RNAlign2D – a novel RNA structural alignment tool based on pseudo-amino acid substitution matrix

Tomasz Woźniak[1], Małgorzata Sajek[2], Jadwiga Jaruzelska[1], Marcin Piotr Sajek[1]

[1]Institute of Human Genetics, Polish Academy of Sciences, Strzeszyńska 32, 60-479 Poznań, Poland

[2]Department of Human Molecular Genetics, Institute of Molecular Biology and Biotechnology, Faculty of Biology, Adam Mickiewicz University, Uniwersytetu Poznańskiego 6, 61-614 Poznań, Poland

Correspondence:

Marcin Sajek, PhD.

Institute of Human Genetics, Polish Academy of Sciences

60-479 Poznań

32, Strzeszyńska street

Email: marcin.sajek@igcz.poznan.pl

Abstract

Motivation

The function of RNA molecules is mainly determined by their secondary structure. Addressing that issue requires creation of appropriate bioinformatic tools that enable alignment of multiple RNA molecules to determine functional domains and/or classify RNA families. The existing tools for RNA multiple alignment that use structural information are relatively slow. Therefore, providing a rapid tool for multiple structural alignment may improve classification of the known RNAs and reveal the function of the newly discovered ones.

Results

Here, we developed an extremely fast Python based RNAlign2D tool. It converts RNA sequence and structure to pseudo-amino acid sequence and uses customizable pseudo-amino acid substitution matrix to align RNA secondary structures and sequences using MUSCLE. It is suitable for RNAs containing modified nucleosides and/or pseudoknots. Our approach is compatible with virtually all protein aligners.

Availability and implementation

RNAlign2D is available from https://github.com/tomaszwozniakihg/rnalign2d. It has been tested on Linux and MacOSX.

Supplementary information

Supplementary data are available at *Bioinformatics* online.

1 Introduction

RNA molecules are central players in various cellular processes, including protein biosynthesis and gene expression regulation (Morris and Mattick, 2014). Their function is mainly determined by the structure, which is often more conserved than the sequence (Capriotti and Marti-Renom, 2010). Thus far, there is data of secondary structures (2D structures) for > 100 000 RNAs and this number is growing (Danaee *et al.*, 2018). Bioinformatic tools for multiple alignment enable identification of motifs and domains which is crucial to predict the RNA function. The information about the structure significantly improves alignment quality. Several tools to align the structure of RNA molecules were developed, including multiple sequence and structure alignment tools, which are usually based on 2D structure prediction algorithms (e.g. TurboFoldII (Tan *et al.*, 2017), MAFFT (Katoh and Toh, 2008)). Alternatively, LocaRNA (Will *et al.*, 2007) and CARNA (Sorescu *et al.*, 2012) tools can be applied which may use fixed 2D structure as an input. However the above tools are relatively slow, especially in case of the long RNA sequences (e.g. 16S rRNA).

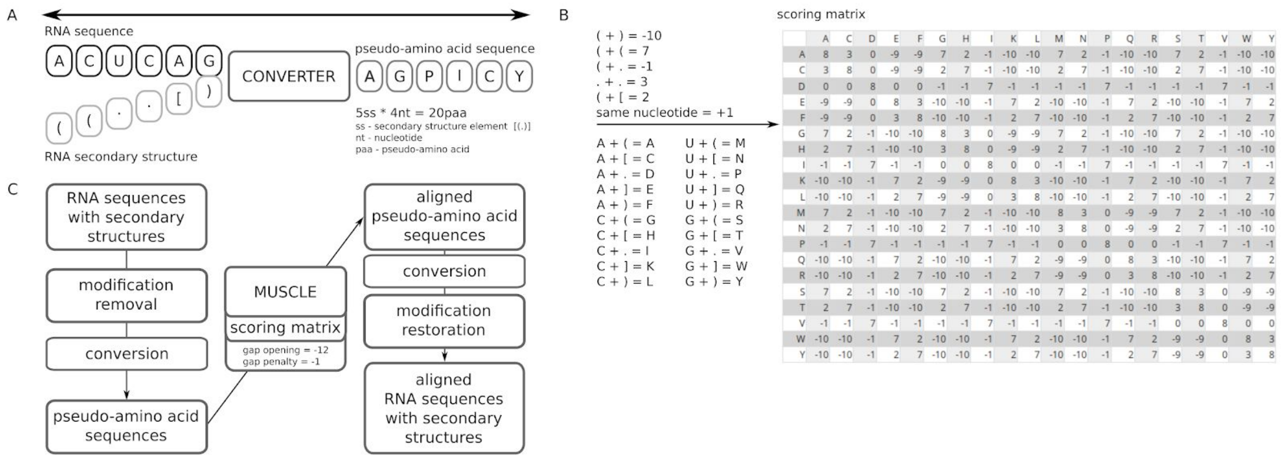
To solve this problem, we developed RNAlign2D, a rapid and accurate Python tool that produces alignments of multiple RNA molecules based mostly on 2D structure information. It does so by using a pseudo-amino acid substitution matrix and MUSCLE aligner (Edgar, 2004). It applies for modified and unmodified RNA sequences, as well as for sequences containing pseudoknots. Finally, the tool we generated can be customized to be compatible with virtually all multiple sequence alignment tools that perform protein alignment.

2 Implementation

2.1 Pseudo-amino acid conversion

A simple description of RNA sequence and the secondary structure can be performed by using only

4 characters - ‘A’, ‘G’, ‘C’ and ‘U’ for the sequence and 3 - ‘(‘, ‘.’, ‘)’ for the 2D structure. In case the sequence and the structure information should be combined, 12 possible nucleotide – dot/bracket pairs can be generated. Addition of ‘[‘ and ‘]’ characters for level 1 pseudoknots results in 20 combinations. For higher-level pseudoknots, the number of combinations increases respectively. Therefore, to make the sequence-structure description simple, we convert it to a pseudo-amino acid sequence. For that purpose two approaches have been used here. Namely, for the sequence without pseudoknots or level 1 pseudoknots only, each nucleotide is combined with its 2D structure element (e.g. ‘A’ and ‘.’, when ‘A’ is in the single-stranded region) and converted to a single pseudo amino acid (Figure 1A). However, for RNAs containing higher-level pseudoknots, only secondary structure elements are converted to pseudo-amino acids because in that case the number of sequence-structure combinations exceeds 20. Details of conversion for all 20 combinations are shown in Figure 1B.



	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	8	3	0	-9	-9	7	2	-1	-10	-10	7	2	-1	-10	-10	7	2	-1	-10	-10
C	3	8	0	-9	-9	2	7	-1	-10	-10	2	7	-1	-10	-10	2	7	-1	-10	-10
D	0	0	8	0	0	-1	-1	7	-1	-1	-1	-1	7	-1	-1	-1	-1	7	-1	-1
E	-9	-9	0	8	3	-10	-10	-1	7	2	-10	-10	-1	7	2	-10	-10	-1	7	2
F	-9	-9	0	3	8	-10	-10	-1	2	7	-10	-10	-1	2	7	-10	-10	-1	2	7
G	7	2	-1	-10	-10	8	3	0	-9	-9	7	2	-1	-10	-10	7	2	-1	-10	-10
H	2	7	-1	-10	-10	3	8	0	-9	-9	2	7	-1	-10	-10	2	7	-1	-10	-10
I	-1	-1	7	-1	-1	0	0	8	0	0	-1	-1	7	-1	-1	-1	-1	7	-1	-1
K	-10	-10	-1	7	2	-9	-9	0	8	3	-10	-10	-1	7	2	-10	-10	-1	7	2
L	-10	-10	-1	2	7	-9	-9	0	3	8	-10	-10	-1	2	7	-10	-10	-1	2	7
M	7	2	-1	-10	-10	7	2	-1	-10	-10	8	3	0	-9	-9	7	2	-1	-10	-10
N	2	7	-1	-10	-10	2	7	-1	-10	-10	3	8	0	-9	-9	2	7	-1	-10	-10
P	-1	-1	7	-1	-1	-1	-1	7	-1	-1	0	0	8	0	0	-1	-1	7	-1	-1
Q	-10	-10	-1	7	2	-10	-10	-1	7	2	-9	-9	0	8	3	-10	-10	-1	7	2
R	-10	-10	-1	2	7	-10	-10	-1	2	7	-9	-9	0	3	8	-10	-10	-1	2	7
S	7	2	-1	-10	-10	7	2	-1	-10	-10	7	2	-1	-10	-10	8	3	0	-9	-9
T	2	7	-1	-10	-10	2	7	-1	-10	-10	2	7	-1	-10	-10	3	8	0	-9	-9
V	-1	-1	7	-1	-1	-1	-1	7	-1	-1	-1	-1	7	-1	-1	0	0	8	0	0
W	-10	-10	-1	7	2	-10	-10	-1	7	2	-10	-10	-1	7	2	-9	-9	0	8	3
Y	-10	-10	-1	2	7	-10	-10	-1	2	7	-10	-10	-1	2	7	-9	-9	0	3	8

Fig. 1. Schematic representation of RNAlign2D work-flow. **A**. Basic concept of RNA sequence-structure conversion to pseudo-amino acid sequence. **B**. Conversion of 20 RNA sequence-structure elements to pseudo-amino acids together with score and default scoring matrix. **C**. Block diagram of RNAlign2D work-flow.

2.2 RNAlign2D tool

After conversion to pseudo-amino acids, the running of multiple sequence alignment programs dedicated to protein sequences provides the most adequate structural RNA alignment. In the case of RNAlign2D tool, the MUSCLE program is used which allows alignment of amino acid sequences, in a reliable and very fast manner. In RNAlign2D, MUSCLE uses a scoring matrix dedicated to RNA structural alignment. Default scoring matrix is shown in Fig. 1B.

RNAlign2D tool performs the following processing steps (Figure 1C): 1/ Removes common modifications from RNA sequences (to do so it uses modification abbreviations from MODOMICS database (Boccaletto *et al.*, 2018)). 2/ Converts secondary structures and sequences to pseudo-amino acid sequences. 3/ Runs MUSCLE program with given sequences, scoring matrix, and penalties. 4/ Converts aligned pseudo-amino acid sequences to RNA sequences and secondary structures. 5/ Restores original modifications for each sequence. RNAlign2D contains an alignment tool, predefined matrices, scoring matrix creation tool, and modification removal tool. It should be installed alongside the MUSCLE program.

2.3 RNAlign2D command-line interface

The program can be run by simply writing in the terminal: rnalign2d -i input_file_name -o output_file_name. There are also additional flags available that allow the user to provide his own scoring matrix, penalties for gap opening and/or extension and choose the running mode (‘simple’ or ‘pseudo’). Additional script ‘create_matrix.py’, allows the user to define customized scoring matrix.

An input file in the ‘simple’ mode is a fasta-like file with the header followed by the line containing a sequence and 2D structure in the dot-bracket format. In the ‘pseudo’ mode, the sequence line is omitted during conversion and alignment. If the structures with higher level pseudoknots are analysed in ‘simple’ mode – higher level pseudoknots are converted to unpaired residues.

3 Validation

RNAlign2D was compared to LocaRNA and CARNA using BraliBase2.0 (Wilm *et al.*, 2006) and RNAStralign (Tan *et al.*, 2017) benchmarks. The sums of positive scores and Positive Predictive Values as well as the running time were calculated for each program. In general, scores were similar for all the programs tested, however they are strongly dependent on the RNA family, e.g. alignment of 16S rRNA family was performed better by RNAlign2D, whereas alignment of telomerase family – by CARNA. Detailed benchmark preparation and the benchmark results are described in Supplementary Information.

4 Conclusion

RNAlign2D performs multiple sequence-structure alignment of RNA molecules with accuracy similar to the other programs tested. The great advantage of RNAlign2D is that it is much faster compared to other programs that have been used so far. This is especially convenient for long RNAs for which it works even several hundred times faster than LocaRNA or CARNA.

Acknowledgements

We thank Dr. David Mathews and Dr. Zen Tan for providing useful scripts for calculations of alignment accuracy and Dr. Tomasz Górecki for discussing the statistical analysis.

References

Boccaletto,P. *et al.* (2018) MODOMICS: a database of RNA modification pathways. 2017 update. *Nucleic Acids Res.*, **46**, D303–D307.

Capriotti,E. and Marti-Renom,M.A. (2010) Quantifying the relationship between sequence and three-dimensional structure conservation in RNA. *BMC Bioinformatics*, **11**, 322.

Danaee,P. *et al.* (2018) BPRNA: Large-scale automated annotation and analysis of RNA secondary structure. *Nucleic Acids Res.*, **46**, 5381–5394.

Edgar,R.C. (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

Katoh,K. and Toh,H. (2008) Improved accuracy of multiple ncRNA alignment by incorporating structural information into a MAFFT-based framework. *BMC Bioinformatics*, **9**, 212.

Morris,K. V. and Mattick,J.S. (2014) The rise of regulatory RNA. *Nat. Rev. Genet.*, **15**, 423–437.

Sorescu,D.A. *et al.* (2012) CARNA-alignment of RNA structure ensembles. *Nucleic Acids Res.*, **40**, 49–53.

Tan,Z. *et al.* (2017) TurboFold II: RNA structural alignment and secondary structure prediction informed by multiple homologs. *Nucleic Acids Res.*, **45**, 11570–11581.

Will,S. *et al.* (2007) Inferring Noncoding RNA Families and Classes by Means of Genome-Scale Structure-Based Clustering. *PLoS Comput. Biol.*, **3**, e65.

Wilm,A. *et al.* (2006) An enhanced RNA alignment benchmark for sequence alignment programs. *Algorithms Mol. Biol.*, **1**, 1–11.