of the probability).

• For each extended alignment, use pplacer statistical support values to place fragment into taxonomy, so that the clade below the placement contains 99% of the probability.

Classify each fragment at the LCA of all placements obtained for the fragment
TIPP+Metaphyler

• Use Metaphyler to perform initial placement of read into taxonomy
• Use TIPP to modify the placement, moving the read further into the clade identified by Metaphyler
Results on rpsB gene (60 bp)
Phylogenetic “Boosters”

- **SATé**: co-estimation of alignments and trees
- **SEPP/TIPP**: phylogenetic analysis of fragmentary data

Algorithmic strategies: divide-and-conquer and iteration to improve the accuracy and scalability of a base method
Part I: UPP

UPP = “Ultra-large multiple sequence alignment using Phylogeny-aware Profiles”


Purpose: highly accurate large-scale multiple sequence alignments, even in the presence of fragmentary sequences.
Summary

• SATé gives better alignments and trees than standard alignment estimation methods

• SEPP can enable alignment of short (fragmentary) sequences into alignments of full-length sequences, and phylogenetic placement into gene trees or taxonomies

• TIPP enables taxon identification of short reads -- not limited to 31 marker genes, and no training is needed.
Overall message

• When data are difficult to analyze, develop better methods - don’t throw out the data.
Acknowledgments

• Guggenheim Foundation Fellowship, Microsoft Research New England, National Science Foundation: Assembling the Tree of Life (ATOL), ITR, and IGERT grants, and David Bruton Jr. Professorship

• NSERC support to Siavash Mirarab

• Collaborators:
  – SATé: Kevin Liu, Serita Nelesen, Sindhu Raghavan, and Randy Linder
  – SEPP: Siavash Mirarab and Nam Nguyen
  – TIPP: Siavash Mirarab, Nam Nguyen, Bo Liu, and Mihai Pop
Megan vs MetaPhyler on 60bp rpsB gene
OBSERVATIONS

• MEGAN is very conservative
• MetaPhyler makes more correct predictions than MEGAN
• Other methods (Liu et al, BMC Bioinformatics 2011) not as sensitive (on these 31 marker genes) as MetaPhyler

Thus, the best taxon identification methods have **high precision** (make accurate predictions), but **low sensitivity** (i.e., they **fail to classify** a large portion of reads) even at higher taxonomy levels.
Profile Hidden Markov Models

- Probabilistic model to represent a family of sequences, represented by a multiple sequence alignment
- Fundamental part of HMMER and other protein databases
- Used for: homology detection, protein family assignment, multiple sequence alignment, phylogenetic placement, protein structure prediction, alignment segmentation, etc.
Given a set $Q$ of query sequences for some gene, a taxonomy $T$, and a set of full-length sequences for the gene,

- Compute reference alignment and tree on the full-length sequences, using SATé
- Use SEPP to place each query sequence into the taxonomy (alignment subsets computed on the reference alignment/tree, then inserted into taxonomy $T$)