

LOCAL PAIRWISE STRUCTURAL RNA ALIGNMENTS BY PRUNING OF THE DYNAMICAL PROGRAMMING MATRIX

Jakob H. Havgaard, Elfar Torarinsson and Jan Gorodkin

University of Copenhagen, Division of Genetics and Bioinformatics, IBHV, Grønnegårdsvej 3, 1870 Frederiksberg C, Denmark, <http://genome.ku.dk/~gorodkin>

Abstract

The Sankoff algorithm for simultaneously folding and aligning RNA sequences is computationally very heavy. Recently a number of groups have applied various constraints to lower the computational requirements to reasonable levels. Whereas the original Sankoff algorithm as well as many of the implementations, only conduct global alignments, the FOLDALIGN implementation makes both local and global structural alignment. The most recent version of FOLDALIGN introduces pruning of the dynamical programming matrix as a simple and effective heuristic which lowers the time and memory requirements significantly without lowering the predictive performance. FOLDALIGN is currently one of few Sankoff algorithms capable of conducting local alignments while being a practical tool. It has also been used in genome-wide screen for putative RNA structures in corresponding, but unaligned regions between human and mouse. In addition to the pairwise version of FOLDALIGN we have also made a multiple alignment method which either takes the pairwise alignments or McCaskill basepair probability matrices as input.

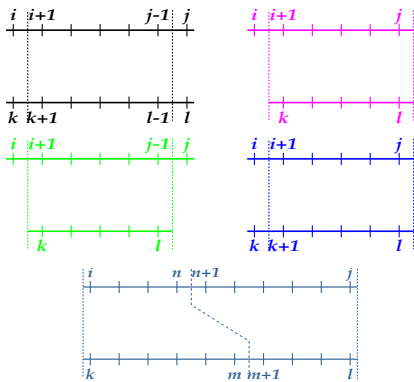
The principles of FOLDALIGN

FOLDALIGN available at: <http://foldalign.ku.dk>.

Algorithm:

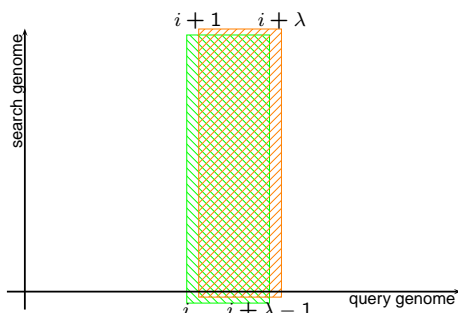
From (Havgaard *et al.*, 2005).

- Based on the work by Sankoff (1985).
- Simultaneously aligns and folds two sequences.
- Maximize a score: **local alignments**.
- Apply a subset of energy parameters.
- Energy parameters merged with base pair substitution scores.
- δ max length difference between any two subsequences.
- λ max motif length during local alignment.



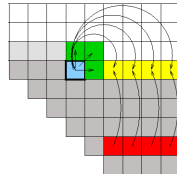
Local alignments

- Keep what is needed in the memory.
- Time cost: chopping up genome into $(2\lambda - 1)$ nt long pieces with λ in overlap.
- To find the best hits: use a BLAST like p-value approach.



Pruning the dynamical programming matrix

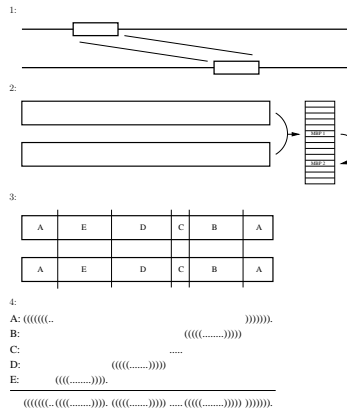
Filling out the dynamical pruning matrix inside out (Havgaard *et al.*, 2007):



- Fill out cells ahead and possibly overwrite.
- Pruning: never fill out cells for scores that does not exceed a length dependent threshold.

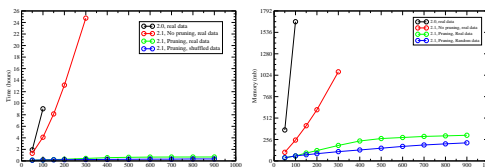
Backtrack: divide and conquer

- Local alignment found (only the score is computed).
- To find the alignment, realign as global alignment. Store branch point (six coordinates) and pointers to the next branch points. For each subalignment a pointer to the last branch point is kept.
- Split the alignment into unbranched segments using branch point pointers and coordinates.
- The smaller segments are realigned and backtracked without using the bifurcation part of the recursion.



Time and memory savings

Drastic reduction of time and memory requirements (due to pruning) as a function of the maximum motif length λ .



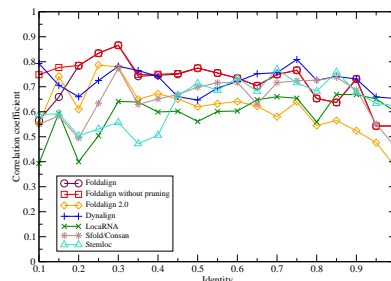
Localization performance same as without pruning.

Performance comparison

Comparing to other methods (global alignment).

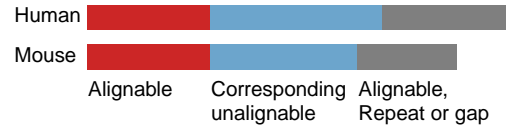
Method	Options	Time (s)	Max (Mb)	Ave (Mb)	Reference
FOLDALIGN 2.1	-global	1752	68	7	(Havgaard <i>et al.</i> , 2007)
FOLDALIGN 2.1	-no-pruning	8663	323	21	(Havgaard <i>et al.</i> , 2007)
FOLDALIGN 2.0	-global -max_diff 25 -score_matrix global.mat	18482	316	167	(Havgaard <i>et al.</i> , 2005)
Dynalign	maxitrace = 1 optimalLonely = 1	7080	17	9	(Harmanci <i>et al.</i> , 2007)
Locarna.pl		170	2	2	(Will <i>et al.</i> , 2007)
Consan	-m mixed80.mod	208146	1581	199	(Dowell <i>et al.</i> , 2006)
stemic	-na 100 -nl 1000	150608	2641	333	(Holmes, 2005)

Data from Dowell and Eddy (2006); ** two pairs could not be aligned.

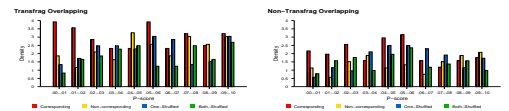


Screen unalignable regions between human and mouse

- Methods relying on sequence based alignments only useful for RNA structure prediction when ID > 60% (Gardner *et al.*, 2005).
- Searching corresponding human mouse regions not aligned in the UCSC browser (Torarinsson *et al.*, 2006).



- Genome wide 100.000 input pairs in our study.
- Screened 10 Chromosomes with transfrag data (Cheng *et al.*, 2005): 2×37.000 input pairs.
- For HSA20. Enrichment (2 fold) of candidates overlapping transfrags.
- For our cut-off criteria ($P < 0.03$) 1297 candidates in the 10 chromosomes (50% false positive rate).
- Experiments on top candidates: found 32 of 36 by RT-PCR and 4 of 12 on northern.
- Database available: http://genome.ku.dk/resources/hm_ncrna_scan.

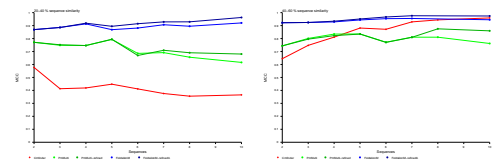


Multiple structural RNA alignments

FOLDALIGNM (Torarinsson *et al.*, 2007) is available at <http://foldalign.ku.dk>.

- Motivation: collect hits from local screen and search for multiple alignments families.
- Much overlap to PMcomp (Hofacker *et al.*, 2004). More efficient implementation; includes a realignment to consensus structure.
- Implementation use FOLDALIGN or McCaskill (1989) base pair probability matrices as input.
- Strength: few sequences and low sequence similarity.

Example on tRNAs:



Perspectives

- Further improvements: Pre-constraints.
- Better handling of structural inserts.
- Extend to handling of local multiple alignments.
- Room for improvement of scoring scheme (more data).
- Better merge scores for structure and alignment.
- Improve selection of candidates (p-value computation).
- Screen more organisms.

References

- Cheng *et al.*, 2005 Science 308:1149–1154.
 Dowell *et al.*, 2006 BMC Bioinformatics 7:400.
 Gardner *et al.*, 2005 Nucleic Acids Res 33:2433–2439.
 Harmanci *et al.*, 2007 BMC Bioinformatics 8:130.
 Havgaard *et al.*, 2005 Bioinformatics, 21:1815–1824.
 Havgaard *et al.*, 2007 PLoS Comput Biol 3:e193.
 Hofacker *et al.*, 2004 Bioinformatics, 20:2222–2227.
 Holmes, 2005 BMC Bioinformatics 6:73.
 McCaskill, 1990 Biopolymers 29:1105–1119.
 Sankoff, 1985 SIAM J Appl Math 45:810–825.
 Torarinsson *et al.*, 2006 Genome Res 16:885–889.
 Torarinsson *et al.*, 2007 Bioinformatics, 23:926–932.
 Will *et al.*, 2007 PLoS Comput Biol 3:e65.