Ultra-large Multiple Sequence Alignment

Tandy Warnow
Founder Professor of Engineering
The University of Illinois at Urbana-Champaign
http://tandy.cs.illinois.edu
Phylogeny
(evolutionary tree)

From the Tree of the Life Website,
University of Arizona
Phylogenies and Applications

Basic Biology:
How did life evolve?

Applications of phylogenies to:
protein structure and function
population genetics
human migrations
metagenomics
Hard Computational Problems

NP-hard problems

Large datasets
100,000+ sequences
thousands of genes

“Big data” complexity:
model misspecification
fragmentary sequences
errors in input data
streaming data
Phylogeny Problem

AGGGGCAT  TAGCCCCA  TAGACTT  TGCACAA  TGCGGCTT

V
W
X
Y

U
V
W
X
Y
Much is known about this problem from a mathematical and empirical viewpoint.
However…
Indels (insertions and deletions)

...ACGGTGCGAGTTACCCA...

Deletion

Mutation

...ACCGAGTCAACCA...
The true multiple alignment

- Reflects historical substitution, insertion, and deletion events
- Defined using transitive closure of pairwise alignments computed on edges of the true tree
Input: unaligned sequences

S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCAGACCGC
S3 = TAGCTGACCAGCGC
S4 = TCACGACCAGACA
Phase 1: Alignment

S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCAGACCGC
S3 = TAGCTGACCGC
S4 = TCACGACCGACA

S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAC---GACCGC--
S3 = TAG-CT-------GACCGC--
S4 = --------TCAC--GACCGACA
Phase 2: Construct tree

S1 = AGGCTATCACCTGACCTCCA  S1 = -AGGCTATCACCTGACCTCCA
S2 = TAGCTATCAGACCGC       S2 = TAG-CTATCAC--GACCGC--
S3 = TAGCTGACCGC           S3 = TAG-CT--------GACCGC--
S4 = TCACGACCGACA         S4 = --------TCAC--GACCGACA
Phylogenomic pipeline

• Select taxon set and markers
• Gather and screen sequence data, possibly identify orthologs
• Compute multiple sequence alignments for each locus, and construct gene trees
• Compute species tree or network:
  – Combine the estimated gene trees, OR
  – Estimate a tree from a concatenation of the multiple sequence alignments
• Get statistical support on each branch (e.g., bootstrapping)
• Estimate dates on the nodes of the phylogeny
• Use species tree with branch support and dates to understand biology
Phylogenomic pipeline

• Select taxon set and markers
• Gather and screen sequence data, possibly identify orthologs
• **Compute multiple sequence alignments for each locus, and construct gene trees**
• Compute species tree or network:
  – Combine the estimated gene trees, OR
  – Estimate a tree from a concatenation of the multiple sequence alignments
• Get statistical support on each branch (e.g., bootstrapping)
• Estimate dates on the nodes of the phylogeny
• Use species tree with branch support and dates to understand biology
Multiple Sequence Alignment (MSA): a scientific grand challenge

\[ S_1 = \text{AGGCTATCACCTGACCTCCA} \]
\[ S_2 = \text{TAGCTATCACGACCGC} \]
\[ S_3 = \text{TAGCTGACCGC} \]
\[ \quad \cdots \]
\[ S_n = \text{TCACGACCGGACA} \]

Novel techniques needed for scalability and accuracy

- NP-hard problems and large datasets
- Current methods do not provide good accuracy
- Few methods can analyze even moderately large datasets

Many important applications besides phylogenetic estimation

\(^1\) Frontiers in Massive Data Analysis, National Academies Press, 2013
This talk

• “Big data” multiple sequence alignment

• SATé (Science 2009, Systematic Biology 2012) and PASTA (RECOMB and J Comp Biol 2015), methods for co-estimation of alignments and trees

• UPP (Genome Biology 2015): ultra-large multiple sequence alignment, using the “Ensemble of HMMs technique”.

• Evaluating BAli-Phy on biological and simulated datasets
First Align, then Compute the Tree

S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCAGACCAGC
S3 = TAGCTGACCAGC
S4 = TCAGCACCAGACA

S1 = -AGGCTATCACCTGACCTCCA
S2 = TAG-CTATCAGACCAGC--
S3 = TAG-CT--------GACCAGC--
S4 = --------TCAGCACCAGACA
Simulation Studies

\[
\begin{align*}
S1 &= \text{-AGGCTATCACCTGACCTCCA} \\
S2 &= \text{TAGCTATCACGACC} \\
S3 &= \text{TAGCTGACC} \\
S4 &= \text{TCACGACC} \\
\end{align*}
\]

Unaligned Sequences

True tree and alignment

Estimated tree and alignment

Compare

Unaligned Sequences
FN: false negative (missing edge)
FP: false positive (incorrect edge)
50% error rate
Two-phase estimation

Alignment methods
- Clustal
- POY (and POY*)
- Probcons (and Probtree)
- Probalign
- MAFFT
- Muscle
- Di-align
- T-Coffee
- Prank (PNAS 2005, Science 2008)
- Opal (ISMB and Bioinf. 2007)
- Infernal (Bioinf. 2009)
- Etc.

Phylogeny methods
- Bayesian MCMC
- Maximum parsimony
- Maximum likelihood
- Neighbor joining
- FastME
- UPGMA
- Quartet puzzling
- Etc.

**RAxML**: heuristic for large-scale ML optimization
1000-taxon models, ordered by difficulty (Liu et al., 2009)
SATé “Family” of methods

• Iterative divide-and-conquer methods
  – Each iteration re-aligns the sequences using the current tree, running preferred MSA methods on small local subsets, and merging subset alignments
  – Each iteration computes an ML tree on the current alignment, under the GTR (Generalized Time Reversible) Markov model of evolution

• Note: these methods are “MSA boosters”, designed to improve accuracy and/or scalability of the base method

• We show results using MAFFT-l-ins-i to align subsets
Re-aligning on a tree

Decompose dataset → Align subsets

Estimate ML tree on merged alignment → Merge sub-alignments
**SATé and PASTA Algorithms**

1. Obtain initial alignment and estimated ML tree.
2. Estimate ML tree on new alignment.
3. Use tree to compute new alignment.
4. Repeat until termination condition, and return the alignment/tree pair with the best ML score.
1000-taxon models, ordered by difficulty – rate of evolution generally increases from left to right.

**SATé-1** 24 hour analysis, on desktop machines

(Similar improvements for biological datasets)

SATé-1 can analyze up to about 8,000 sequences.
1000-taxon models ranked by difficulty

SATé-1 and SATé-2 (Systematic Biology, 2012)

SATé-1: up to 8K
SATé-2: up to ~50K
SATé variants differ only in the decomposition strategy.

1. Decompose dataset
2. Align subsets
3. Merge sub-alignments
4. Estimate ML tree on merged alignment
PASTA merging: Step 1

Compute a spanning tree connecting alignment subsets
Use Opal (or Muscle) to merge adjacent subset alignments in the spanning tree.
PASTA merging: Step 3

Use transitivity to merge all pairwise-merged alignments from Step 2 into final alignment on entire dataset

**Overall:** $O(n \log(n) + L)$
Tree accuracy

1 million sequences:

- PASTA finished one iteration in 15 days
- PASTA tree had 6% error, compared to 5.6% when using true alignment
- Starting tree had 8.4% error
1kp: Thousand Transcriptome Project

First study (Wickett, Mirarab, et al., PNAS 2014) had ~100 species and ~800 genes, gene trees and alignments estimated using SATé, and a coalescent-based species tree estimated using ASTRAL.

Second study: Plant Tree of Life based on transcriptomes of ~1200 species, and more than 13,000 gene families (most not single copy).

Challenges:
Species tree estimation from conflicting gene trees
Gene tree estimation of datasets with > 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.
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:
Alignment of datasets with > 100,000 sequences with many 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.
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
One Hidden Markov Model for the entire alignment?
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
• 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?

HMM 1
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
RNASim: alignment error

![Graph showing mean alignment error across different datasets and methods.]

All methods given 24 hrs on a 12-core machine

Note: Mafft was run under default settings for 10K and 50K sequences and under Parttree for 100K sequences, and fails to complete under any setting for 200K sequences. Clustal-Omega only completes on 10K dataset.
Note: Mafft was run under default settings for 10K and 50K sequences and under Parttree for 100K sequences, and fails to complete under any setting for 200K sequences. Clustal-Omega only completes on 10K dataset.
Notes:

- We show alignment error using average of SP-FN and SP-FP.
- UPP variants have better alignment scores than PASTA.
- (Not shown: Total Column Scores – PASTA more accurate than UPP)
- No other methods tested could complete on these data
RNAsim Million Sequences: tree error

Using 12 processors:

- UPP(Fast,NoDecomp) took 2.2 days,
- UPP(Fast) took 11.9 days, and
- PASTA took 10.3 days
UPP vs. PASTA: impact of fragmentation

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

Under low rates of evolution, PASTA can still be highly accurate (data not shown).

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 time used (in hours) given 12 processors
Co-estimation would be much better!!!

S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACC CGC
S3 = TAGCTGACC CGC
S4 = TCACGACC CGACA

S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACC CGC
S3 = TAGCTGACC CGC
S4 = TCACGACC CGACA
What about BAli-Phy?

BAli-Phy (Redelings and Suchard): leading method for statistical co-estimation of alignments and trees

Like Bayesian phylogeny estimation, it is expected to be the most rigorous and accurate technique for estimating trees and alignments!
**BAli-Phy: Better than PASTA!**

**Alignment Accuracy (TC score)**

Simulated nucleotide datasets with 100 or 200 sequences (unpublished data from Mike Nute’s PhD dissertation).

*Averages over 10 replicates*
But: BAli-Phy is limited to small datasets

From [www.bali-phy.org/README.html](http://www.bali-phy.org/README.html), 5.2.1. Too many taxa?

“BAli-Phy is quite CPU intensive, and so we recommend using 50 or fewer taxa in order to limit the time required to accumulate enough MCMC samples. (Despite this recommendation, data sets with more than 100 taxa have occasionally been known to converge.) We recommend initially pruning as many taxa as possible from your data set, then adding some back if the MCMC is not too slow.”
Re-aligning on a tree

Decompose dataset

Align subsets: MAFFT

Estimate ML tree on merged alignment

Merge sub-alignments
Re-aligning on a tree

Decompose dataset

Align subsets: BAli-Phy??

Estimate ML tree on merged alignment

Merge sub-alignments
Results on 1000-sequence datasets (Comparing default PASTA to PASTA+BAli-Phy)

Decomposition to 100-sequence subsets, one iteration of PASTA+BAli-Phy
Results on 10,000-sequence datasets

(Comparing UPP variants where the backbone alignment is computed using either default PASTA or PASTA+BAli-Phy)
Benchmarking Statistical Multiple Sequence Alignment

Nute, Saleh, and Warnow 2018

Study design

• Goal: Evaluate Bali-Phy (Redelings and Suchard) on both biological and simulated datasets, in comparison to leading alignment methods on small protein sequence datasets (at most 27 sequences)

• Metrics: Modeller score (precision), SP-score (recall), Expansion ratio (normalized alignment length), and running time

• Datasets: 120 simulated datasets (6 model conditions) and 1192 biological datasets (4 biological benchmarks)

• Specific note: For each dataset, Bali-Phy was run independently on 32 processors for 48 hours, the burn-in was discarded, and the posterior decoding (PD) alignment was then computed. These Bali-Phy analyses used 230 CPU years on Blue Waters (supercomputer at NCSA).
Modeler vs SP-Score on 120 Simulated Datasets

BAli-Phy is best!
Expansion Ratios on 120 Simulated Datasets

BAli-Phy is best!
Modeler score vs SP-score on 1192 biological datasets

T-Coffee and PROMALS are best!

BAli-Phy good for Modeler score, but not so good for SP-Score (e.g., MAFFT better)
Modeler Score on 1192 Biological datasets

BAli-Phy has the best modeler score
SP-score on 1192 Biological datasets

BAli-Phy not competitive for SP-score (but best method depends on % ID)
Expansion Ratio on 1192 Biological datasets

BAli-Phy under-aligns

Method
- BAliPhy-PD
- Clustal
- ContraAlign
- DIAlign
- KAalign
- MAFFT-G
- MAFFT-Homologs
- Muscle
- Prank
- Prime
- ProbAlign
- ProbCons
- ProMals
- TCoffee

Identity Bin
- 0 ≤ ID < 0.15
- 0.15 ≤ ID < 0.25
- 0.25 ≤ ID < 0.5
- 0.5 ≤ ID < 1
Running Time on 4 biological datasets with 17 sequences each

Table 3: Running time information of a single 17-sequence data set in each of the biological benchmarks for different alignment methods, with methods roughly sorted by running time from fastest to slowest. The running times are rounded to the nearest hundredth of a second, and reflect wall clock time. The time reported for most methods is based on a single processor. However, BAli-Phy was run 32 independent times, and the running time reported is for a single run; MAFFT uses 4 threads, and Clustal-Omega uses 12 threads.

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>MattBench</th>
<th>Homstrad</th>
<th>Sisyphus</th>
<th>BAliBASE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data set</td>
<td>SF054</td>
<td>proteasome</td>
<td>AL00048098</td>
<td>BALBS213</td>
</tr>
<tr>
<td>Max. Seq. Len.</td>
<td>270</td>
<td>250</td>
<td>117</td>
<td>688</td>
</tr>
<tr>
<td>DiAlign</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Prime</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>KAlign</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
</tr>
<tr>
<td>Clustal-Omega</td>
<td>0.4</td>
<td>0.3</td>
<td>0.1</td>
<td>1.5</td>
</tr>
<tr>
<td>Muscle</td>
<td>0.5</td>
<td>0.4</td>
<td>0.1</td>
<td>1.0</td>
</tr>
<tr>
<td>MAFFT-G-INS-i</td>
<td>0.7</td>
<td>0.7</td>
<td>0.3</td>
<td>2.0</td>
</tr>
<tr>
<td>ProbAlign</td>
<td>1.7</td>
<td>1.4</td>
<td>0.4</td>
<td>7.9</td>
</tr>
<tr>
<td>ProbCons</td>
<td>3.1</td>
<td>2.6</td>
<td>0.6</td>
<td>12.6</td>
</tr>
<tr>
<td>CONTRAlign</td>
<td>5.8</td>
<td>6.2</td>
<td>1.4</td>
<td>42.0</td>
</tr>
<tr>
<td>Prank</td>
<td>48.5</td>
<td>1:16.1</td>
<td>9.4</td>
<td>4:14.7</td>
</tr>
<tr>
<td>Promals</td>
<td>14:11.5</td>
<td>12:22.1</td>
<td>5:06.2</td>
<td>24:03.2</td>
</tr>
<tr>
<td>T-Coffee</td>
<td>46:47.2</td>
<td>58:04.7</td>
<td>7:06.5</td>
<td>59:18.8</td>
</tr>
<tr>
<td>BAli-Phy</td>
<td>48:00:00.0</td>
<td>48:00:00.0</td>
<td>48:00:00.0</td>
<td>48:00:00.0</td>
</tr>
</tbody>
</table>

BAli-Phy benefits from a long running time.
Therefore, we used >2 months for each dataset.
Observations

• Bali-Phy is much more accurate than all other methods on simulated datasets

• Bali-Phy is generally less accurate than the top half of these methods on biological datasets, especially with respect to SP-score (recall)

• Average percent pairwise ID impacts all the measures of accuracy for all methods, and changes relative performance
We do not know why there is a difference in accuracy.

Most likely not an issue of failure of the MCMC analyses to converge (48 hours, 32 processors, small numbers of sequences).

Possible explanations:

1. Model misspecification (proteins don’t evolve under the Bali-Phy model)
2. Structural alignments and evolutionary alignments are different
3. The structural alignments are not correct (perhaps over-aligned)

All these explanations are likely true, but the relative contributions are unknown.
Final comments

• MSA is challenging, but algorithmic techniques can improve accuracy and scalability:
  – Dataset size can be addressed using good divide-and-conquer approaches.
  – Heterogeneity in sequence length can be addressed using “local alignment” approaches, such as profile HMMs, with ensembles of profile HMMs providing improved accuracy.

• Yet the differences between performance on biological and simulated datasets is troubling.
The Tree of Life: **Multiple Challenges**

Scientific challenges:

- Ultra-large multiple-sequence alignment
- Gene tree estimation
- Metagenomic classification
- 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: applied probability theory, graph theory, supercomputing, and heuristics

Testing: simulations and real data
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)

Evaluating BAli-Phy: Mike Nute and Ehsan Saleh (PhD students 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

Papers available at http://tandy.cs.illinois.edu/MSAproject.html