Clustal Omega

Danielle Campbell

BioE/CS598AGB

Clustal Omega

- Progressive alignment
  - Uses mBed to construct guide tree
    - $O(N \log_2 N)$, as opposed to $O(N^2)$
  - Uses HHalign to perform the alignment
    - Converts sequences/profiles to HMMs
    - Aligns HMMs for each node moving up the tree
- External profile alignment (EPA) optional
  - "Softens up" each pairwise alignment at each node by also comparing to external HMM

Sievers *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* 7: 539

Clustal Omega balances speed with accuracy

Table II Prefab results

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.

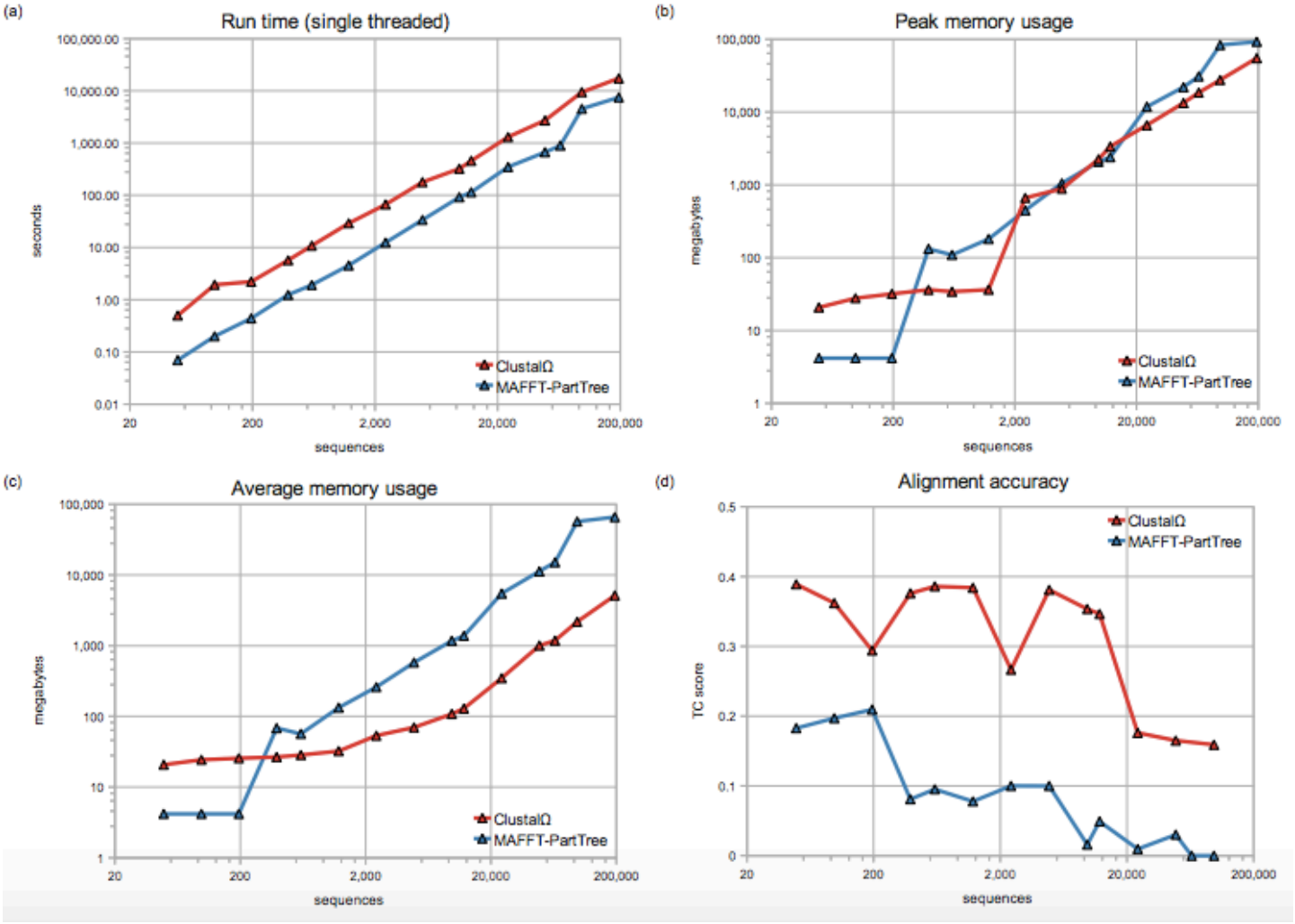
Table III HomFam benchmarking results

	$93 \leqslant N \leqslant 2957$ (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.

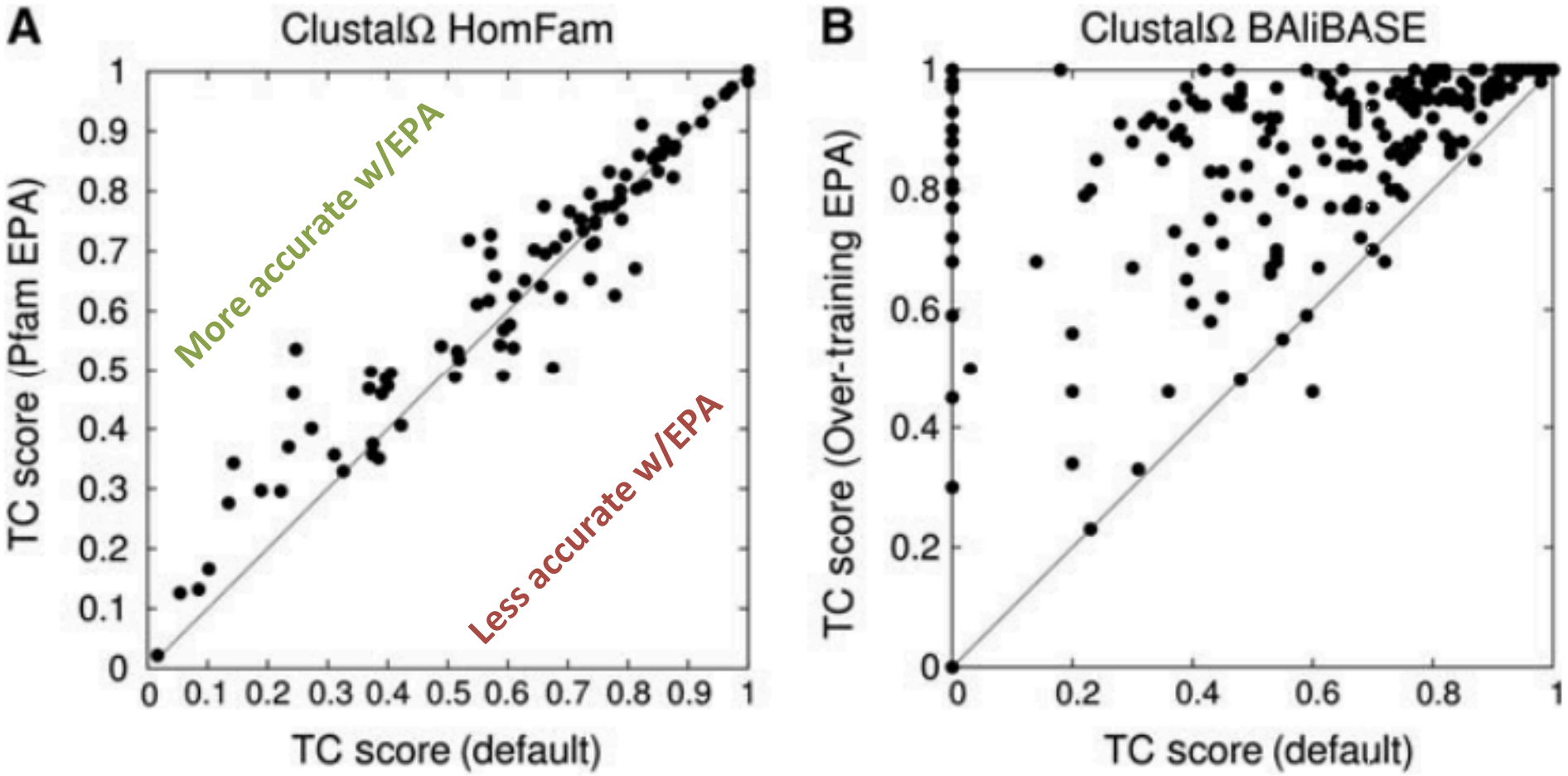
Sievers *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* 7: 539

Clustal Omega is ideal for alignments of many sequences



Sievers *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* 7: 539

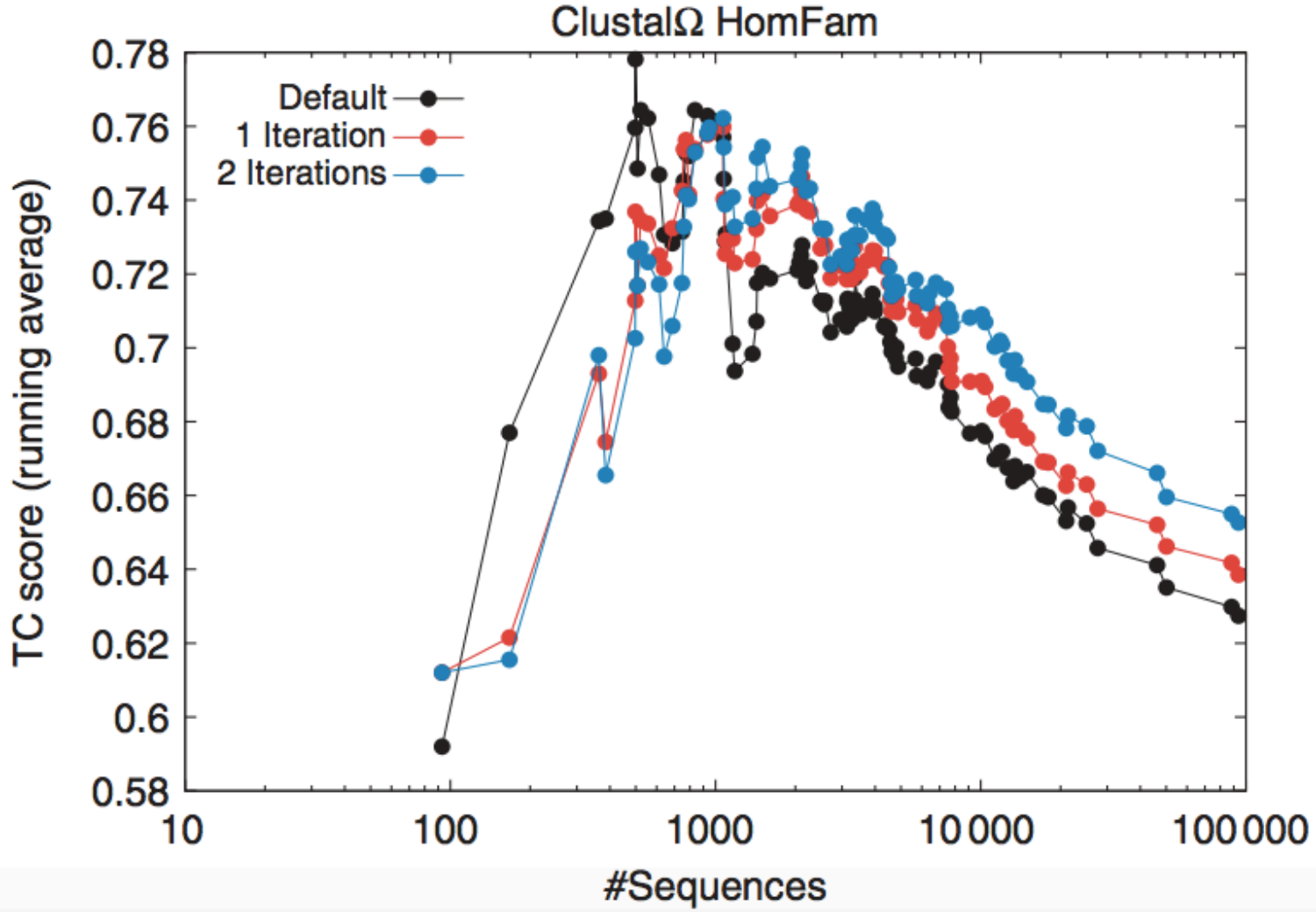
EPA generally increases accuracy



"we use the benchmark reference alignments themselves as external profiles"

Sievers *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* 7: 539

Iterating Clustal Omega is only beneficial for databases with more than 1000 sequences



Sievers *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* 7: 539

Conclusions & Criticisms

Authors say Clustal Omega:

- is fast, scalable, accurate
- allows the user to take advantage of precomputed data (EPA)
- benefits from multiple iterations

- EPA has small effects with realistic data
- iterations are detrimental on small datasets

Sievers *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* 7: 539

Sievers *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* 7: 539