Scalability of Alignment Quality with Number of Sequences

Fabian Sievers

Conway Institute of Biomolecular & Biomedical Research, University College Dublin, Ireland

Keywords

Multiple Sequence Alignment, MSA, Alignment Quality, Total Column Score, Scalability, Large Alignments, Guide-Trees, Progressive Alignment, Iteration, Homology, Clustal-Omega

Abstract

Multiple Sequence Alignments (MSAs) of >100,000 sequences are getting commonplace. At present, there are no systematic analyses concerning the scalability of the alignment quality as the number of aligned sequences is increased. We bench-marked many widely used MSA packages using protein families with known structures. We found that the accuracy of alignments decreases as more sequences are added indiscriminately; this is true for all packages and large numbers of sequences. For small numbers of carefully selected sequences a modest improvement is possible. The reason for this deterioration is mostly due to 'attrition' during the profile alignment stage rather than problems in guide-tree construction. This effect can be attenuated through iteration or external profile alignment. This suggests that the availability of high quality curated alignments will have to complement algorithmic and/or software developments in the long-term.

Materials

The following MSA programs were benchmarked:

1. Clustal Omega, v1.0.3
2. ClustalW2, v2.1
3. DIALIGN 2.2.1
4. FSA 1.15.5
5. Kalign 2.04
6. MAFFT 6.857
7. MSAProbs 0.9.4
8. MUMMALS 1.01
9. MUSCLE v3.8.31
10. Opal v2.0.0
11. Pagan v.0.38
12. POA V2 v1.0.0
13. PRANK v.100802
14. Probalign v1.4
15. PROBCONS v1.12
16. PSAlign
17. SATé v1.4.0
18. T-Coffee v8.99

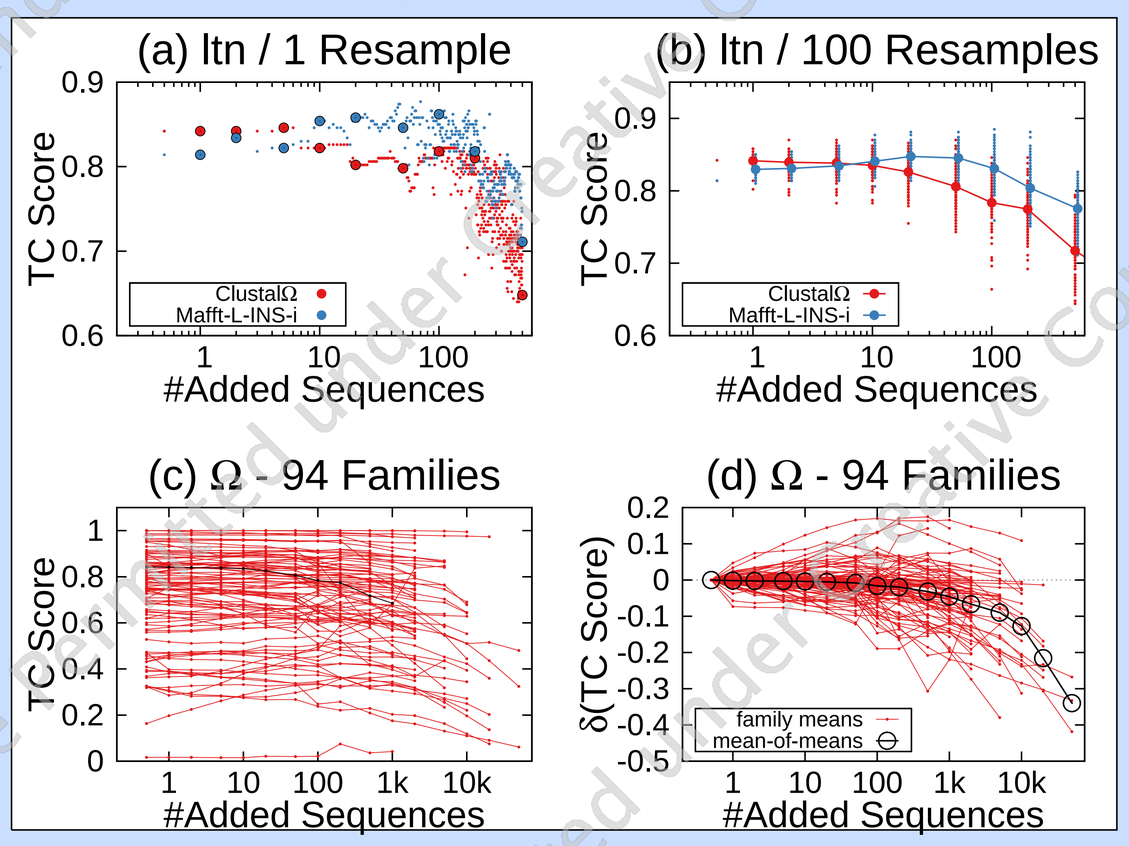
Benchmark data-sets were constructed by successivly adding homologous Pfam sequences to Homstrad reference sequences.

- Compiled 94 families, where Homstrad and Pfam entry had one-to-one match
- Number of reference sequences 5 to 41, alignment lengths 39 to 938
- Between 88 and 93,675 non-reference sequences (three families with >50,000 sequences)

Methods

• Setup:

- Incrementally add 0,1,2,5,10,... non-reference (Pfam) sequences to (Homstrad) references
- Use Total-Column (TC) Score to measure accuracy of alignment (sum-of-pairs similar)
- Measure TC Score for all positions; results for Core Columns are similar
- Express all TC scores wrt base score, where only reference sequences are aligned
- Repeat with differently randomised sequences, average
- Constructing biggest possible (default) tree and successively populating leaves focuses on profile-profile alignment accuracy
- Constructing guide-tree from scratch, pruning away non-reference sequences and re-aligning reference sequences focuses on guide-tree construction



(a) TC scores for example family when non-reference sequences are added one-by-one. (b) TC scores for 100 different re-samples of example family, mean with bullets. (c) Mean TC scores for all 94 HomFam families. (d) Mean TC scores shifted by base alignment scores, mean-of-means with black circles.

• Iteration:

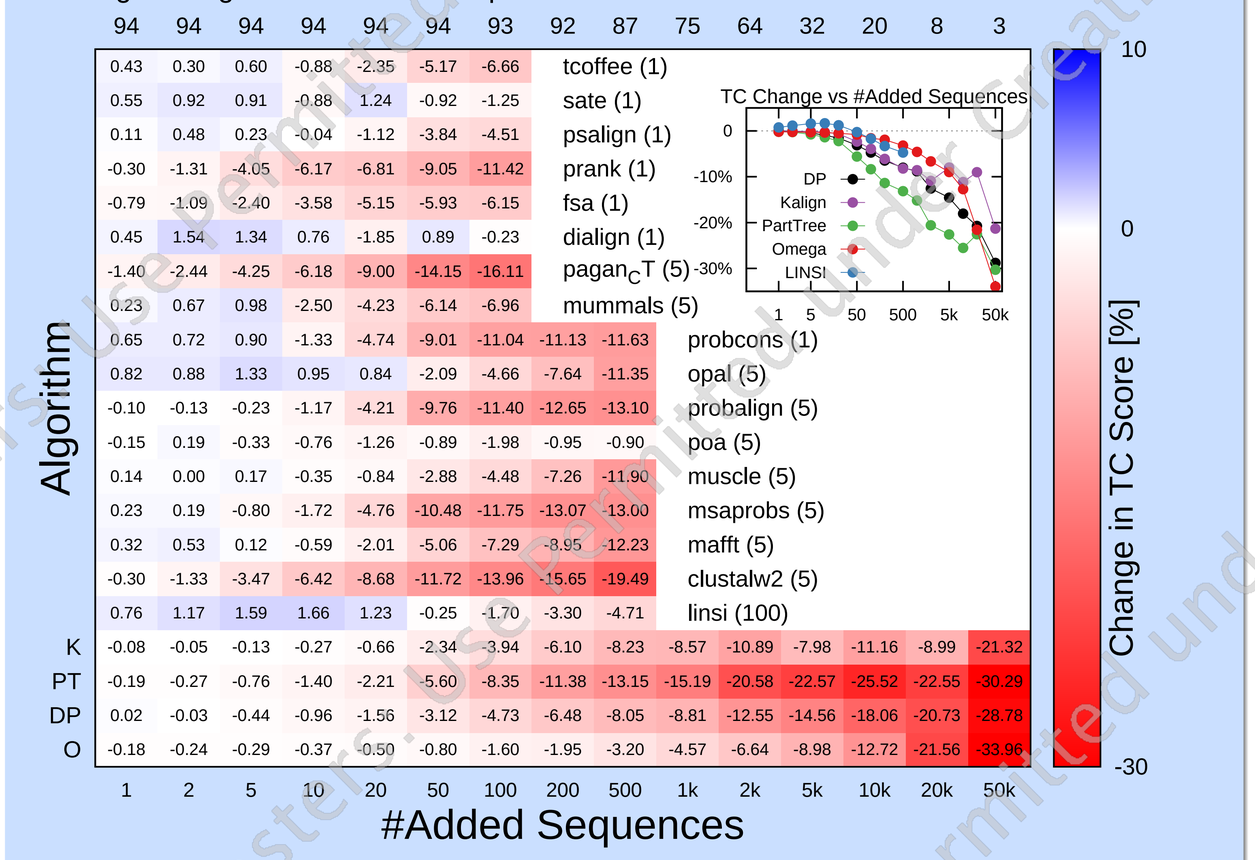
- Re-construct guide-tree, based on preliminary full alignment (not just pair-wise alignments)
- Use Hidden Markov Model (HMM) of preliminary alignment as guide (Clustal-Omega only)
- Bisect preliminary alignment and realign sequences from first group onto second group (e.g., Mafft & Muscle)

• Homology Extension & External Profile Alignment:

- Group homologous sequences into (a) very similar, (b) somewhat similar, (c) medium and (d) dissimilar wrt minimum k-tuple distance to any of the references
- Only add sequences from one 'band' of homologous sequences
- External Profile Alignment (EPA) uses external HMM to guide alignment
- EPA can come from small, locally produced alignments or large data-base like Pfam

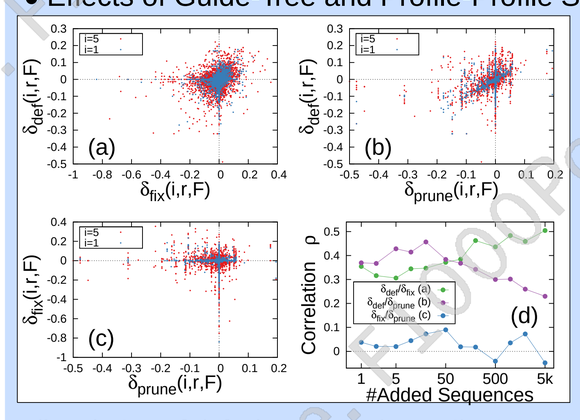
Results

• Change in Alignment Score as Sequences Added

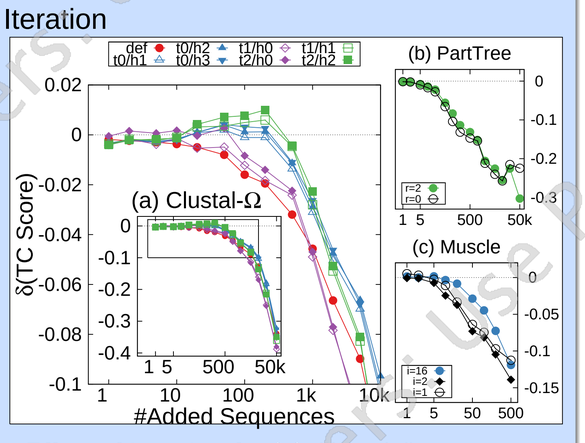


Change in HomFam alignment score wrt base alignment as non-reference sequences added. #Added sequences along bottom x-axis, #families along top x-axis. Alignment algorithm along the y-axis, K = Kalign, PT = MAFFT-PartTree, DP = DP-PartTree, O = Clustal Omega. #Number of re-samples R in parentheses, R=100 for K, PT, DP and O. Improvement blue, deterioration red. Top right hand inset: O (red), K (purple), PT (green) DP (black), MAFFT L-INS-i (blue)

• Effects of Guide-Tree and Profile-Profile Stage & Iteration

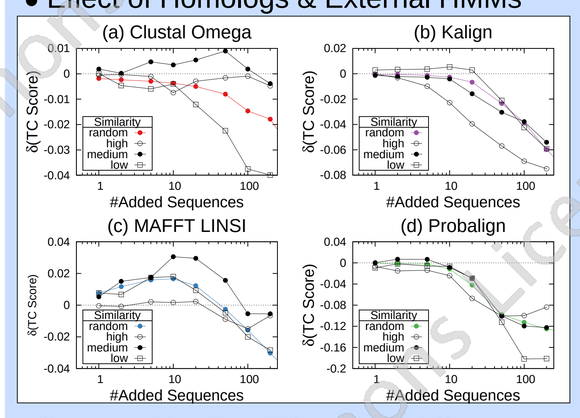


Correlation of default score and contributions from tree building and profile alignment. (a) default/fixed tree, (b) default/pruned tree, (c) fixed/pruned tree, (d) correlation with number of added sequences

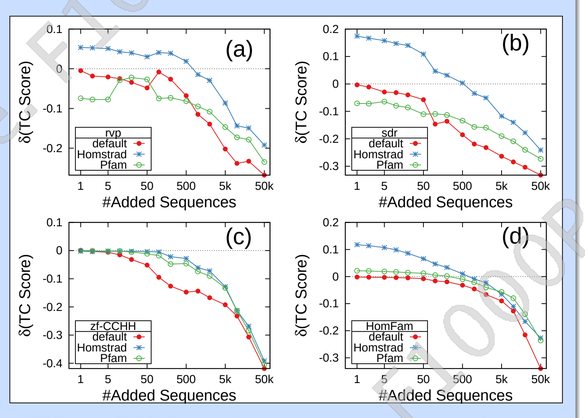


Effect of iteration for (a) Clustal-Omega, default(red), guide-tree iteration (purple), HMM iteration (blue), combined (green); (b) Mafft PartTree, default (gree), no iteration (black); (c) Muscle, default (blue), fewer iterations (black)

• Effect of Homologs & External HMMs



Change in HomFam TC score for different algorithms and different sampling schemes. (a) Clustal Omega, (b) Kalign, (c) MAFFT L-INS-i, (d) Probalign. Random sampling with bullets and thicker lines, sampling of sequences of high similarity with circles, of low similarity with crosses and of in-between similarity with diamonds and boxes



Effect of EPA on alignment accuracies of most numerous families: (a) rvp, (b) sdr, (c) zf-CCHH, (d) average of all HomFam. Default (as before – red), over-fitted (HMM derived from reference alignment – blue), Pfam (HMM downloaded from Pfam – green).

Discussion

- Alignment quality goes down for all aligners as more sequences are added
- Effects of guide-tree construction and profile-profile stage decoupled
- Guide-tree phase more important for fewer, profile-profile phase for more sequences
- Iteration, homology extension and EPA can delay onset of quality decay
- Very similar and very dissimilar sequences do not improve score much in homology extension, best results for intermediate similarities
- For very large numbers of sequences none of the above can rescue quality
- EPA appears to have highest potential, so construction of curated reference alignments must accompany algorithm development

References

- Sievers F, Dineen D, Wilm A and Higgins DG, Making automated multiple alignments of very large numbers of protein sequences, Bioinformatics, Volume 29, Issue 8, Pp. 989-995.
- Sievers F, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. 2011;7:539.