Supplementary material: A benchmark of multiple sequence alignment programs upon structural RNAs

Paul P. Gardner[a] Andreas Wilm[b] Stefan Washietl[c]

[a]*Department of Evolutionary Biology, University of Copenhagen, Universitetsparken 15, 2100 Copenhagen Ø, Denmark*

[b]*Institut für Physikalische Biologie, Heinrich-Heine-Universität Düsseldorf, Universitätsstr. 1, D – 40225 Düsseldorf, Germany*

[c]*Institut für Theoretische Chemie und Molekulare Strukturbiologie, Universität Wien, Währingerstraße 17, A-1090 Wien, Austria*

Supplementary Materials

Program	**Version**	**Reference**	**Short description**
Align-m	2.1	[1]	Uses a non-progressive local approach to guide a global alignment.
ClustalW	1.82	[2, 3]	The classic progressive alignment program.
DIALIGN	2.2	[4–6]	Aligns gap-free segments as a whole without introducing gaps.
Handel	0.1 (dart)	[7, 8]	Phylogenetic alignment using a evolutionary hidden Markov model based on Thorne-Kishino-Felsenstein evolutionary model.
MAFFT	4.22	[9]	Rapid group-to-group alignment by fast Fourier transformation.
MUSCLE	3.51	[10, 11]	Employs a draft progressive step followed by an improved progressive and iterative refinement steps.
PCMA	2.0	[12]	Progressive method which aligns highly similar sequences as ClustalW and divergent groups by the T-Coffee strategy.
POA	2	[13]	Represents alignments as graphs which are directly aligned without the need for profiles.
ProAlign	0.5	[14]	Probabilistic progressive alignment combining a pair hidden Markov model and an evolutionary model.
Prrn	3.0 (scc)	[15]	Doubly nested randomised iterative alignment where group-to-group alignments are repeated to improve the overall score.
T-Coffee	1.37	[16]	Uses an alignment library to seek for maximum consistency of each residue pair with all other pairs of this library and guides the progressive step by means of this library.
Dynalign	second edition	[17, 18]	Simultaneously aligns and predicts the lowest free energy RNA secondary structure common to two sequences.
Foldalign	2.0.0	[19]	Structurally aligns two sequences using a light weight energy model in combination with RIBOSUM-like score matrices.
PMcomp	N/A	[20]	Computes and aligns base-pair probability matrices (calculated using McCaskill's algorithm [21]).
Stemloc	0.2 (dart)	[22, 23]	Alignment of RNA sequences using pre-folding and pre-alignment envelope heuristics.

Table 1
This table summarises the alignment methods used in this study.

Label	Command	Reference
Sequence Alignment		
Align-m (1)	`align_m -m RNA2`	[1]
Align-m (2)	`align_m -m RNA2 -p2m_Fmin 0.7 -p2m_nseq_min 5`	
Align-m (3)	`align_m -m RNA2 -s2p_go 10 -s2p_ge 1`	
Align-m (4)	`align_m -m RNA2 -s2p_go 10 -s2p_ge 1 -p2m_Fmin 0.7 -p2m_nseq_min 5`	
Align-m (5)	`align_m -m RNA2 -s2p_w 3`	
ClustalW	`clustalw -type=dna -align`	[2, 3]
ClustalW (qt)	`clustalw -type=dna -align -quicktree`	
DIALIGN	`dialign2-2 -n`	[4–6]
DIALIGN (it)	`dialign2-2 -n -it`	
DIALIGN (o)	`dialign2-2 -n -o`	
DIALIGN (it,o)	`dialign2-2 -n -it -o`	
Handel	`handalign.pl`	[7, 8]
MAFFT (fftnsi)	`fftnsi`	[9]
MAFFT (fftns)	`fftns`	
MAFFT (nwnsi)	`nwnsi`	
MAFFT (nwns)	`nwns`	
MUSCLE	`muscle`	[10, 11]
MUSCLE (nj)	`muscle -cluster1 neighborjoining -cluster2 neighborjoining`	
MUSCLE (mi32)	`muscle -maxiters 32`	
MUSCLE (nj,mi32)	`muscle -maxiters 32 -cluster1 neighborjoining -cluster2 neighborjoining`	
MUSCLE (m6)	`muscle -maxtrees 6`	
MUSCLE (nj,mt6)	`muscle -maxtrees 6 -cluster1 neighborjoining -cluster2 neighborjoining`	
MUSCLE (mi32,mt6)	`muscle -maxiters 32 -maxtrees 6`	
MUSCLE (nj,mi32,mt6)	`muscle -maxiters 32 -maxtrees 6 -cluster1 neighborjoining -cluster2 neighborjoining`	
PCMA	`pcma`	[12]
PCMA (agi20)	`pcma -ave_grp_id=20`	
PCMA (agi60)	`pcma -ave_grp_id=60`	
POA	`poa -v blosum80.mat`	[13]
POA (g)	`poa -do_global -v blosum80.mat`	
POA (p)	`poa -do_progressive -v blosum80.mat`	
POA (g,p)	`poa -do_global -do_progressive -v blosum80.mat`	
ProAlign (bw400)	`java -Xmx256m -jar ProAlign_0.5a0.jar -bwidth=400`	[14]
Prrn	`prrn`	[15]
Prrn (S10)	`prrn -S10`	
T-Coffee	`t_coffee`	[16]
T-Coffee (c)	`t_coffee -in=Mlalign_id_pair,Mclustalw_pair`	
T-Coffee (f)	`t_coffee -in=Mlalign_id_pair,Mfast_pair`	
T-Coffee (s)	`t_coffee -in=Mlalign_id_pair,Mslow_pair`	

Table 2
This table summarises parameters and references for applied sequence alignment methods corresponding to abbreviations used in the body of the manuscript. Due to space constraints we only display those parameters affecting algorithm methodology and not those for data input/output.

Label	Command	Reference
Structural Alignment		
Dynalign	`dynalign len2-len1+5 0.4 5 20 2 1 0`*	[17,18]
Foldalign	`foldalign -global -max_diff 25 -score_matrix global.fmat`	[19]
PMcomp	`pmcomp.pl`	[20]
PMcomp (fast)	`pmcomp.pl --fast`	[20,24]
Stemloc (slow)	`stemloc --global --multiple -verbose --nfold 1000 --norndfold`	[22,23]
Stemloc (fast)	`stemloc --global --multiple -verbose --nfold 110 --norndfold`	
Statistics		
SPS	`bali_score`	[25]
SCI	`RNAz`	[26]
Percent Sequence Identity	`alistat`	[27]

Table 3
This table summarises parameters and references for applied structural alignment methods corresponding to abbreviations used in the body of the manuscript. Due to space constraints we only display those parameters affecting algorithm methodology and not those for data input/output. $*$Note that $len1$ corresponds to the length of the shortest sequence and $len2$ corresponds to the length of the longest sequence.

References

[1] Van Walle, I., Lasters, I., and Wyns, L. (2004) Align-m: a new algorithm for multiple alignment of highly divergent sequences. *Bioinformatics,* **20**(9), 1428–1435.

[2] Thompson, J., Higgins, D., and Gibson, T. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.. *Nucl. Acids Res.,* **22**, 4673–4680.

[3] Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T. J., Higgins, D. G., and Thompson, J. D. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucl. Acids Res.,* **31**(13), 3497–3500.

[4] Morgenstern, B., Frech, K., Dress, A., and Werner, T. (1998) DIALIGN: finding local similarities by multiple sequence alignment. *Bioinformatics,* **14**(3), 290–294.

[5] Morgenstern, B. (1999) DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics,* **15**(3), 211–218.

[6] Morgenstern, B. (2004) DIALIGN: multiple DNA and protein sequence alignment at BiBiServ. *Nucl. Acids Res.,* **32**(suppl_2), W33–36.

[7] Holmes, I. and Bruno, W. J. (2001) Evolutionary HMMs: a Bayesian approach to multiple alignment. *Bioinformatics,* **17**(9), 803–820.

[8] Holmes, I. (2003) Using guide trees to construct multiple-sequence evolutionary HMMs. *Bioinformatics,* **19**(90001), 147i–157.

[9] Katoh, K., Misawa, K., Kuma, K.-i., and Miyata, T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucl. Acids Res.,* **30**(14), 3059–3066.

[10] Edgar, R. C. (2004) Muscle: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics,* **5**(1), 113.

[11] Edgar, R. C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucl. Acids Res.,* **32**(5), 1792–1797.

[12] Pei, J., Sadreyev, R., and Grishin, N. V. (2003) PCMA: fast and accurate multiple sequence alignment based on profile consistency. *Bioinformatics,* **19**(3), 427–428.

[13] Lee, C., Grasso, C., and Sharlow, M. F. (2002) Multiple sequence alignment using partial order graphs. *Bioinformatics,* **18**(3), 452–464.

[14] Löytynoja, A. and Milinkovitch, M. C. (2003) A hidden Markov model for progressive multiple alignment. *Bioinformatics,* **19**(12), 1505–1513.

[15] Gotoh, O. (1996) Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments.. *J. Mol. Biol.,* **264**, 823–838.

[16] Notredame, C., Higgins, D., and J., H. (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment.. *J. Mol. Biol.,* **302**, 205–217.

[17] Mathews, D. and Turner, D. (2002) Dynalign: An algorithm for finding the secondary structure common to two RNA sequences. *J. Mol. Biol.,* **317**(2), 191–203.

[18] Mathews, D. (2005) Predicting a set of minimal free energy RNA secondary structures common to two sequences. *Bioinformatics,* **(Advance Access published February 24)**.

[19] Hull Havgaard, J., Lyngsø, R., Stormo, G., and Gorodkin, J. (2005) Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%. *Bioinformatics,* **(Advance Access published January 18)**.

[20] Hofacker, I., Bernhart, S., and Stadler, P. (2004) Alignment of RNA base pairing probability matrices. *Bioinformatics,* **20**(14), 2222–2227.

[21] McCaskill, J. S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structures. *Biopolymers,* **29**, 1105–1119.

[22] Holmes, I. (2004) A probabilistic model for the evolution of RNA structure. *BMC Bioinformatics,* **5**(166).

[23] Holmes, I. (2005) Accelerated probabilistic inference of RNA structure evolution. *BMC Bioinformatics,* **6**(1), 73.

[24] Bonhoeffer, S., McCaskill, J., Stadler, P., and Schuster, P. (1993) RNA multi-structure landscapes. A study based on temperature dependent partition functions. *Eur Biophys J,* **22**(1), 13–24.

[25] Thompson, J., Plewniak, F., and Poch, O. (1999) A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res,* **27**(13).

[26] Washietl, S., Hofacker, I., and Stadler, P. (2005) Fast and reliable prediction of noncoding RNAs. *Proc Natl Acad Sci U S A,* **102**, 2454–2459.

[27] Eddy, S. SQUID - C function library for sequence analysis..