Challenges with Multiple Sequence Alignment

Tandy Warnow
University of Illinois at Urbana-Champaign
Multiple Sequence Alignment (MSA): *an important grand challenge*¹

\[
\begin{align*}
S1 &= AGGCTATCACCTGACCTCCA \\
S2 &= TAGCTATCACGACCGC \\
S3 &= TAGCTGACCGGC \\
\vdots \\
Sn &= TCACGACCGACA \\
\end{align*}
\]

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
Challenges

• Large numbers of sequences
• High rates of substitutions and indels
• Fragmentary sequences
• Statistical co-estimation (e.g., BAli-Phy) is great on simulated datasets, but not on protein benchmarks, and we don’t know why
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
Input: unaligned sequences

S1 = AGGCTATCACCCTGACCTCCA
S2 = TAGCTATCAGACCGC
S3 = TAGCTGACCGC
S4 = TCACGACCACGC
Phase 1: Alignment

S1 = AGGCTATCACCCTGACCTCCA  
S2 = TAGCTATCACCACCACGC  
S3 = TAGCTGACCGC  
S4 = TCACGACCACCA

S1 = -AGGCTATCACCCTGACCTCCA  
S2 = TAG-CTATCACCACGC--  
S3 = TAG-CT--------ACCGC--  
S4 = ""---TCAC---ACCGACCA
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
Simulation Studies

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

Unaligned Sequences

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

True tree and alignment

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

Compare

Estimated tree and alignment

S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCAGCGACGC
S3 = TAGCTGACCGC
S4 = TCACGACCGACA
## 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., Science 2009)
Challenges

• Large numbers of sequences
• High rates of substitutions and indels
Re-aligning on a tree

1. Estimate ML tree on merged alignment
2. Decompose dataset
3. Align subsets
4. Merge sub-alignments
Re-aligning on a tree

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

Algorithmic parameter: how to align subsets. Default: MAFFT L-INS-i.
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 – rate of evolution generally increases from left to right

SATé-1 24-hour analysis, on desktop machines (using MAFFT on subsets)
(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
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

- 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.
Challenges

• Large numbers of sequences
• High rates of substitutions and indels
• 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 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
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 time used (in hours) given 12 processors
Lessons (so far)

• First check your dataset for sequence length heterogeneity!
• If sequences are mostly the same length, PASTA is fine
• Otherwise compute an UPP alignment:
  • PASTA on the full-length sequences
  • then an ensemble of profile Hidden Markov Models to align the remaining sequences (e.g., fragments)
Co-estimation would be much better!!!
Modeler vs SP-Score 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)
Challenges

• Large numbers of sequences
• High rates of substitutions and indels
• Fragmentary sequences
• Statistical co-estimation (e.g., BAli-Phy) is great on simulated datasets, but not on protein benchmarks, and we don’t know why
The Tutorial (by Kodi Taraszka)

• PASTA for large-scale MSA and tree estimation
• Default settings, and when to change them
• Using the GUI
Re-aligning on a tree

1. Decompose dataset
2. Align subsets
3. Estimate ML tree on merged alignment
4. 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

Alignment

Tree

Repeat until termination condition, and return the alignment/tree pair with the best ML score
PASTA: easy to use GUI

https://github.com/smirarab/pasta