Multiple sequence alignment methods: evidence from data

Tandy Warnow
Alignment Error/Accuracy

• SPFN: percentage of homologies in the true alignment that are *not* recovered (false negative homologies)
• SPFP: percentage of homologies in the estimated alignment that are false (false positive homologies)

• TC: total number of columns correctly recovered
• SP-score: percentage of homologies in the true alignment that are recovered
• Pairs score: 1-(avg of SP-FN and SP-FP)
Benchmarks

• Simulations: can control everything, and true alignment is not disputed
  – Different simulators

• Biological: can’t control anything, and reference alignment might not be true alignment
  – BAliBASE, HomFam, Prefab
  – CRW (Comparative Ribosomal Website)
Alignment Methods (Sample)

- Clustal-Omega
- MAFFT
- Muscle
- Opal
- Prank/Pagan
- Probcons

Co-estimation of trees and alignments
- Bali-Phy and Alifritz (statistical co-estimation)
- SATe-1, SATe-2, and PASTA (divide-and-conquer co-estimation)
- POY and Beetle (treelength optimization)
Other Criteria

- Tree topology error
- Tree branch length error
- Gap length distribution
- Insertion/deletion ratio
- Alignment length
- Number of indels
How does the guide tree impact accuracy?

• Does improving the accuracy of the guide tree help?
• Do all alignment methods respond identically? (Is the same guide tree good for all methods?)
• Do the default settings for the guide tree work well?
Alignment criteria

• Does the relative performance of methods depend on the alignment criterion?
• Which alignment criteria are predictive of tree accuracy?
• How should we design MSA methods to produce best accuracy?
Choice of best MSA method

- Does it depend on type of data (DNA or amino acids?)
- Does it depend on rate of evolution?
- Does it depend on gap length distribution?
- Does it depend on existence of fragments?
ITS2 sequences are added into this backbone MSA, using the aligned to build a backbone MSA, and then the new ITS1 and as e to full-length sequences are first consuming method, L-INS-i, is not always the most accurate appropriate to the problem of interest. The most time-
cases reordering according to similarity using the between
776
step
adheref
98h)
0.9994 3.76h 36.2min
0.9707 23.4 daysb 2.43 daysb
0.9604 1.32h 1.44h
0.2779 15.5h 1.60h
0.9969 6.67 days 18.3h

Case 2

mafft --6merpair --addfragments frags existingmsa
mafft --6merpair --add

Case 3

mafft --6merpair --addfragments frags existingmsa


*Wall-clock time with 10 cores. Command-line argument for parallel processing is --thread 10.

bFull command-line options are as follows: mafft --localpair --weighti 0 --add frags existingmsa.

Table 2. Comparison of Different Options Using the 16S.B.ALL Data Set (Mirarab et al. 2012).

<table>
<thead>
<tr>
<th>Data</th>
<th>Method</th>
<th>Accuracy</th>
<th>CPU Time</th>
<th>Actual Timea</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case 1</td>
<td>mafft --multipair --addfragments frags existingmsa</td>
<td>0.9969</td>
<td>6.67 days</td>
<td>18.3 h</td>
</tr>
<tr>
<td></td>
<td>mafft --6merpair --addfragments frags existingmsa</td>
<td>0.9949</td>
<td>3.76 h</td>
<td>36.2 min</td>
</tr>
<tr>
<td></td>
<td>mafft --localpair --add frags existingmsa</td>
<td>0.9707</td>
<td>23.4 daysb</td>
<td>2.43 daysb</td>
</tr>
<tr>
<td></td>
<td>mafft --6merpair --add frags existingmsa</td>
<td>0.9604</td>
<td>1.32 h</td>
<td>1.44 h</td>
</tr>
<tr>
<td></td>
<td>profile alignment</td>
<td>0.2779</td>
<td>15.5h</td>
<td>1.60 h</td>
</tr>
<tr>
<td>Case 2</td>
<td>mafft --6merpair --addfragments frags existingmsa</td>
<td>0.9969</td>
<td>4.54 h</td>
<td>33.8 min</td>
</tr>
<tr>
<td>Case 3</td>
<td>mafft --6merpair --addfragments frags existingmsa</td>
<td>0.9949</td>
<td>1.79 days</td>
<td>5.91 h</td>
</tr>
</tbody>
</table>

Note.—The estimated alignments were compared with the CRW alignment to measure the accuracy (the number of correctly aligned letters/the number of aligned letters in the CRW alignment). Calculations were performed on a Linux PC with 2.67 GHz Intel Xeon E7-8837/256 GB RAM (for the case marked with superscript alphabet “b”), or on a Linux PC with 3.47 GHz Intel Xeon XS690/48 GB RAM (for the other cases).
Case 1: 13,822 sequences in the existing alignment × 13,821 fragments;
Case 2: 1,000 sequences in the existing alignment × 138,210 fragments;
Case 3: 13,822 sequences in the existing alignment × 138,210 fragments.

From Katoh and Standley, 2013 (dealing with fragmentary sequences)
Important!

• Each method can be run in different ways – so you need to know the exact command used, to be able to evaluate performance. (You also need to know the version number!)
Clustal-Omega study

• Clustal-Omega (Seivers et al., Molecular Systems Biology 2011) is the latest in the Clustal family of MSA methods
• Clustal-Omega is designed primarily for amino acid alignment, but can be used on nucleotide datasets
• Alignment criterion: TC (column score)
• Datasets: biological with structural alignments
Table II Prefab results

<table>
<thead>
<tr>
<th>Aligner</th>
<th>0 &lt; %ID ≤ 100 families</th>
<th>0 ≤ %ID ≤ 20 families</th>
<th>20 ≤ %ID ≤ 40 families</th>
<th>40 ≤ %ID ≤ 70 families</th>
<th>70 ≤ %ID ≤ 100 families</th>
<th>Total time (s) (1682 families)</th>
<th>Consistency</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSAprobs</td>
<td>0.737</td>
<td>0.591</td>
<td>0.889</td>
<td>0.965</td>
<td>0.971</td>
<td>51 286.00</td>
<td>Yes</td>
</tr>
<tr>
<td>Mafft</td>
<td>0.721</td>
<td>0.569</td>
<td>0.876</td>
<td>0.961</td>
<td>0.979</td>
<td>4544.45</td>
<td>Yes</td>
</tr>
<tr>
<td>(auto)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Probalign</td>
<td>0.719</td>
<td>0.563</td>
<td>0.881</td>
<td>0.961</td>
<td>0.977</td>
<td>35 117.30</td>
<td>Yes</td>
</tr>
<tr>
<td>Probcons</td>
<td>0.717</td>
<td>0.562</td>
<td>0.876</td>
<td>0.955</td>
<td>0.972</td>
<td>46 908.30</td>
<td>Yes</td>
</tr>
<tr>
<td>T-Coffee</td>
<td>0.710</td>
<td>0.558</td>
<td>0.865</td>
<td>0.950</td>
<td>0.972</td>
<td>175 789.00</td>
<td>Yes</td>
</tr>
<tr>
<td>ClustalΩ</td>
<td>0.700</td>
<td>0.535</td>
<td>0.866</td>
<td>0.967</td>
<td>0.980</td>
<td>1698.06</td>
<td>No</td>
</tr>
<tr>
<td>MUSCLE</td>
<td>0.677</td>
<td>0.507</td>
<td>0.850</td>
<td>0.946</td>
<td>0.976</td>
<td>2068.56</td>
<td>No</td>
</tr>
<tr>
<td>Mafft</td>
<td>0.677</td>
<td>0.513</td>
<td>0.836</td>
<td>0.961</td>
<td>0.979</td>
<td>225.56</td>
<td>No</td>
</tr>
<tr>
<td>Kalign</td>
<td>0.649</td>
<td>0.474</td>
<td>0.817</td>
<td>0.957</td>
<td>0.979</td>
<td>80.81</td>
<td>No</td>
</tr>
<tr>
<td>ClustalW2</td>
<td>0.617</td>
<td>0.430</td>
<td>0.797</td>
<td>0.933</td>
<td>0.975</td>
<td>3433.53</td>
<td>No</td>
</tr>
<tr>
<td>Dialign</td>
<td>0.595</td>
<td>0.398</td>
<td>0.783</td>
<td>0.940</td>
<td>0.974</td>
<td>18 909.70</td>
<td>No</td>
</tr>
<tr>
<td>PRANK</td>
<td>0.586</td>
<td>0.390</td>
<td>0.767</td>
<td>0.951</td>
<td>0.978</td>
<td>351 498.00</td>
<td>No</td>
</tr>
<tr>
<td>FSA</td>
<td>0.534</td>
<td>0.277</td>
<td>0.791</td>
<td>0.965</td>
<td>0.976</td>
<td>229 391.00</td>
<td>No</td>
</tr>
</tbody>
</table>

Total column scores (TC) are shown for different percent identity ranges; the second column is the average score over all test cases. The total run time in seconds is shown in the second last column. The last column indicates if the method is consistency based.

TC Score shown (larger is better) on Prefab structural benchmark of AA alignments
Note that best performing method depends on the “%ID” (measure of similarity)

From Seivers et al., Molecular Systems Biology 2011
BAliBASE is a collection of structurally-based alignments of amino acid sequences

From Seivers et al., Molecular Systems Biology 2011
The columns show total column score (TC) and total run time in seconds for groupings of small (<3000 sequences), medium (3000–10 000 sequences) and large (> 10 000 sequences) HomFam test cases.

HomFam is a set of structurally-based alignments of sets of amino acid sequences

From Seivers et al., Molecular Systems Biology 2011
Observations

• Relative and absolute accuracy (wrt TC score) impacted by degree of heterogeneity and dataset size
• Some methods cannot run on large datasets
• On small datasets, Clustal-Omega not as accurate as best methods (Probalign, MAFFFT, and MSAProbs)
• On large datasets, Clustal-Omega more accurate than other methods
Questions

• How do the different co-estimation methods compare with respect to tree error and alignment error?
  – POY and BeeTLe (tree-length optimization methods)
  – BAli-Phy and Alifritz (statistical co-estimation methods)
  – SATe-1, SATe-2, and PASTA (iterative)
Results about treelength

• Yes – Solving treelength using affine gap penalties is better than using simple gap penalties.
• However - alignment accuracy is very low.
• Tree accuracy is good, if compared to maximum parsimony (MP) analyses of good alignments
• Tree accuracy is bad, if compared to maximum likelihood (ML) analyses of good alignments
• Not examined: better gap penalties
SATe “Family”

• Iterative divide-and-conquer methods
  – Each iteration uses the current tree with divide-and-conquer, to produce an alignment (running preferred MSA methods on subsets, and aligning alignments together)
  – Each iteration computes an ML tree on the current alignment, under Markov models of evolution that do not consider indels
SATe-I and SATe-II

• SATe (Simultaneous Alignment and Tree Estimation) was introduced in Liu et al., Science 2009; SATe-II (Liu et al. Systematic Biology 2012) was an improvement in accuracy and speed.

• Basic approach: iterate between alignment and tree estimation (using standard ML analysis on alignments)

• Stop after 24 hours, and return alignment/tree pair with best ML score

• Designed and tested only on nucleotide sequences
Obtain initial alignment and estimated ML tree

Estimate ML tree on new alignment

Use tree to compute new alignment
SATé iteration
(actual decomposition produces 32 subproblems)

Decompose based on input tree

Estimate ML tree on merged alignment

Align subproblems

Merge subproblems
1000 taxon models, ordered by difficulty

24 hour SATé-I analysis, on desktop machines

(*Similar improvements for biological datasets*)
Table 2 shows alignment accuracy on the AA datasets. Due to dataset sizes, Muscle and SATe-II failed to complete on two of the HomFam datasets, so we separate out the results for these two datasets from the remaining 17 HomFam datasets.

PASTA had the best pairs score or was tied for the best pairs score for both HomFam and AA-10 datasets. Mafft had the best TC score for HomFam(17), but PASTA was very close. For HomFam(2), PASTA had the best TC score and Mafft was a close second. On AA-10 datasets, SATe-II had the best TC score and was closely trailed by Mafft and PASTA.

Comparison to SATe-II on 50,000-taxon dataset.

SATe-II could not finish even one iteration on the RNASim with 50,000 sequences running for 24 hr and given 12 CPUs on TACC. However, we were able to...
run two iterations of SATe-II on a separate machine with no running time limits (12 Quad-Core AMD Opteron processors, 256GB of RAM memory). Given 12 CPUs, two iterations of SATe-II took 137 hr, compared to 10 hr for PASTA. However, the resulting SATe-II alignment recovered only 30 columns entirely correctly while PASTA recovered 311 columns. The pairs score of SATe-II was extremely poor (38.2%), while PASTA was quite accurate (81.0%). The tree produced by SATe-II had higher error than PASTA (12.6% versus 8.2% FN rate).

Impact of varying algorithmic parameters. We compared results obtained using four different starting trees: a random tree, the ML tree on the Mafft-PartTree alignment, PASTA's default starting tree, and the true (model) tree (see Table 3). After one iteration, PASTA alignments and trees based on our starting tree or true tree had roughly the same accuracy, and the starting tree based on Mafft-PartTree resulted in only a slightly worse tree (1% higher FN rate). However, using a random tree resulted in much higher tree error rates (52.3% error), and alignments that were also less accurate. Interestingly, after three iterations of PASTA, no noticeable difference could be detected between results from various starting trees. Thus, PASTA is robust to the choice of the starting tree.

We also evaluated the impact of changing the alignment subset size (Table 4); these analyses showed that using alignment subsets of only 50 sequences improved the TC score and running time substantially, and only slightly changed the pairs score or tree error score. Although these analyses were performed only for two datasets, they suggest the possibility that improved results might be obtained through smaller alignment subsets.

Running Time. Figure 3 compares the running time (in hours) of different alignment methods. Note that PASTA was faster than SATe-II in all cases and could analyze datasets that SATe-II could not (i.e., the Table 2. Alignment Accuracy on AA Datasets

<table>
<thead>
<tr>
<th>Method</th>
<th>AA-10</th>
<th>HomFam(17)</th>
<th>HomFam(2)</th>
<th>Pairs score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clustal-O</td>
<td>78</td>
<td>88</td>
<td>29</td>
<td>0.76</td>
</tr>
<tr>
<td>Muscle</td>
<td>48</td>
<td>51</td>
<td>X</td>
<td>0.70</td>
</tr>
<tr>
<td>Mafft</td>
<td>81</td>
<td>103</td>
<td>32</td>
<td>0.76</td>
</tr>
<tr>
<td>Initial</td>
<td>54</td>
<td>95</td>
<td>16</td>
<td>0.75</td>
</tr>
<tr>
<td>SATé-II</td>
<td>83</td>
<td>73</td>
<td>X</td>
<td>0.75</td>
</tr>
<tr>
<td>PASTA</td>
<td>80</td>
<td>102</td>
<td>36</td>
<td>0.76</td>
</tr>
</tbody>
</table>

We show TC (the number of correctly aligned sites, left) and the pairs score (the average of the SP-score and modeler score, right). X indicates that a method failed to run on a particular dataset given the computational constraints. “Initial” corresponds to the alignment approach used to obtain the starting tree of PASTA (HMMER failed to align one sequence in the 16S.T dataset). All values shown are averages over all datasets in each category.

Boldface indicates the best values for each model condition.

Comparison of PASTA to SATe-II and other alignments on AA datasets. From Mirarab et al., J. Computational Biology 2014
Table S26: Comparison of alignment errors for two-phase and coestimation methods. SATé<sup>BML</sup> is the best likelihood method for SATé run until no improvements can be found with CT-5 proposals and with either all two-phase starting/tree alignment pairs or just RAxML(ClustalW). The four model conditions for which ALIFRITZ had not yet reported any ML trees are marked “N/A”. We report results for both the posterior decoding alignment and the MAP alignment from BAli-Phy’s MCMC walk. \( n = 1 \) for all values in the table.

<table>
<thead>
<tr>
<th>Model</th>
<th>SATé&lt;sup&gt;BML&lt;/sup&gt;</th>
<th>MAFFT</th>
<th>Prank+GT</th>
<th>Muscle</th>
<th>ClustalW</th>
<th>BAli-Phy posterior-decoding</th>
<th>BAli-Phy MAP</th>
<th>ALIFRITZ</th>
</tr>
</thead>
<tbody>
<tr>
<td>100L1</td>
<td>21.3</td>
<td>20.1</td>
<td>41.7</td>
<td>30.6</td>
<td>54.1</td>
<td>12.4</td>
<td>15.3</td>
<td>N/A</td>
</tr>
<tr>
<td>100L2</td>
<td>1.7</td>
<td>1.9</td>
<td>1.7</td>
<td>2.4</td>
<td>12.9</td>
<td>1.1</td>
<td>1.8</td>
<td>11.0</td>
</tr>
<tr>
<td>100M1</td>
<td>31.8</td>
<td>29.2</td>
<td>63.2</td>
<td>39.3</td>
<td>56.9</td>
<td>29.0</td>
<td>29.0</td>
<td>N/A</td>
</tr>
<tr>
<td>100M2</td>
<td>12.1</td>
<td>17.5</td>
<td>13.1</td>
<td>15.9</td>
<td>39.1</td>
<td>6.1</td>
<td>7.9</td>
<td>N/A</td>
</tr>
<tr>
<td>100M3</td>
<td>3.3</td>
<td>4.0</td>
<td>3.1</td>
<td>3.3</td>
<td>8.5</td>
<td>2.3</td>
<td>3.0</td>
<td>16.4</td>
</tr>
<tr>
<td>100S1</td>
<td>27.8</td>
<td>29.4</td>
<td>35.5</td>
<td>39.4</td>
<td>40.9</td>
<td>12.0</td>
<td>13.6</td>
<td>N/A</td>
</tr>
<tr>
<td>100S2</td>
<td>13.4</td>
<td>19.5</td>
<td>13.4</td>
<td>18.0</td>
<td>27.3</td>
<td>6.9</td>
<td>9.2</td>
<td>47.5</td>
</tr>
<tr>
<td>Average</td>
<td>15.9</td>
<td>17.4</td>
<td>24.5</td>
<td>21.3</td>
<td>34.2</td>
<td>10.0</td>
<td>11.4</td>
<td>N/A</td>
</tr>
</tbody>
</table>

✓ ✓ ✗

Alignment error is average of SPFN and SPFP. However, Bali-Phy could not run on datasets with 500 or 1000 sequences. Results from Liu et al., Science 2009.
Table S27: **Comparison of tree errors for two-phase and coestimation methods.** SATé\textsuperscript{BML} is the best likelihood method for SATé run until no improvements can be found with CT-5 proposals and with either all two-phase starting/tree alignment pairs or just RAxML(ClustalW). The four model conditions for which ALIFRITZ had not yet reported any ML trees are marked “N/A”. We report results for both the majority consensus tree and MAP tree computed from from BAli-Phy’s partially completed MCMC walk. Since the consensus tree is not usually binary, we also give the “FP” rate for BAli-Phy. $n = 1$ for all values in the table.

<table>
<thead>
<tr>
<th>Model</th>
<th>RAxML(TrueAln)</th>
<th>SATé\textsuperscript{BML}</th>
<th>RAxML(MAFFT)</th>
<th>RAxML(Prank+GT)</th>
<th>RAxML(Muscle)</th>
<th>RAxML(ClustalW)</th>
<th>BAi-Phy majority consensus</th>
<th>BAi-Phy MAP</th>
<th>ALIFRITZ</th>
</tr>
</thead>
<tbody>
<tr>
<td>100L1</td>
<td>12.4</td>
<td>15.5</td>
<td>12.4</td>
<td>29.9</td>
<td>12.4</td>
<td>26.8</td>
<td>16.5</td>
<td>15.5</td>
<td>N/A</td>
</tr>
<tr>
<td>100L2</td>
<td>2.1</td>
<td>2.1</td>
<td>2.1</td>
<td>2.1</td>
<td>2.1</td>
<td>4.3</td>
<td>2.1</td>
<td>2.1</td>
<td>4.3</td>
</tr>
<tr>
<td>100M1</td>
<td>3.1</td>
<td>13.4</td>
<td>17.5</td>
<td>33.0</td>
<td>14.4</td>
<td>25.8</td>
<td>42.3</td>
<td>41.2</td>
<td>N/A</td>
</tr>
<tr>
<td>100M2</td>
<td>6.2</td>
<td>5.2</td>
<td>6.2</td>
<td>6.2</td>
<td>5.2</td>
<td>6.2</td>
<td>5.2</td>
<td>5.2</td>
<td>N/A</td>
</tr>
<tr>
<td>100M3</td>
<td>5.2</td>
<td>5.2</td>
<td>6.2</td>
<td>4.1</td>
<td>4.1</td>
<td>6.2</td>
<td>5.2</td>
<td>4.1</td>
<td>6.2</td>
</tr>
<tr>
<td>100S1</td>
<td>11.5</td>
<td>11.5</td>
<td>13.5</td>
<td>24.0</td>
<td>13.5</td>
<td>18.8</td>
<td>17.7</td>
<td>17.7</td>
<td>N/A</td>
</tr>
<tr>
<td>100S2</td>
<td>3.2</td>
<td>2.2</td>
<td>3.2</td>
<td>2.2</td>
<td>4.3</td>
<td>7.5</td>
<td>2.2</td>
<td>2.2</td>
<td>7.5</td>
</tr>
<tr>
<td>Average</td>
<td>6.2</td>
<td>7.8</td>
<td>8.7</td>
<td>14.5</td>
<td>8.0</td>
<td>13.6</td>
<td>13.0</td>
<td>12.6</td>
<td>N/A</td>
</tr>
</tbody>
</table>

✓ ✓ ✓ ✓ ✓ ✓

**Problem:** BAli-Phy failure to converge, despite multi-week analyses. Results from Liu et al., Science 2009.
Results for co-estimation methods

• Optimizing treelength (POY and BeeTLe) doesn’t produce good alignments, and trees are not as good as those obtained using ML on standard MSA methods.

• Statistical co-estimation of alignments and trees under models of evolution that include indels can produce highly accurate alignments and trees – but running time is a big issue.

• SATé and PASTA are iterative techniques for co-estimating alignments and trees, and produce good results... but have no statistical guarantees.
Impact of guide tree

• Most MSA methods use “progressive alignment” techniques, that
  – First compute a guide tree $T$
  – Align the sequences from the bottom-up using the guide tree

• Hence, there is a potential for the guide tree to impact the final alignment.

• Many authors have studied this issue... here’s our take on it (Nelesen et al., PSB 2008)
Nelesen et al., PSB 2008

• Pacific Symposium on Biocomputing, 2008
• MSA methods:
  – ClustalW, Muscle, Probcons, MAFFT, and FTA (Fixed Tree Alignment, using POY on the guidetree)
• Guide trees:
  – Default for each method
  – Two different UPGMA trees
  – Probtree (ML on Probcons+GT alignment)
• Examined results on simulated datasets with respect to alignment error and tree error
nelesen
alignment, and let \( \hat{A} \) be the estimated alignment. Then the SP-error rate is

\[ \left| \text{Pairs}(A^*) - \text{Pairs}(\hat{A}) \right| \]

expressed as a percentage; thus the SP-error is the percentage of the pairs of truly homologous nucleotides that are unpaired in the estimated alignment. However, it is possible for the SP-error rate to be 0, and yet have different alignments.

3.2. Results.

We first examine the guide trees with respect to their topological accuracy. As shown in Figure 2, the accuracy of guide trees differs significantly, with the ProbCons default tree generally the least accurate, and our "probtree" guide tree the most accurate; the two UPGMA guide trees have very similar accuracy levels.

Figure 2. Guide tree topological error rates, averaged over all model conditions and replicates. (1) ClustalW default, (2) ProbCons default, (3) Muscle default, (4) upgma1, (5) upgma2, and (6) probtree.

In Figure 3 we examine the accuracy of the alignments obtained using different MSA methods on these guide trees. Surprisingly, despite the large differences in topological accuracy of the guide trees, alignment accuracy (measured using SP-error) for a particular alignment method varies relatively little between alignments estimated from different guide trees.

For example, two ClustalW alignments or two Muscle alignments will have essentially the same accuracy scores, independent of the guide tree. The biggest factor impacting the SP-error of the alignment is the MSA method. Generally, ProbCons is the most accurate and ClustalW is the least.

We then examined the impact of changes in guide tree on the accuracy of the resultant RAxML-based phylogeny (see Figure 4). In all cases, for a given MSA method, phylogenetic estimations obtained when the guide tree

Figure from Nelesen et al., Pacific Symposium on Biocomputing, 2008
Figure 3. SP-error rates of alignments. M(guide tree) indicates multiple sequence alignment generated using the indicated guide tree.

Figure 4. Missing edge rate of estimated trees. R(M(guide tree)) indicates RAxML run on the alignment generated by the multiple sequence alignment method using the guide tree indicated. R(true-aln) indicates the tree generated by RAxML when given the true alignment.

Figure from Nelesen et al., Pacific Symposium on Biocomputing, 2008
Observations

• Guide tree choice did not seem to affect alignment SP error
• Guide tree choice affected tree error – but impact depended on dataset size (25 vs. 100) and MSA method.
• Probcons very impacted by guide tree (and that may be because its own default guide tree is poorly chosen).
• FTA very impacted by guide tree. Note that FTA on the true tree is MORE accurate than ML on the true alignment.
• For analyses of 100-taxon datasets, Probtree is a good guide tree.
Another study...

• Prank (Loytynoja and Goldman, Science 2008) is a “phylogeny aware” progressive alignment strategy.
• Their study focused on evaluating MSAs with respect to TC score, but also atypical criteria, such as:
  – Gene tree branch length estimation
  – Alignment length estimation (compression issue)
  – Insertion/deletion ratio
  – Number of insertions/deletions
• They explored very small simulated datasets, evolving sequences down trees.
Alignment accuracy errors are reduced by a phylogeny-aware algorithm. (A) Inferred branch lengths at different depths in the tree (B) Errors (C) insertion/deletion ratio, (D) gap overlap, (E) total length of the alignment, and (F) the proportion of columns correctly recovered. (G) The phylogeny-aware method PRANK is preferred by, some molecular biologists work-specifically designed for evolutionary analyses inferences. We believe that alignment methods specifically designed for evolutionary analyses are number of (A) insertions and (B) deletions, (C) insertion/deletion ratio, (D) gap overlap, (E) total length of the alignment, and (F) the proportion of columns correctly recovered. From traditional alignment methods grow with increasing evolutionary distances and more difficult alignments (close-intermediate-distant, white to blue gradient) and also with a denser sampling and increasingly similar sequences (close-2X-4X, white to red gradient).

Fig. 3. Alignment accuracy errors are reduced by a phylogeny-aware algorithm. (A) Inferred branch lengths at different depths in the tree (B) Errors (C) insertion/deletion ratio, (D) gap overlap, (E) total length of the alignment, and (F) the proportion of columns correctly recovered. (G) The phylogeny-aware method PRANK is preferred by, some molecular biologists work-specifically designed for evolutionary analyses inferences. We believe that alignment methods specifically designed for evolutionary analyses are number of (A) insertions and (B) deletions, (C) insertion/deletion ratio, (D) gap overlap, (E) total length of the alignment, and (F) the proportion of columns correctly recovered. From traditional alignment methods grow with increasing evolutionary distances and more difficult alignments (close-intermediate-distant, white to blue gradient) and also with a denser sampling and increasingly similar sequences (close-2X-4X, white to red gradient).

From Loytyjoja and Goldman, Science 2008:
From Loytynoja and Goldman, Science 2008
Observations

• Most alignment methods “over-align” (produce compressed alignments)
• Prank avoids this through its “phylogeny-aware” strategy
• Compression results in
  – Over-estimations of branch lengths
  – Under-estimation of insertions
• Clustal is least accurate, other methods in between
Results so far

• Relative accuracy depends on the alignment criterion – TC and sum-of-pairs scores do not necessarily correlate well.
• Tree accuracy is also not that well correlated with alignment accuracy.
• Different alignment criteria are optimized using different techniques
• Accuracy on AA (amino acid) datasets not the same as accuracy on NT (nucleotide) datasets.
• Dataset properties that impact accuracy:
  – Dataset size
  – Heterogeneity (rate of evolution)
  – Perhaps other things (gap length distribution?) – and note, we have not yet examined fragmentary datasets
• Exact command matters (always check details)
General trends

• Treelength-based optimization currently not as accurate as some standard techniques (e.g., ML on MAFFT alignments)

• Many methods give excellent results on small datasets – Probcons, Probalign, Bali-Phy, etc... but most are not in use because of dataset size limitations

• Large datasets best using PASTA or UPP? (maybe)

• Co-estimation under statistical models might be the way to go, IF...
Research Projects

• Design your own MSA method, or just modify an existing one in some simple way (e.g., different guide tree)
• Test existing MSA methods with respect to different criteria (e.g., extend Prank study to more methods and datasets)
• Develop different MSA criteria that are more appropriate than TC, SPFN, SPFP
• Compare different MSA methods on some biological dataset
• Parallelize some MSA method
• Consider how to combine MSAs on the same input
Treeelength optimization

• POY is the most well-known method for co-estimating alignments and trees using treelength criteria (however – note that the developers of POY say to ignore the alignment and only use the tree).

• The accuracy of the final tree depends on the edit distance formulation – as noted by several studies. Affine gap penalties are more biologically realistic than simple gap penalties.

• We developed BeeTLe (Better Tree Length), a heuristic that is guaranteed to always be as least as accurate as POY for the treelength criterion.
Tree length questions

• Is it better to use affine than simple gap penalties?
• Does POY solve its treelength problem? Is BeeTLe actually better (as promised)?
• How accurate are the alignments?
• How accurate are the trees, compared to
  – MP analyses of good alignments
  – ML analyses of good alignments
whether POY (or any method based upon treelength optimization) is reliable for estimating highly accurate trees or alignments, the more important question is which approaches are likely to produce the most accurate trees and alignments? The study we presented suggests strongly that treelength optimization is unlikely to produce trees or alignments that are as accurate as maximum likelihood on the leading alignment methods; it also showed that SATe-II trees and alignments were even more accurate than maximum likelihood trees on leading alignments. Thus, parsimony-style co-estimation (as in POY and BeeTLe) produced trees and alignments that are inferior to the co-estimation approach in SATe-II. It makes sense, therefore, to discuss SATe-II’s co-estimation technique.

The technique used by SATe-II to co-estimate trees and alignments uses iteration combined with divide-and-conquer; each iteration involves the estimation of a new alignment (produced using divide-and-conquer) and then uses RAxML to produce an ML tree on that new alignment. However, the ML model used in estimating the tree is GTR + Gamma, and so indels are treated in the standard way, which is as missing data – rather than treating...
Maximum Parsimony (MP) on different alignments

Simulated 100-sequence DNA datasets with varying rates of evolution
Results from Liu and Warnow, PLoS ONE 2012
Improved upon MP (MAFFT)'s (using similar statistical tests). Averages and standard error bars are shown; correction for multiple tests, estimation methods (i.e., RAxML for ML analysis and PAUP* for be obtained by a more careful analysis. Obtained by a more careful search through treespace. As a result it is likely that even shorter trees would be some of the neighbors of these trees) for treelength, and returns the shortest tree. Therefore, it is possible that there are currently available methods that might yield more even more accurate trees than those tested in this study. Thus, it is possible that there are currently available methods that might yield more accurate trees than those tested in this study. We also did not explore the particular, likelihood-based methods such as MrBayes [39], Phyml alignments using other phylogeny estimation methods. In particular, likelihood-based methods such as MrBayes [39], Phyml, Probtree, Prank (MP) and a handful of alignment methods (i.e., MAFFT, SATe [40], GARLI [41], FastTree [42,43], and Metapiga2 [44] might detrimental. In other words, although it is important to understand whether co-estimation of trees and alignments is beneficial or estimating trees and alignments from unaligned sequences, and in this study. We report missing branch rates for BeeTLe-Affine in Simulated 100-sequence DNA datasets with varying rates of evolution Results from Liu and Warnow, PLoS ONE 2012
PASTA study

- PASTA (RECOMB 2014 and J. Computational Biology 2014) is the replacement of SATe-1 (Liu et al., Science 2009) and SATe-2 (Liu et al., Systematic Biology 2012)
- Alignment criteria: “Pairs” score and Total Column (TC) score
- Evaluated on simulated and biological datasets (both nucleotide and amino acid)
- Alignment methods compared: “Initial” (an HMM-based technique), Clustal-Omega, MAFFT, and SATe
SATe Family

- **SATe-I (2009):**
  - Up to about 10,000 sequences
  - Good accuracy and reasonable speed
  - “Center-tree” decomposition

- **SATe-II (2012)**
  - Up to about 50,000 sequences
  - Improved accuracy and speed
  - Centroid-edge recursive decomposition

- **PASTA (2014)**
  - Up to 1,000,000 sequences
  - Improved accuracy and speed
  - Combines centroid-edge decomposition with transitivity merge
FIG. 1. Algorithmic design of PASTA. The first six boxes show the steps involved in one iteration of PASTA. The last two boxes show the meaning of transitivity for homologies defined by a column of an MSA, and how the concept of transitivity can be used to merge two compatible and overlapping alignments. MSA, multiple sequence alignment.
SATé-I
vs. SATé-II

SATé-II
• Faster and more accurate than SATé-I
• Longer analyses or use of ML to select tree/alignment pair slightly better results
PASTA variants – impact of alignment subset size

Table 4. Impact of Alignment Subset Size

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Subset size</th>
<th>Tree error</th>
<th>Alignment accuracy</th>
<th>Running times</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>FN</td>
<td>Pairs score</td>
<td>TC</td>
</tr>
<tr>
<td>RNASim 10K</td>
<td>200</td>
<td>10.7%</td>
<td>88.8%</td>
<td>145</td>
</tr>
<tr>
<td>RNASim 10K</td>
<td>100</td>
<td><strong>10.4%</strong></td>
<td>87.4%</td>
<td>185</td>
</tr>
<tr>
<td>RNASim 10K</td>
<td>50</td>
<td>10.7%</td>
<td>88.6%</td>
<td><strong>210</strong></td>
</tr>
<tr>
<td>16S.T</td>
<td>200</td>
<td>8.2%</td>
<td><strong>82.7%</strong></td>
<td>121</td>
</tr>
<tr>
<td>16S.T</td>
<td>100</td>
<td>8.1%</td>
<td>82.0%</td>
<td>125</td>
</tr>
<tr>
<td>16S.T</td>
<td>50</td>
<td><strong>7.9%</strong></td>
<td>79.0%</td>
<td><strong>129</strong></td>
</tr>
</tbody>
</table>

We report tree error and alignment accuracy on one replicate of the 10K RNASim dataset and also on the 16S.T dataset, using three iterations of PASTA in which we explore the impact of changing the subset size from 200 (the default) to 100 and 50; all other algorithmic parameters use default values. Boldface indicates the best performance on the data.
Table 2 shows alignment accuracy on the AA datasets. Due to dataset sizes, Muscle and SATe-II failed to complete on two of the HomFam datasets, so we separate out the results for these two datasets from the remaining 17 HomFam datasets. PASTA had the best pairs score or was tied for the best pairs score for both HomFam and AA-10 datasets. Mafft had the best TC score for HomFam(17), but PASTA was very close. For HomFam(2), PASTA had the best TC score and Mafft was a close second. On AA-10 datasets, SATe-II had the best TC score and was closely trailed by Mafft and PASTA.

**Comparison to SATe-II on 50,000-taxon dataset.**

SATe-II could not finish even one iteration on the RNASim with 50,000 sequences running for 24 hr and given 12 CPUs on TACC. However, we were able to

---

**Comparison of PASTA to SATe-II and other methods on nucleotide datasets, with respect to tree error.** Figure from Mirarab et al., J. Computational Biology 2014