Multiple sequence alignment methods of long non-coding RNAs

Suzana Ilinca Tudose
Bioinformatics And System Biology, M.Sc. Undergraduate
Albert Ludwig's University
Freiburg

August 20, 2011

Abstract

In this study we aim to compare the results of current multiple alignment tools and a self developed alignment pipeline in the special context of long non-coding RNAs (lncRNA).

1 Introduction

Long ncRNA is a rapidly advancing field of genetics, with yet only briefly studied roles (in gene regulation), organization, conservation or medical implications. It is however expected that they will play a great role in further genetic studies and progress, since they have such great potential. Given their (sometimes impressive) length or other particularities, we were inclined to think that they might need some rather special alignment algorithms. We will present in the following pages the results we got comparing alignments sets from the Ensemble epo alignmets, by Galaxy Multiz blocks and alignments generated by our own pipeline.

Outline I will now present the work done and describe the way the aligning pipeline works [Section 2], then show and discuss the results [Section 3]. In the end I will try to draw some conclusions in Section 4.

2 Work description

We based our study on 93 human lncRNAs we got from the lncRNAs data base (`http://lncrnadb.com/`), all the annotated ones available at the poin when we started. Also we restricted our study to 19 species, available in all three alignment aproaches.

Our alignment algorithm is basically to get possible homologous sequences for each species and to align them to the current set of sequences. Before moving to the next specie, we decide whether to keep the new sequence in the set or not, given some threshold values for the (normalized) sum of pairs score and the percentage of gaps it introduces in the alignment. Thus we enforce a certain minimum quality to our final alignemnts.

We find the potential homologous sequences using the online BLAT tool. We have a small Perl script, that does it automatically for us, for every specie - sequence pair that is need. Although the parsing of the HTML colorful encoding of the results gave us a little headache (because they used different ways of encoding the same thing and we had to check each the almost 2000 cases). We decided not to use the stand alone version of BLAT because the online one performs a lot faster on our machine (up to one minute instead of even 20 minutes sometimes) but also because we didn't need any se pecific parameter settings.

Using the stand alone BLAT makes it not trivial at all to decide which blocks should belong together or to simply simulate the best result on the web, generally in the case when several exons/blocks are found. The domcumentation tells us how to set the parameters and how to compute the score, but we still eventually got some results that scored better than the best one on the web. So we would allways get the same score as indicated by the online tool, but often it was not the best one. Also we didn't find information about how to best chose the blocks that should belong together.

For the Multiz alignments we used the online tool for stitching together the small blocks (about 3000 for our set). At the end we still got several blocks for some lncRNAs - one for each exon. The same problem we had with the epo alignments and we had to find a way merge them into one

multiple alignment per sequence, so that the scores are really comparable. It was a laborious work and we found many expected and unexpected problems along the way.

3 Results

The first encouraging result we got is depicted in [*Figure 1*]. As one can easily see, our alignment approach performed well from the score point of view. It is by far the most satisfying. But this is however not the "whole picture". We need to know of course how this alignments look like. If we got good scores on a 2-sequence alignment while the other approaches have 19, it is deffinitely not what we needed. In [*Figure 2*] we have plotted the percent of bases in the alignment against the number of species in the alignment.

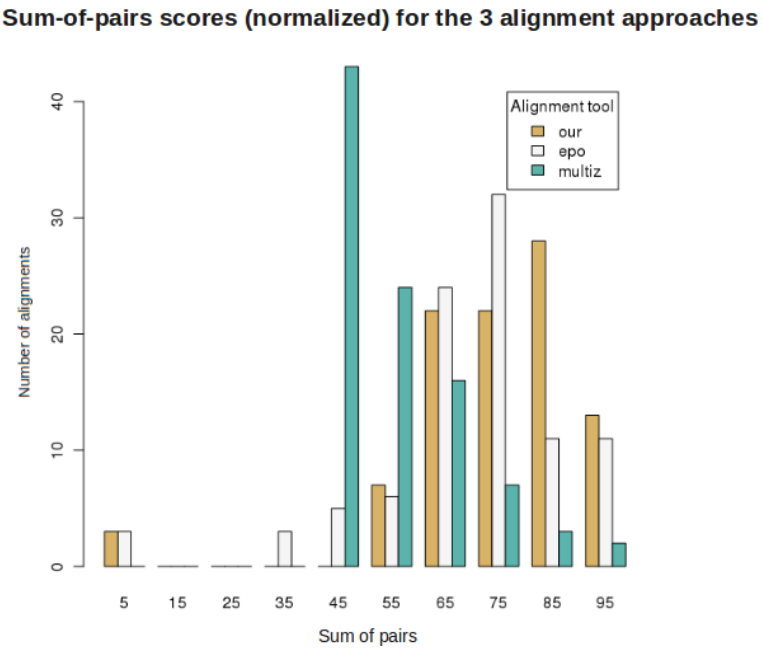

Figure 1: Number of alignlents with same score

Ecouraging as it is, that the results in the upper right corner of [*Figure 2*] are achieved using our pipeline, at this point we cannot consider our approach competitive. Both epo and multiz have performed better.

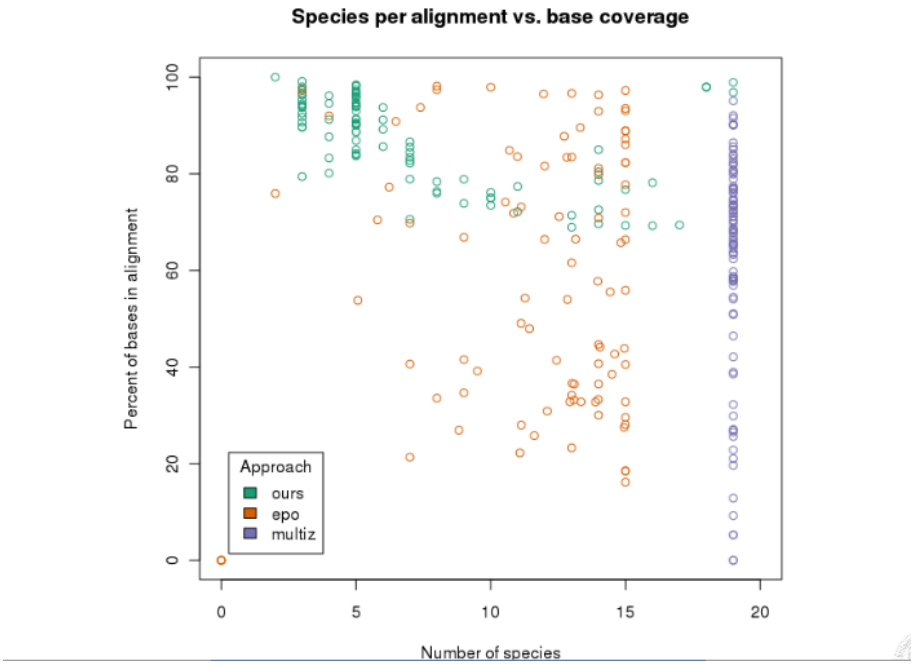

Figure 2: Percent of bases vs. the number of species

In the summary table in [*Figure 3*] we can see some other indicators to how each approach performed. I do find our results here very encouraging as well, as we think that many homologous sequences were not reported by our algorithm because of the windowsize of BLAT. This is not adjustable in the online version, so in order to improve the results we should switch again to using a local BLAT and adjusting its parameters accordingly.

One major draw back at the multiz alignments, besides the excessive fragmentation of the resulting blocks is the fact that in the fasta file one gets after the stitching of the blocks there is no information at all about the sequences chosen (except the specie). I find this very unsatisfying as it doesn't leave you the possibility to do a whole serie of checkings e.g. synteny. There is no possibility to change that from the web interface. In the epo alignments we found one case of synteny loss.

	our	multiz	epo
Loss of synteny	no	?	yes
Avg. # of sequences per alignment	7.1	19	11.6
Avg. sum-of-scores per alignment	77.8	56.1	68.5
Avg. % of bases in an alignment	87.5	62.1	57.3
mature mRNA alignments	yes	no	no

Figure 3: Result summary table

4 Conclusions

We worked hard, and achieved very little...

In the context of lncRNAs the epo alignments are probably the most balanced. Our approach would also be useful if only alignments with high similarity are needed (so the quality of the alignment is more important than the number of species in it). Multiz is very fragmented and offers too little information about the sequence, so personally I would take it as the last possibility.

As for further work I consider it is worth trying to do the same but using a local BLAT search. The main advantage of that would be the possibility to adjust the tile size but htere are certainly other parameters that are worth playing with. I think it would be useful to find a small set of parameters that would be interesting to study and have a script to run the whole thing with these different values for the parameters and pipe the results to a machine learning algorithm to find the best values.

5 References

Long non-coding RNAs: insights into functions, Nature Reviews Genetics 10, 155-159 (March 2009), Tim R. Mercer, Marcel E. Dinger & John S. Mattick

http://genome.ucsc.edu/

http://main.g2.bx.psu.edu/

http://www.ensembl.org/index.html

http://www.ebi.ac.uk/Tools/msa/muscle/

www.lncRNAdb.org