Using Ensembles of Hidden Markov Models for Grand Challenges in Bioinformatics

Tandy Warnow
Founder Professor of Engineering
The University of Illinois at Urbana-Champaign
http://tandy.cs.illinois.edu
Phylogenomics

From the Tree of the Life Website,
University of Arizona
1kp: Thousand Transcriptome Project

- Plant Tree of Life based on transcriptomes of ~1200 species
- More than 13,000 gene families (most not single copy)

Challenge:
Alignments and trees on > 100,000 sequences
1000-taxon models, ordered by difficulty (Liu et al., Science 324(5934):1561-1564, 2009)
Re-aligning on a tree

Decompose dataset

Align subsets

Estimate ML tree on merged alignment

Merge sub-alignments
SATé and PASTA Algorithms

Obtain initial alignment and estimated ML tree

Estimate ML tree on new alignment

Use tree to compute new alignment

Repeat until termination condition, and return the alignment/tree pair with the best ML score
1000 taxon models, ordered by difficulty, Liu et al., Science 324(5934):1561-1564, 2009

24 hour SATé-I analysis, on desktop machines

(Similar improvements for biological datasets)
SATé-2 better than SATé-1

SATé-1 (Liu et al., Science 2009): can analyze up to 8K sequences
SATé-2 (Liu et al., Systematic Biology 2012): can analyze up to ~50K sequences
PASTA: Mirarab, Nguyen, and Warnow, J Comp. Biol. 2015
- Simulated RNAsim datasets from 10K to 200K taxa
- Limited to 24 hours using 12 CPUs
- Not all methods could run (missing bars could not finish)
Comparing default PASTA to PASTA+BAli-Phy on simulated datasets with 1000 sequences

Decomposition to 100-sequence subsets, one iteration of PASTA+BAli-Phy
1kp: Thousand Transcriptome Project

- Plant Tree of Life based on transcriptomes of ~1200 species
- More than 13,000 gene families (most not single copy)

**Challenge:**
Alignments and trees on > 100,000 sequences
1KP dataset: more than 100,000 p450 amino-acid sequences, many fragmentary
1KP dataset: more than 100,000 p450 amino-acid sequences, many fragmentary.

All standard multiple sequence alignment methods we tested performed poorly on datasets with fragments.
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.
Profile HMMs

- Generative model for representing a MSA
- Consists of:
  - Set of states (Match, insertion, and deletion)
  - Transition probabilities
  - Emission probabilities
Profile Hidden Markov Model for DNA sequence alignment
HMMs for MSA

- Given seed alignment (e.g., in PFAM) and a collection of sequences for the protein family:
  - Represent seed alignment using HMM
  - Align each additional sequence to the HMM
  - Use transitivity to obtain MSA
HMMs for MSA

- Given seed alignment (e.g., in PFAM) and a collection of sequences for the protein family:
  - Represent seed alignment using HMM
  - Align each additional sequence to the HMM
  - Use transitivity to obtain MSA
- Can we do something like this without a seed alignment?
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.
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.

Uses an ensemble of HMMs
Simple idea (not UPP)

• Select random subset of sequences, and build “backbone alignment”

• Construct a Hidden Markov Model (HMM) on the backbone alignment

• Add all remaining sequences to the backbone alignment using the HMM
PASTA: even better than SATé-2

Starting tree is based on UPP(simple):
one profile HMM for Backbone alignment

PASTA: Mirarab, Nguyen, and Warnow, J Comp. Biol. 2015
- Simulated RNAsim datasets from 10K to 200K taxa
- Limited to 24 hours using 12 CPUs
- Not all methods could run (missing bars could not finish)
Select random subset of sequences, and build “backbone alignment”

• Construct a Hidden Markov Model (HMM) on the backbone alignment

• Add all remaining sequences to the backbone alignment using the HMM

This approach works well if the dataset is small and has low evolutionary rates, but is not very accurate otherwise.
One Hidden Markov Model for the entire alignment?
One Hidden Markov Model for the backbone alignment?
Or 2 HMMs?
Or 4 HMMs?
Or all 7 HMMs?
UPP Algorithmic Approach

1. Select random subset of full-length sequences, and build “backbone alignment”

2. Construct an “Ensemble of Hidden Markov Models” on the backbone alignment

3. Add all remaining sequences to the backbone alignment using the Ensemble of HMMs
Evaluation

• Simulated datasets (some have fragmentary sequences):
  – 10K to 1,000,000 sequences in RNASim – complex RNA sequence evolution simulation
  – 1000-sequence nucleotide datasets from SATé papers
  – 5000-sequence AA datasets (from FastTree paper)
  – 10,000-sequence Indelible nucleotide simulation

• Biological datasets:
  – Proteins: largest BaliBASE and HomFam
  – RNA: 3 CRW datasets up to 28,000 sequences
Using 12 processors:

- UPP(Fast, NoDecomp) took 2.2 days,
- UPP(Fast) took 11.9 days, and
- PASTA took 10.3 days
UPP is very robust to fragmentary sequences

Under high rates of evolution, PASTA is badly impacted by fragmentary sequences (the same is true for other methods).

UPP continues to have good accuracy even on datasets with many fragments under all rates of evolution.

Performance on fragmentary datasets of the 1000M2 model condition
UPP Running Time

Wall-clock align time (hr) vs. Number of sequences.

- UPP(Fast)

Wall-clock time used (in hours) given 12 processors.
Other Applications of the Ensemble of HMMs

SEPP (phylogenetic placement, Mirarab, Nguyen, and Warnow PSB 2014)

**TIPP** (metagenomic taxon identification, Nguyen, Mirarab, Liu, Pop, and Warnow, Bioinformatics 2014)

**HIPPI** (protein classification and remote homology detection), RECOMB-CG 2016 and BMC Genomics 2016 (Nguyen, Nute, Mirarab, and Warnow)
The NIH Human Microbiome Project

25,000 human genes, 1,000,000 bacterial genes
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 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.
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
HIPPI

- Hierarchical Profile HMMs for Protein family Identification
- Nguyen, Nute, Mirarab, and Warnow, RECOMB-CG 2016 and BMC-Genomics 2016
- Uses an ensemble of HMMs to classify protein sequences
- Tested on HMMER
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)
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 Sci-Phy and 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 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 Sci-Phy and FlowerPower by Sjolander.)

- The basic idea can be used with any kind of probabilistic model, doesn’t have to be restricted to profile HMMs.
The Tree of Life: *Multiple Challenges*

Scientific challenges:

- Ultra-large multiple-sequence alignment
- Alignment-free phylogeny estimation
- Supertree estimation
- Estimating species trees from many gene trees
- Genome rearrangement phylogeny
- Reticulate evolution
- Visualization of large trees and alignments
- Data mining techniques to explore multiple optima
- Theoretical guarantees under Markov models of evolution

Techniques:

machine learning, applied probability theory, graph theory, combinatorial optimization, supercomputing, and heuristics
Acknowledgments

PASTA and UPP: Nam Nguyen (now postdoc at UIUC) and Siavash Mirarab (now faculty at UCSD), undergrad: Keerthana Kumar (at UT-Austin)
PASTA+BAli-Phy: Mike Nute (PhD student at UIUC)

**Current NSF grants:** ABI-1458652 (multiple sequence alignment)
Grainger Foundation (at UIUC), and UIUC
TACC, UTCS, Blue Waters, and UIUC campus cluster

PASTA, UPP, SEPP, and TIPP are available on github at https://github.com/smirarab/; see also PASTA+BAli-Phy at http://github.com/MGNute/pasta
Table 2. Average alignment SP-error, tree error, and TC score across most full-length datasets

<table>
<thead>
<tr>
<th>Method</th>
<th>ROSE NT</th>
<th>RNASim 10K</th>
<th>Indelible 10K</th>
<th>ROSE AA</th>
<th>CRW 10 AA</th>
<th>10 AA (17)</th>
<th>HomFam (2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>UPP</td>
<td>7.8 (1)</td>
<td>9.5 (1)</td>
<td>1.7 (2)</td>
<td>2.9 (1)</td>
<td>12.5 (1)</td>
<td>24.2 (1)</td>
<td>23.3 (1)</td>
</tr>
<tr>
<td>PASTA</td>
<td>7.8 (1)</td>
<td>15.0 (2)</td>
<td>0.4 (1)</td>
<td>3.1 (1)</td>
<td>12.8 (1)</td>
<td>24.0 (1)</td>
<td>22.5 (1)</td>
</tr>
<tr>
<td>MAFFT</td>
<td>20.6 (2)</td>
<td>25.5 (3)</td>
<td>41.4 (3)</td>
<td>4.9 (2)</td>
<td>28.3 (2)</td>
<td>23.5 (1)</td>
<td>25.3 (2)</td>
</tr>
<tr>
<td>Muscle</td>
<td>20.6 (2)</td>
<td>64.7 (5)</td>
<td>62.4 (4)</td>
<td>5.5 (3)</td>
<td>30.7 (3)</td>
<td>30.2 (2)</td>
<td>48.1 (4)</td>
</tr>
<tr>
<td>Clustal</td>
<td>49.2 (3)</td>
<td>35.3 (4)</td>
<td>X</td>
<td>6.5 (4)</td>
<td>43.3 (4)</td>
<td>24.3 (1)</td>
<td>27.7 (3)</td>
</tr>
</tbody>
</table>

Average /Delta1 FN error

<table>
<thead>
<tr>
<th>Method</th>
<th>ROSE NT</th>
<th>RNASim 10K</th>
<th>Indelible 10K</th>
<th>ROSE AA</th>
<th>CRW 10 AA</th>
<th>10 AA (17)</th>
<th>HomFam (2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>UPP</td>
<td>1.3 (1)</td>
<td>0.8 (1)</td>
<td>0.3 (1)</td>
<td>1.8 (1)</td>
<td>7.8 (2)</td>
<td>3.4 (2)</td>
<td>NA</td>
</tr>
<tr>
<td>PASTA</td>
<td>1.3 (1)</td>
<td>0.4 (1)</td>
<td>&lt;0.1 (1)</td>
<td>1.3 (1)</td>
<td>5.1 (1)</td>
<td>3.3 (1)</td>
<td>NA</td>
</tr>
<tr>
<td>MAFFT</td>
<td>5.8 (2)</td>
<td>3.5 (2)</td>
<td>24.8 (3)</td>
<td>4.5 (3)</td>
<td>10.1 (3)</td>
<td>2.3 (1)</td>
<td>NA</td>
</tr>
<tr>
<td>Muscle</td>
<td>8.4 (3)</td>
<td>7.3 (3)</td>
<td>32.5 (4)</td>
<td>3.1 (2)</td>
<td>5.5 (1)</td>
<td>12.6 (3)</td>
<td>NA</td>
</tr>
<tr>
<td>Clustal</td>
<td>24.3 (4)</td>
<td>10.4 (4)</td>
<td>X</td>
<td>4.2 (3)</td>
<td>34.1 (4)</td>
<td>3.5 (2)</td>
<td>NA</td>
</tr>
</tbody>
</table>

Average TC score

<table>
<thead>
<tr>
<th>Method</th>
<th>ROSE NT</th>
<th>RNASim 10K</th>
<th>Indelible 10K</th>
<th>ROSE AA</th>
<th>CRW 10 AA</th>
<th>10 AA (17)</th>
<th>HomFam (2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>UPP</td>
<td>37.8 (1)</td>
<td>0.5 (2)</td>
<td>11.0 (3)</td>
<td>2.6 (2)</td>
<td>1.4 (1)</td>
<td>11.4 (1)</td>
<td>47.3 (1)</td>
</tr>
<tr>
<td>PASTA</td>
<td>37.8 (1)</td>
<td>2.3 (1)</td>
<td>48.0 (1)</td>
<td>5.4 (1)</td>
<td>2.3 (1)</td>
<td>12.1 (1)</td>
<td>46.1 (2)</td>
</tr>
<tr>
<td>MAFFT</td>
<td>31.4 (2)</td>
<td>0.4 (2)</td>
<td>7.8 (4)</td>
<td>0.6 (3)</td>
<td>0.7 (2)</td>
<td>12.1 (1)</td>
<td>45.5 (2)</td>
</tr>
<tr>
<td>Muscle</td>
<td>9.8 (3)</td>
<td>&lt;0.0 (2)</td>
<td>18.3 (2)</td>
<td>2.7 (2)</td>
<td>0.7 (2)</td>
<td>10.5 (2)</td>
<td>27.7 (4)</td>
</tr>
<tr>
<td>Clustal</td>
<td>5.7 (4)</td>
<td>0.2 (2)</td>
<td>X</td>
<td>3.1 (2)</td>
<td>0.1 (2)</td>
<td>11.8 (1)</td>
<td>38.6 (3)</td>
</tr>
</tbody>
</table>

We report the average alignment SP-error (the average of SFPN and SFPF errors) (top), average ΔFN error (middle), and average TC score (bottom), for the collection of full-length datasets. All scores represent percentages and so are out of 100. Results marked with an X indicate that the method failed to terminate within the time limit (24 hours on a 12-core machine). Muscle failed to align two of the HomFam datasets; we report separate average results on the 17 HomFam datasets for all methods and the two HomFam datasets for all but Muscle. We did not test tree error on the HomFam datasets (therefore, the ΔFN error is indicated by "NA"). The tier ranking for each method is shown parenthetically.
1. Pre-processing seed alignments

i) Compute an ML tree on each seed alignment
ii) Build ensemble of HMMs for each seed alignment, using its ML tree
iii) Collect HMMs into database

2. Classification of query sequences

i) Score query sequence against all HMMs in database, keeping only scores above inclusion threshold
ii) Rank families by best scoring HMM within family
iii) Assign query sequence to top ranking family