New methods for estimating multiple sequence alignments with high accuracy and scalability

Tandy Warnow
Departments of Computer Science and Bioengineering
Carl R. Woese Institute for Genomic Biology
The University of Illinois at Urbana-Champaign
http://tandy.cs.illinois.edu
Phylogenomic Pipelines

• Finding related genomic regions (homology detection)
• Multiple sequence alignment
• Maximum Likelihood phylogeny estimation for single genes
• Species tree estimation from multiple conflicting genes
• Answer biological questions
Input: unaligned sequences

\[
\begin{align*}
S1 &= \text{AGGCTATCACCTGACCTCCA} \\
S2 &= \text{TAGCTATCACGACCGC} \\
S3 &= \text{TAGCTGACCGC} \\
S4 &= \text{TCACGACCGACA}
\end{align*}
\]
Phase 1: Alignment

\[
S_1 = \text{AGGCTATCACCTGACCTCCA} \quad S_1 = \text{AGGCTATCACCTGACCTCCA}
\]
\[
S_2 = \text{TAGCTATCAGCCCGC} \quad S_2 = \text{TAGCTATCAGCCCGC}
\]
\[
S_3 = \text{TAGCTGAGCCGC} \quad S_3 = \text{TAGCTGAGCCGC}
\]
\[
S_4 = \text{TCACGACCACA} \quad S_4 = \text{TCACGACCACA}
\]
Phase 2: Construct tree

S1 = AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC
S3 = TAGCTGACC
S4 = TCACGACC

S1 = -AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGACCGC
S3 = TAGCTGACC
S4 = TCACGACC

S1

S2

S3

S4
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
Avian Phylogenomics Project

E Jarvis, HHMI
MTP Gilbert, Copenhagen
G Zhang, BGI
T. Warnow, UT-Austin
S. Mirarab, UT-Austin
Md. S. Bayzid, UT-Austin

Jarvis, Mirarab, et al., *Science* 2014 (14,000 loci, each with about 50 sequences)

- **Multiple sequence alignment for each locus computed using PASTA**
- Gene trees computed using RAxML
- Species trees computed using ExaML and coalescent-based pipeline (statistical binning and MP-EST)

Plus many many other people...
1kp: Thousand Transcriptome Project

Wickett, Mirarab, et al. PNAS 2014 (about 100 species and 800 genes):

- Multiple sequence alignment for each locus computed using PASTA
- Gene trees computed using RAxML
- Species tree estimation computed using concatenation (RAxML) and coalescent-based method (ASTRAL).

G. Ka-Shu Wong
U Alberta

J. Leebens-Mack
U Georgia

N. Wickett
Northwestern

N. Matasci
iPlant

T. Warnow,
UIUC

S. Mirarab,
UT-Austin

N. Nguyen,
UT-Austin

Plus many many other people…
1kp: Thousand Transcriptome Project

- Second analysis underway, much larger dataset (~1200 species and ~1000 loci), with potentially multiple copies of each gene in each individual (due to gene duplication and loss)

Challenge: Construct multiple sequence alignment and tree for more than 100,000 sequences
## Two-phase estimation

<table>
<thead>
<tr>
<th>Alignment methods</th>
<th>Phylogeny methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Clustal</td>
<td>• Bayesian MCMC</td>
</tr>
<tr>
<td>• Di-align</td>
<td>• Maximum parsimony</td>
</tr>
<tr>
<td>• Infernal</td>
<td>• Maximum likelihood</td>
</tr>
<tr>
<td>• MAFFT</td>
<td>• Neighbor joining</td>
</tr>
<tr>
<td>• Muscle</td>
<td>• FastME</td>
</tr>
<tr>
<td>• Opal</td>
<td>• UPGMA</td>
</tr>
<tr>
<td>• PAGAN</td>
<td>• Quartet puzzling</td>
</tr>
<tr>
<td>• PASTA</td>
<td>• Etc.</td>
</tr>
<tr>
<td>• POY (and POY*)</td>
<td></td>
</tr>
<tr>
<td>• Prank</td>
<td></td>
</tr>
<tr>
<td>• Probalign</td>
<td></td>
</tr>
<tr>
<td>• Probcons (and Probtree)</td>
<td></td>
</tr>
<tr>
<td>• T-Coffee</td>
<td></td>
</tr>
<tr>
<td>• <em>UPP</em></td>
<td></td>
</tr>
<tr>
<td>• Etc.</td>
<td></td>
</tr>
</tbody>
</table>

**RAxML**: heuristic for large-scale ML optimization
Simulation Studies

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

S1 = -AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGCCGAC
S3 = TAGCTGACCGC
S4 = TCACGACCGACA

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

S1 = -AGGCTATCACCTGACCTCCA
S2 = TAGCTATCACGCCGAC
S3 = TAGCTGACCGC
S4 = TCACGACCGACA
Quantifying Error

FN: false negative (missing edge)
FP: false positive (incorrect edge)

50% error rate
1000-taxon models, ordered by difficulty (Liu et al., 2009)
Consequences of MSA difficulties

- Biologists restrict dataset size or use inadequate methods to be able to analyze their data.
- Alignment accuracy is reduced, which impacts downstream analyses:
  - Protein structure and function prediction
  - Metagenomic taxon identification
  - Phylogeny estimation
  - Detection of positive selection
- Scientific discoveries are jeopardized.
Multiple Sequence Alignment (MSA): *an important grand challenge*¹

S1 = AGGCTATCACCTGACCTCCA  
S2 = TAGCTATCACGACCGC  
S3 = TAGCTGACCGC  
...  
Sn = TCACGACCGACA

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
Today’s talk: MSA estimation

- **PASTA**: divide-and-conquer method to “boost” a MSA method

- Default PASTA (using MAFFT): can align 1,000,000 sequences with high accuracy and speed
Today’s talk: MSA estimation

- PASTA: divide-and-conquer method to “boost” a MSA method
- Default PASTA (using MAFFT): can align 1,000,000 sequences with high accuracy and speed
- UPP: improves on PASTA in the presence of substantial sequence length heterogeneity - uses an ensemble of Hidden Markov Models.
1000-taxon models, ordered by difficulty (Liu et al., 2009)
Key Observations

• Datasets that are large and have high rates of evolution are difficult to align accurately.
• However, datasets with slow rates of evolution can be aligned with high accuracy.
• Not all MSA methods can run on large datasets (and some cannot even run on moderate-sized datasets).

These observations suggest *divide-and-conquer* to boost MSA methods to larger datasets.
Re-aligning on a tree (boosting an MSA method)

Decompose dataset

Align subsets

Merge subset-alignments

Estimate ML tree on merged alignment
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
(Similar improvements for biological datasets)
SATé-1 can analyze up to about 8,000 sequences.
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: even better than SATé-2

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)
PASTA Running Time and Scalability

- One iteration
- Using
  - 12 cpus
  - 1 node on Lonestar TACC
  - Maximum 24 GB memory
- Showing wall clock running time
  - ~ 1 hour for 10k taxa
  - ~ 17 hours for 200k taxa
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
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, many of which are fragmentary
Multiple Sequence Alignment (MSA): an important grand challenge

\[ S_1 = \text{AGGCTATCACCTGACCTCCA} \quad S_1 = -\text{AGGCTATCACCTGACCTCCA} \]
\[ S_2 = \text{TAGCTATCACGACGCGC} \quad S_2 = \text{TAG-CTATCAC--GACCGC--} \]
\[ S_3 = \text{TAGCTGACCGC} \quad S_3 = \text{TAG-CT--------GACCGC--} \]
\[ \ldots \]
\[ S_n = \text{TCACGACCGACA} \quad S_n = \text{--------TCAC--GACCGACA} \]

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
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
One Hidden Markov Model for the entire alignment?
Or 2 HMMs?
Or 4 HMMs?
Or all 7 HMMs?
UPP Algorithmic Approach

1. Select small 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
RNASim Million Sequences: tree error

Using 12 TACC processors:

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

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.

1000M2 model condition
UPP Running Time

Wall clock align time (hr)

Number of sequences

Wall-clock time used (in hours) given 12 processors
PASTA and UPP: boosters of MSA methods

• PASTA
  – Combines iteration and divide-and-conquer to “boost” a preferred MSA method to large datasets; we showed results based on MAFFT

• UPP
  – Step 1: Constructs a “backbone” tree and an alignment on a small random subset of the sequences
  – Step 2: Aligns all the remaining sequences to the backbone alignment
  – We showed results where default PASTA computed the backbone alignment and tree (which is based on MAFFT).
PASTA and UPP: boosters of MSA methods

- **PASTA**
  - Combines iteration and divide-and-conquer to “boost” a preferred MSA method to large datasets; *we showed results based on MAFFT*

- **UPP**
  - Step 1: Constructs a “backbone” tree and an alignment on a small random subset of the sequences
  - Step 2: Aligns all the remaining sequences to the backbone alignment
  - *We showed results where default PASTA computed the backbone alignment and tree.*

**Note:** PASTA and UPP can be used with any MSA method.
Statistical co-estimation would be much better!!!

\[
\begin{align*}
S1 &= AGGCTATCACCTGACCTCCA & S1 &= -AGGCTATCACCTGACCTCCA \\
S2 &= TAGCTATCAGGACCGC & S2 &= TAG-CTATCAC--GACCGC-- \\
S3 &= TAGCTGACCGC & S3 &= TAG-CT----GACCGC-- \\
S4 &= TCACGACCGACA & S4 &= --------TCAC--GACCGACA \\
\end{align*}
\]
BAli-Phy: leading statistical co-estimation method

- BAli-Phy (Redelings and Suchard, 2005):
  - Statistical co-estimation of the sequence alignment and the tree, using Bayesian MCMC.
  - Output can be a multiple sequence alignment, a phylogeny, or both, and can give estimate of uncertainty in each one.
Each dataset has 100 or 200 sequences. To run BAli-Phy on a single dataset, we used 32 Blue Waters processors and ran BAli-Phy on each processor for 24 hours. We then collected all the samples from all the processors, and computed the posterior decoding (PD) alignment.

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

- BAli-Phy is computationally intensive:
  - 63 sequence dataset (Gaya et al., 2011) took 3 weeks
  - Largest dataset analyzed had 117 sequences (McKenzie et al., 2014)

- BAli-Phy is not scalable:
  - Our study shows it breaks somewhere before 500 sequences (numerical issues possibly)

- 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.
Acknowledgments

Papers available at [http://tandy.cs.illinois.edu/papers.html](http://tandy.cs.illinois.edu/papers.html)
PASTA and UPP at [https://github.com/smirarab](https://github.com/smirarab)

**Funding:** NSF ABI-1458652 and III:AF:1513629, a Founder Professorship from the Grainger Foundation, and HHMI (to S.M.)

**Computational support:** TACC (PASTA and UPP)