Review: Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega

CS466, Spring 2017
Aayushee Jain

Background:

- Clustal-W : Progressive alignment with guide trees
  - Complexity $O(N^2)$
  - Limit of a few thousands
  - Mistakes made initially can not be corrected later.

- T-Coffee : Consistency principle
  - Increased accuracy by 5-10%
  - Limit of a few hundred sequences

- Clustal-Omega is not only more accurate but can align any number of sequences.
- Also supports reusability.

Clustal-Omega:

- Clustal Omega is an iterative progressive alignment approach that uses HMM
- Uses mBed to generate guide tree. This method is faster than the tradition approach and has time complexity O(NlogN).
- Uses Hhalign to increase accuracy.
- Performs iteration on both guide tree and HMM alignments

Results: BAliBASE

Table I BAliBASE results

Aligner	Av score (218 families)	BB11 (38 families)	BB12 (44 families)	BB2 (41 families)	BB3 (30 families)	BB4 (49 families)	BB5 (16 families)	Tot time (s)	Consistency
MSAprobs	0.607	0.441	0.865	0.464	0.607	0.622	0.608	12 382.00	Yes
Probalign	0.589	0.453	0.862	0.439	0.566	0.603	0.549	10 095.20	Yes
MAFFT (auto)	0.588	0.439	0.831	0.450	0.581	0.605	0.591	1475.40	Mostly (203/218)
Probcons	0.558	0.417	0.855	0.406	0.544	0.532	0.573	13 086.30	Yes
Clustal Ω	0.554	0.358	0.789	0.450	0.575	0.579	0.533	539.91	No
T-Coffee	0.551	0.410	0.848	0.402	0.491	0.545	0.587	81 041.50	Yes
Kalign	0.501	0.365	0.790	0.360	0.476	0.504	0.435	21.88	No
MUSCLE	0.475	0.318	0.804	0.350	0.409	0.450	0.460	789.57	No
MAFFT (default)	0.458	0.258	0.749	0.316	0.425	0.480	0.496	68.24	No
FSA	0.419	0.270	0.818	0.187	0.259	0.474	0.398	53 648.10	No
Dialign	0.415	0.265	0.696	0.292	0.312	0.441	0.425	3977.44	No
PRANK	0.376	0.223	0.680	0.257	0.321	0.360	0.356	128 355.00	No
ClustalW	0.374	0.227	0.712	0.220	0.272	0.396	0.308	766.47	No

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.

Sievers F, Wilm A, Dineen D, et al. 2011

The BAliBASE benchmark is comprised of 218 families, grouped into

- 38+44 families exhibiting variability and length (BB11+BB12),
- 41 families containing orphan sequences (BB2),
- 30 families with sub-families (BB3),
- 49 families with extensions (BB4) and
- 16 families with insertions (BB5).

Results: Prefab

Aligner	0<%ID≤100 (1682 families)	0≤%ID≤20 (912 families)	20≤%ID≤40 (563 families)	40≤%ID≤70 (117 families)	70≤%ID≤100 (90 families)	Total time (s) (1682 families)	Consistency
MSAprobs	0.737	0.591	0.889	0.965	0.971	51 286.00	Yes
MAFFT (auto)	0.721	0.569	0.876	0.961	0.979	4544.45	Yes
Probalign	0.719	0.563	0.881	0.961	0.977	35 117.30	Yes
Probcons	0.717	0.562	0.876	0.955	0.972	46 908.30	Yes
T-Coffee	0.710	0.558	0.865	0.950	0.972	175 789.00	Yes
Clustal Ω	0.700	0.535	0.866	0.967	0.980	1698.06	No
MUSCLE	0.677	0.507	0.850	0.946	0.976	2068.56	No
MAFFT	0.677	0.513	0.836	0.961	0.979	225.56	No
Kalign	0.649	0.474	0.817	0.957	0.979	80.81	No
ClustalW2	0.617	0.430	0.797	0.933	0.975	3433.53	No
Dialign	0.595	0.398	0.783	0.940	0.974	18 909.70	No
PRANK	0.586	0.390	0.767	0.951	0.978	351 498.00	No
FSA	0.534	0.277	0.791	0.965	0.976	229 391.00	No

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.

- The alignments are scored by comparing to reference alignments of two sequences.
- The entire benchmark set can be grouped according to the pairwise identities of the reference sequences.

Sievers F, Wilm A, Dineen D, et al. 2011

Results: HomFab

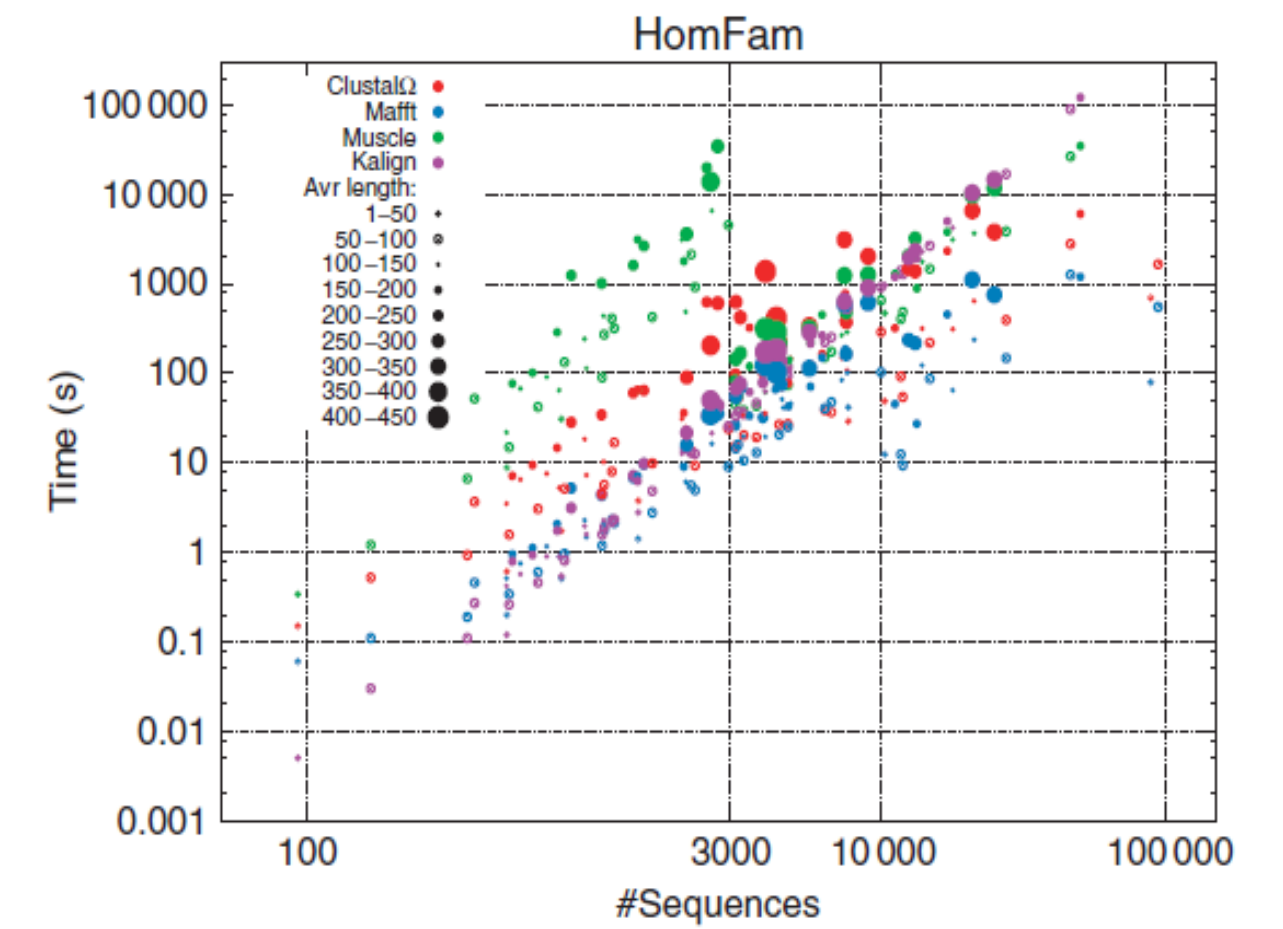
HomFam benchmarking results

	$93 \leqslant N \leqslant 2957$ (41 families)	$3127 \leqslant N \leqslant 9105$ (33 families)	$10\,099 \leqslant N \leqslant 50\,157$ (18 families)
Aligner	TC/t (s)	TC/t (s)	TC/t (s)
Clustal Ω	0.708/2114.0	0.639/11 719.5	0.464/27 328.9
Kalign	0.569/324.9	0.563/6752.0	0.420/286 711.0
MAFFT default	0.550/238.9	0.462/3115.4	–/–
MAFFT –parttree	–/–	–/–	0.253/6119.4
MUSCLE default	0.533/104 587.0	–/–	–/–
MUSCLE –maxiters 2	–/–	0.416/8239.2	0.216/110 292.0

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.

Sievers F, Wilm A, Dineen D, et al. 2011

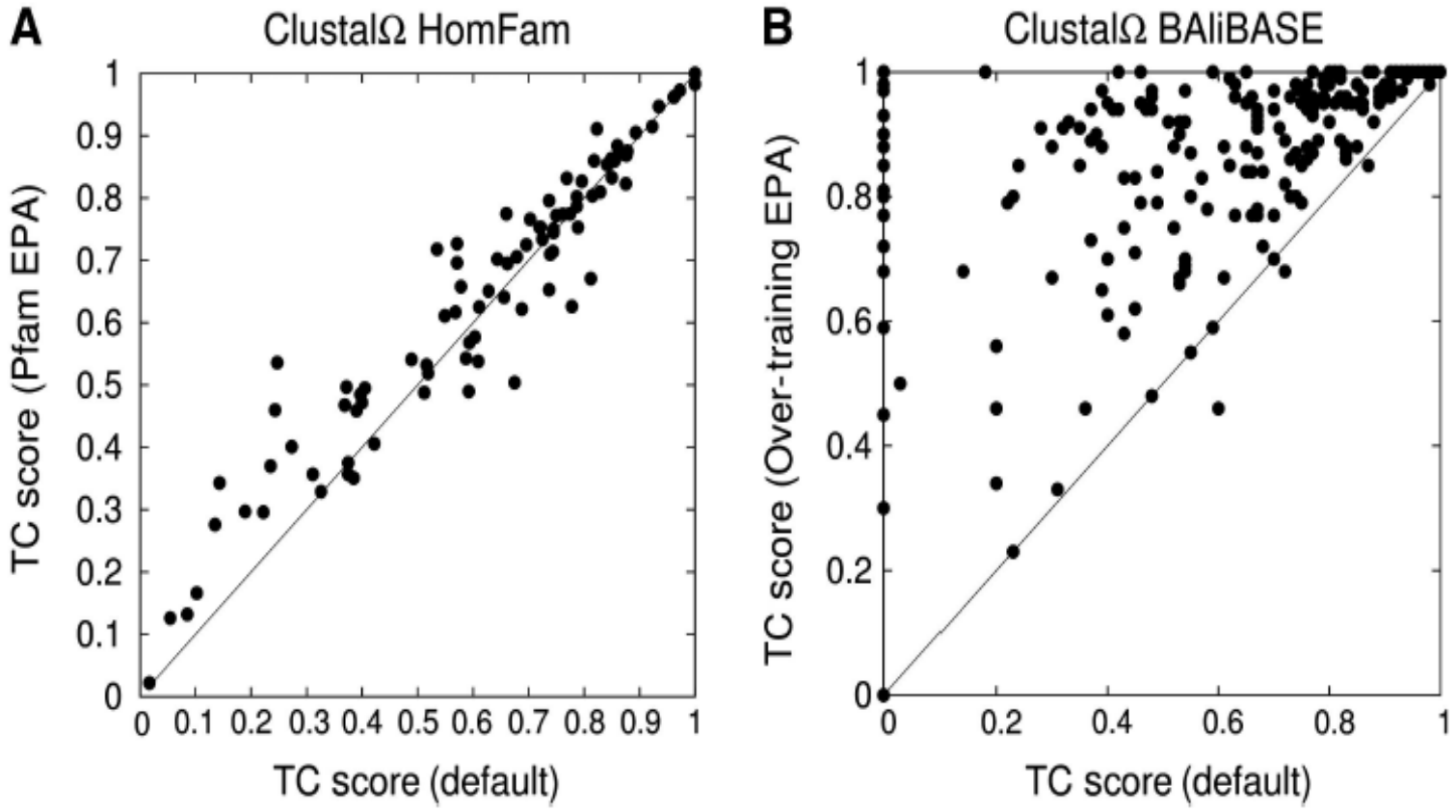
Discussion: Comparison of Alignment time



- Clustal Omega (red), MAFFT (blue), MUSCLE (green) and Kalign (purple)
- Both axes have logarithmic scales.
- Clustal Omega and Kalign were run with default flags over the entire range.
- MUSCLE was run with –maxiters 2 for *N*>3000 sequences.
- MAFFT was run with --parttree for *N*>10 000 sequences.

Sievers F, Wilm A, Dineen D, et al. 2011

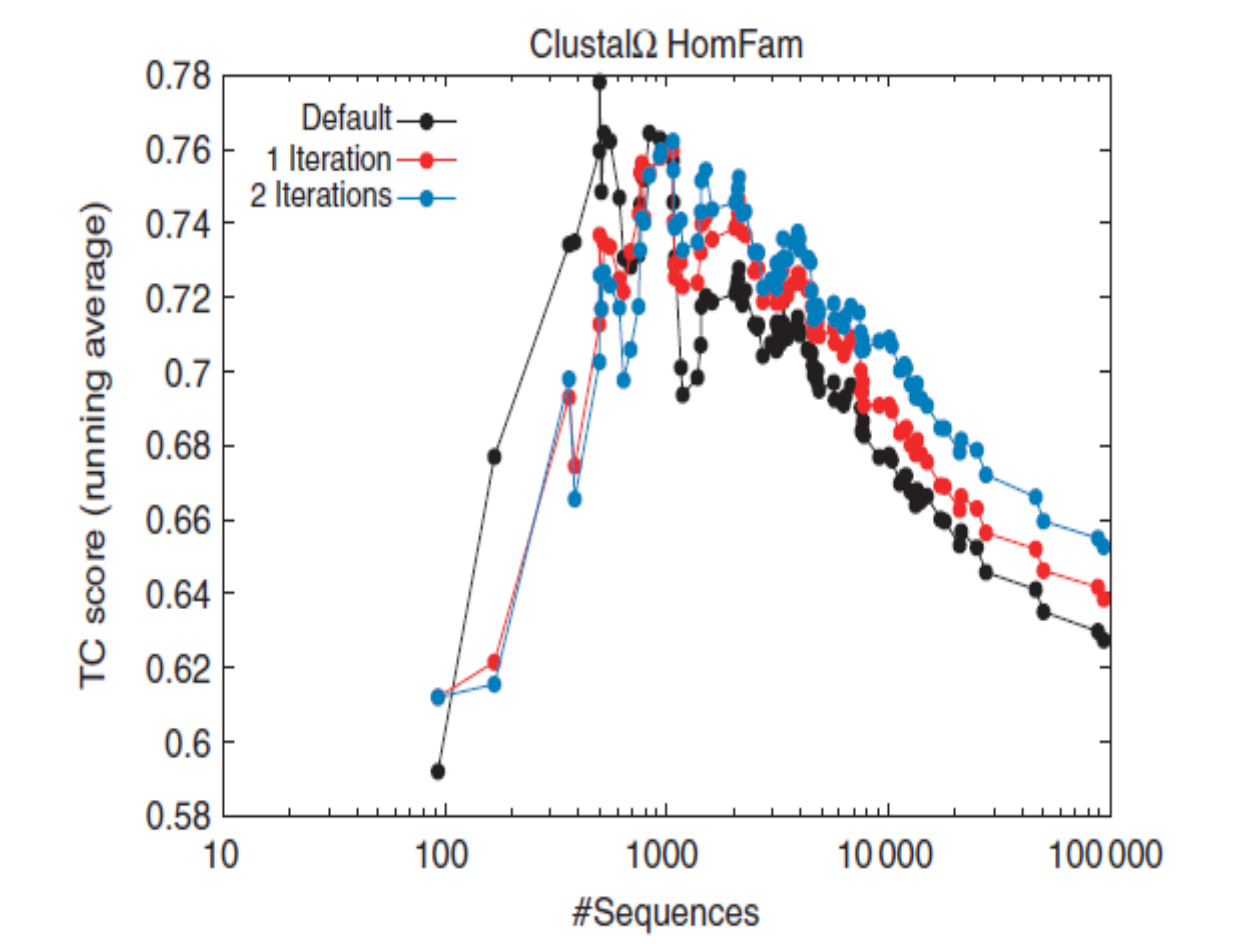
Discussion: EPA for HomFam and BAliBASE



- Average improvement in (**A**) 2.5%. HMMs taken from Pfam, benchmarking carried out using corresponding structure-based alignment in Homstrad.
- Average improvement in (**B**) over 30%. Here, test sets and EPA-HMMs were both derived from BAliBASE reference alignments.

Sievers F, Wilm A, Dineen D, et al. 2011

Discussion: Iteration of HomFam alignments.



- Points represent cumulative running averages of TC scores.
- Clustal Omega default results in black, results after 1 iteration in red, after 2 iterations in blue.
- Iterations are combined HMM/guide tree iterations; *x* axis, logarithmic and *y* axis, linear scale.

Sievers F, Wilm A, Dineen D, et al. 2011

Conclusion:

- Clustal Omega uses a modified iterative progressive alignment method and can align over 10,000 sequences quickly and accurately
- Clustal Omega is very useful for finding evidence of conserved function in DNA and protein sequences.
- Clustal Omega can be used to find promoters and other cis-regulatory elements.
- Clustal-Omega outputs can be read directly through several alignment viewers including Jalview.

Sievers F, Wilm A, Dineen D, et al. 2011

REFERENCES

- Sievers F, Wilm A, Dineen D, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology. 2011;7:539. doi:10.1038/msb.2011.75.