Discussion of 6 papers
Things to consider for studies comparing methods

• Pay attention to:
  – Datasets
  – Criteria
  – Competing methods
  – Whether differences are substantial and interesting

• Also consider how reproducible the study is.

See Appendix C in the textbook.
Datasets

• What datasets were used? How large? What properties do they have?
• Do the datasets have properties similar to biological datasets that would be commonly encountered?
• Do the datasets favor some particular methods?
• Are the datasets “easy” or “hard”?
• Can you find the datasets?
• Anything particularly noteworthy?
Criteria

• What criteria did the authors use? Is it a standard criterion?
• What software used to calculate scores?
• Are there other criteria that would be better?
• Do the criteria used favor some method?
Competing methods

• New methods should always be compared to the best available methods at that time. So check:
  – What competing methods did they use? What commands and version numbers?
  – Are there other methods they should have examined?
Evaluating differences

• Showing that method X is better than method Y is not enough.
  – How big are the differences? Enough to make a difference in some biological discovery?
  – How often is it better, and under what circumstances?
  – Are the differences statistically significant?
Reproducibility

• A study can’t be reproduced or even interpreted unless you really can answer these questions positively:

1. Do you know the method version numbers and commands?
2. Can you obtain the datasets analyzed in the study, or regenerate them (and would this be easy to do)?
3. Do you have access to the codes used to evaluate the results, or can you produce the same evaluation yourself?
Six papers you read

• "Who watches the watchmen?", by Iantorno et al. (2014), DOI 10.1007/978-1-62703-646-7_4
• "Phylogeny-Aware Gap Placement Prevents Errors in Sequence Alignment and Evolutionary Analysis", by Loytynoja and Goldman (2008), DOI: 10.1126/science.1158395
• "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega" by Sievers et al. (2011), DOI: 10.1038/msb.2011.75
• "BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny", by Suchard and Redelings (2006), DOI: 10.1093/bioinformatics/btl175
• "Split-Inducing Indels in Phylogenomic Analysis by Donath and Stadler (2011)
“Who watches the watchmen?”

• What is this about?
• Any performance study?
• What did you think?
“Phylogeny-aware gap placement”

• What is this about?
• Any performance study?
• What did you think?
“BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny”

• What is this about?
• Any performance study?
• What did you think?
“M-Coffee...”

• What is this about?
• Any performance study?
• What did you think?
“Split-inducing Indels...”

• What is this about?
• Any performance study?
• What did you think?
“Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal-Omega”

• What is this about?
• Any performance study?
• What did you think?
• "Who watches the watchmen?", by Iantorno et al. (2014), DOI 10.1007/978-1-62703-646-7_4
• "Phylogeny-Aware Gap Placement Prevents Errors in Sequence Alignment and Evolutionary Analysis", by Loytynoja and Goldman (2008), DOI: 10.1126/science.1158395
• "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega" by Sievers et al. (2011), DOI: 10.1038/msb.2011.75
• "BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny", by Suchard and Redelings (2006), DOI: 10.1093/bioinformatics/btl175
• "Split-Inducing Indels in Phylogenomic Analysis” by Donath and Stadler (2011)
Clustal-Omega
Clustal-Omega

• Discussion of “Fast scalable generation of high-quality protein multiple sequence alignments using Clustal-Omega” by Sievers et al., published in Molecular Systems Biology 7:539 (2011)
  – Design of Clustal-Omega
  – Performance on data
  – Main innovation: guide tree construction
  – Reviews in “transparent process”
  – Comparison to other studies (time permitting)
Clustal-Omega study

- Clustal-Omega 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
Clustal-Omega

Given unaligned sequences:

• Build a guide tree
• Perform progressive alignment on the guide tree, using “consistency” to inform the alignment mergers

If desired, you can:

• Iterate (construct profile HMM based on alignment, and then re-align)
• Use external profile alignment (EPA)
Evaluation

• Default Clustal-Omega (no iteration, no EPA) compared to standard methods on protein benchmarks

• Variants of Clustal-Omega compared to default Clustal-Omega

Criteria:

• TC (column score) on “core” columns

• Running time
Datasets

• BAliBASE ("small")
• PreFab (50 sequences in each set, but evaluated based on pairwise alignments)
• HomFam (much larger!)
TC scores on BAliBASE sequences

Table 1 BAliBASE results

<table>
<thead>
<tr>
<th>Aligner</th>
<th>BB11 (38 families)</th>
<th>BB12 (44 families)</th>
<th>BB2 (41 families)</th>
<th>BB3 (30 families)</th>
<th>BB4 (49 families)</th>
<th>BB5 (16 families)</th>
<th>Tot time (s)</th>
<th>Consistency</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSAprobs</td>
<td>0.607</td>
<td>0.441</td>
<td>0.865</td>
<td>0.464</td>
<td>0.607</td>
<td>0.622</td>
<td>0.608</td>
<td>12382.00</td>
</tr>
<tr>
<td>Probalign</td>
<td>0.589</td>
<td>0.453</td>
<td>0.862</td>
<td>0.439</td>
<td>0.566</td>
<td>0.603</td>
<td>0.549</td>
<td>10095.20</td>
</tr>
<tr>
<td>MAFFT (auto)</td>
<td>0.588</td>
<td>0.439</td>
<td>0.831</td>
<td>0.450</td>
<td>0.581</td>
<td>0.605</td>
<td>0.591</td>
<td>1475.40</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Probcons</td>
<td>0.558</td>
<td>0.417</td>
<td>0.855</td>
<td>0.406</td>
<td>0.544</td>
<td>0.532</td>
<td>0.573</td>
<td>13086.30</td>
</tr>
<tr>
<td>Clustal Ω</td>
<td>0.554</td>
<td>0.358</td>
<td>0.789</td>
<td>0.450</td>
<td>0.575</td>
<td>0.579</td>
<td>0.533</td>
<td>539.91</td>
</tr>
<tr>
<td>T-Coffee</td>
<td>0.551</td>
<td>0.410</td>
<td>0.848</td>
<td>0.402</td>
<td>0.491</td>
<td>0.545</td>
<td>0.587</td>
<td>81041.50</td>
</tr>
<tr>
<td>Kalign</td>
<td>0.501</td>
<td>0.365</td>
<td>0.790</td>
<td>0.360</td>
<td>0.476</td>
<td>0.504</td>
<td>0.435</td>
<td>21.88</td>
</tr>
<tr>
<td>MUSCLE</td>
<td>0.475</td>
<td>0.318</td>
<td>0.804</td>
<td>0.350</td>
<td>0.409</td>
<td>0.450</td>
<td>0.460</td>
<td>789.57</td>
</tr>
<tr>
<td>MAFFT (default)</td>
<td>0.458</td>
<td>0.258</td>
<td>0.749</td>
<td>0.316</td>
<td>0.425</td>
<td>0.480</td>
<td>0.496</td>
<td>68.24</td>
</tr>
<tr>
<td>FSA</td>
<td>0.419</td>
<td>0.270</td>
<td>0.818</td>
<td>0.187</td>
<td>0.259</td>
<td>0.474</td>
<td>0.398</td>
<td>53648.10</td>
</tr>
<tr>
<td>Dialign</td>
<td>0.415</td>
<td>0.265</td>
<td>0.696</td>
<td>0.292</td>
<td>0.312</td>
<td>0.441</td>
<td>0.425</td>
<td>3977.44</td>
</tr>
<tr>
<td>PRANK</td>
<td>0.376</td>
<td>0.223</td>
<td>0.680</td>
<td>0.257</td>
<td>0.321</td>
<td>0.360</td>
<td>0.356</td>
<td>128355.00</td>
</tr>
<tr>
<td>ClustalW</td>
<td>0.374</td>
<td>0.227</td>
<td>0.712</td>
<td>0.220</td>
<td>0.272</td>
<td>0.396</td>
<td>0.308</td>
<td>766.47</td>
</tr>
</tbody>
</table>

The figures are total column scores produced using bali score on core columns only. The average score over all families is given in the second column. The results for BAliBASE subgroupings are in columns 3–8. The total run time for all 218 families is given in the second last column. The last column indicates whether the method is consistency based.
TC scores on PreFab sequences

Table II Prefab results

<table>
<thead>
<tr>
<th>Aligner</th>
<th>0 &lt; %ID ≤ 100 (1682 families)</th>
<th>0 ≤ %ID ≤ 20 (912 families)</th>
<th>20 ≤ %ID ≤ 40 (563 families)</th>
<th>40 ≤ %ID ≤ 70 (117 families)</th>
<th>70 ≤ %ID ≤ 100 (90 families)</th>
<th>Total time (s) (1682 families)</th>
<th>Consistency</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSAprobs</td>
<td>0.737</td>
<td>0.756</td>
<td>0.769</td>
<td>0.777</td>
<td>0.785</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>0.971</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>0.717</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>Clustal Ω</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>225.06</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.

Note that best performing method depends on the “%ID” (measure of similarity)

From Seivers et al., Molecular Systems Biology 2011
TC scores on HomFam sequences

Table III  HomFam benchmarking results

<table>
<thead>
<tr>
<th>Aligner</th>
<th>93 ≤ N ≤ 2957 (41 families)</th>
<th>3127 ≤ N ≤ 9105 (33 families)</th>
<th>10 099 ≤ N ≤ 50157 (18 families)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TC/t (s)</td>
<td>TC/t (s)</td>
<td>TC/t (s)</td>
</tr>
<tr>
<td>Clustal Ω</td>
<td>0.708/2114.0</td>
<td>0.639/11 719.5</td>
<td>0.464/27 328.9</td>
</tr>
<tr>
<td>Kalign</td>
<td>0.569/324.9</td>
<td>0.563/6752.0</td>
<td>0.420/286 711.0</td>
</tr>
<tr>
<td>MAFFT default</td>
<td>0.550/238.9</td>
<td>0.462/3115.4</td>
<td>/ /</td>
</tr>
<tr>
<td>MAFFT –parttree</td>
<td>/ /</td>
<td>/ /</td>
<td>0.253/6119.4</td>
</tr>
<tr>
<td>MUSCLE default</td>
<td>0.533/104 587.0</td>
<td>/ /</td>
<td>/ /</td>
</tr>
<tr>
<td>MUSCLE –maxiters 2</td>
<td>/ /</td>
<td>0.416/8239.2</td>
<td>0.216/110 292.0</td>
</tr>
</tbody>
</table>

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.

Note the reduced set of methods due to dataset size.

From Seivers et al., Molecular Systems Biology 2011
Observations (so far)

- 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 Probalign, MAFFT v6.857, and MSAprobs.
- On large datasets, Clustal-Omega more accurate than other methods that it was compared to.
Clustal-Omega variants

• Iterating using profile HMMs
• Using EPA (profiles based on external structural alignments)
EPA for HomFam and BAiLBASE.

A  ClustalΩ HomFam

B  ClustalΩ BAiLBASE

© as stated in the article, figure or figure legend

Fabian Sievers et al. Mol Syst Biol 2011;7:539
Iteration of HomFam alignments.

Fabric Sievers et al. Mol Syst Biol 2011;7:539
Alignment time for Clustal Omega (red), MAFFT (blue), MUSCLE (green) and Kalign (purple) against the number of sequences of HomFam test sets.

Fabian Sievers et al. Mol Syst Biol 2011;7:539
Additional Observations

- Iterating using profile HMMs improves TC scores for Clustal-Omega.
- Using EPA (profiles based on external structural alignments) improves TC scores for Clustal-Omega.
- MAFFT-parttree is faster than Clustal-Omega.
- The key to Clustal-Omega’s speed is its guide tree.
Some concerns

• Other criteria (1-SPFN, 1-SPFP, impact on tree error, etc.)
• Other datasets
• Newer methods (or updated methods)
• Restriction to methods that complete quickly
TC scores on HomFam sequences

**Table III** HomFam benchmarking results

<table>
<thead>
<tr>
<th>Aligner</th>
<th>93 ≤ N ≤ 2957 (41 families)</th>
<th>3127 ≤ N ≤ 9105 (33 families)</th>
<th>10 099 ≤ N ≤ 50157 (18 families)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clustal Ω</td>
<td>0.708/2114.0</td>
<td>0.639/11 719.5</td>
<td>0.464/27 328.9</td>
</tr>
<tr>
<td>Kalign</td>
<td>0.569/324.9</td>
<td>0.563/6752.0</td>
<td>0.420/286 711.0</td>
</tr>
<tr>
<td>MAFFT default</td>
<td>0.550/238.9</td>
<td>0.462/3115.4</td>
<td>–/–</td>
</tr>
<tr>
<td>MAFFT –parttree</td>
<td>–/–</td>
<td>–/–</td>
<td>–/–</td>
</tr>
<tr>
<td>MUSCLE default</td>
<td>0.533/104 587.0</td>
<td>–/–</td>
<td>0.253/6119.4</td>
</tr>
<tr>
<td>MUSCLE –maxiters 2</td>
<td>–/–</td>
<td>0.416/8239.2</td>
<td>0.216/110 292.0</td>
</tr>
</tbody>
</table>

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.

Note the reduced set of methods due to dataset size.

From Seivers et al., Molecular Systems Biology 2011
Methods they didn’t explore

• PASTA and UPP: designed explicitly for large datasets. (They tested SATe-1, the precursor to SATe-II and PASTA, but not on the large datasets because it could not be run efficiently.)

• Improved versions of MAFFT since their study

• And who knows what else...
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.

<table>
<thead>
<tr>
<th>Method</th>
<th>AA-10</th>
<th>HomFam(17)</th>
<th>HomFam(2)</th>
<th>AA-10</th>
<th>HomFam(17)</th>
<th>HomFam(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clustal-O</td>
<td>78</td>
<td>88</td>
<td>29</td>
<td>0.76</td>
<td>0.72</td>
<td>0.71</td>
</tr>
<tr>
<td>Muscle</td>
<td>48</td>
<td>51</td>
<td>X</td>
<td>0.70</td>
<td>0.52</td>
<td>X</td>
</tr>
<tr>
<td>Mafft</td>
<td>81</td>
<td>103</td>
<td>32</td>
<td>0.76</td>
<td>0.75</td>
<td>0.79</td>
</tr>
<tr>
<td>Initial</td>
<td>54</td>
<td>95</td>
<td>16</td>
<td>0.75</td>
<td>0.71</td>
<td>0.81</td>
</tr>
<tr>
<td>SATé-II</td>
<td>83</td>
<td>73</td>
<td>X</td>
<td>0.75</td>
<td>0.64</td>
<td>X</td>
</tr>
<tr>
<td>PASTA</td>
<td>80</td>
<td>102</td>
<td>36</td>
<td>0.76</td>
<td>0.78</td>
<td>0.83</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.

Mafft version 7.143b run in default mode (even on HomFam (17) datasets).

Comparison of PASTA to SATe-II and other alignments on AA datasets.

From Mirarab et al., J. Computational Biology 2014
MAFFT vs. Clustal-Omega

• In the Clustal-Omega paper, Clustal-Omega had better TC scores than MAFFT.

• In the PASTA paper, MAFFT had better TC scores than Clustal-Omega.
Why the difference?

• Two versions of MAFFFT (Clustal-Omega paper used version 6.857, PASTA paper used version 7.143b)
• Two different commands possibly (MAFFT-parttree vs. MAFFFT-default)

Lesson: Version number and command impacts absolute and relative performance!
Reviewer 1

• “The overall message appears to be that ClustalOmega appears to be the fastest and most accurate algorithm for very large alignments, while it is still substantially outperformed by other methods for smaller alignments.”

• “Looking at the results presented here (again, from the standpoint of a "consumer"), I would not personally expect to make much use of this algorithm and will probably continue to use MAFFT or MUSCLE, mainly since they're still performing similarly well or better for smaller alignments, which are the main use for myself. I expect this to be true for the majority of users. However, this may change in the future and especially for those who work on major sequencing / data repository centers (such as the EBI), as sequencing of more species ramps up.”
The main novelty of this program compared to previous versions of Clustal is its ability to deal with huge data sets. This is mainly achieved by using an efficient clustering algorithm to efficiently calculate a guide tree for progressive alignment... constructing the guide tree for progressive alignment is the bottleneck of most MSA programs: calculating a tree from the input sequences with traditional methods takes (at least) \(O(n^2)\) time while progressively aligning the sequences according to this tree can be done in linear time."
Comments about EPA

• Reviewer 1: “It also offers the Profile Alignment use-case, which is likely to be more important in the future”

• Reviewer 2: “A second interesting novelty of Clustal Omega is the option to use previously calculated "External Profile Alignments" (EPA) to improve the alignment procedure. This is done using an HMM approach developed by one of the co-authors. Since for many protein families, there are now profile HMMs available, it is a very sensible idea to use this information for improved MSA, rather than relying on sequence similarity alone, as do traditional alignment methods.”
Overall

- Main features of Clustal-Omega are the fast guide tree calculation, the use of iteration, and EPA. (Both EPA and iteration use profile HMMs.)
- But improvements in accuracy compared to other methods only seems to be apparent on large datasets when restricting the set of methods to those that complete quickly. Newer methods (and even new versions of old methods like MAFFT) seem to outperform Clustal-Omega for accuracy, even if they are not as fast.
- Accuracy was only assessed using TC score, which is not necessarily the whole story.
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.
From Loytyjoja and Goldman, Science 2008:
Alignment accuracy errors are reduced by a phylogeny-aware algorithm. From 1 to 4 for the intermediate sets indicate that alignment errors lead to overestimated 28 March 2008; accepted 15 May 2008 20 JUNE 2008 SCIENCE VOL 320 280 20 JUNE 2008

Fig. 3.
The phylogeny-aware method PRANK sequence sampling and increasingly similar sequences (close-2X-4X, white to red gradient). 

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
Research questions

• Clustal-Omega’s main advantage seems to be its fast guide tree. But the guide tree has a large impact on accuracy – can we improve the accuracy of Clustal-Omega by giving it a better guide tree?

• Iteration using profile HMMs helps accuracy; can we use this technique in other methods?

• EPAs are useful. What are the best ways to use external structurally-defined alignments?