CIAlign - A highly customisable command line tool to Clean and Interpret multiple sequence Alignments

Charlotte Tumescheit, Andrew E. Firth, Katherine Brown

Division of Virology, Department of Pathology, Cambridge University

ct518@cam.ac.uk

https://github.com/KatyBrown/CIAlign

Scan me

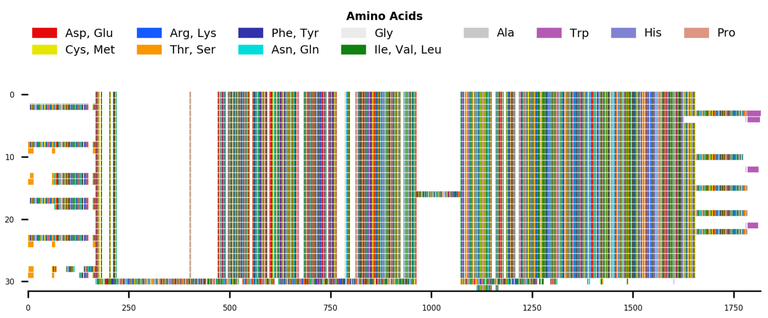


Figure 1: Mini alignment of an amino acid MSA (Input).

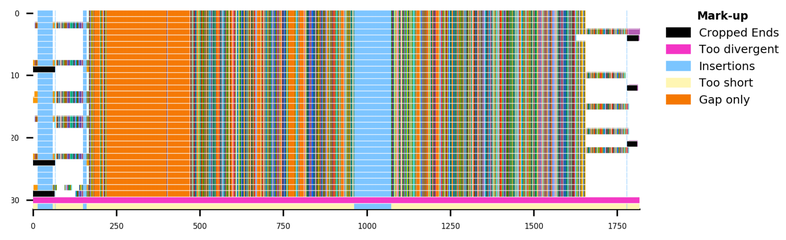


Figure 2: Markup of areas that are removed by CIAlign where the colour corresponds to the function that removed it.

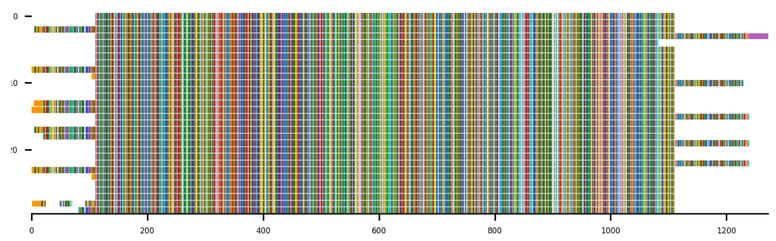


Figure 3: Resulting "cleaned" Mini alignment (Output), after CIAlign has been applied with all functions with default parameters

Motivation: Multiple Sequence Alignments (MSAs) are essential for many biological analyses, e.g. structure prediction, phylogenetic analysis, contig assembly, etc.

Problem: Poorly aligned or large gap regions in MSAs due to divergent or incomplete sequences (especially at the ends), insertions and deletions. This slows down computation and can impact conclusions without being biologically meaningful. Another common problem is length induced visualisation difficulty.

Solution: Development of a user-friendly, highly customisable tool that removes poorly aligned regions from an MSA, clearly shows what has been removed and gives a clear visualisation of even large MSAs.

Main features

- **Clean**
  - Remove sources of noise from an already aligned MSA
- **Visualise**
  - *Mini alignments* of amino acid or nucleotide MSAs, where each residue is represented by a coloured rectangle
  - For Input, Output and a markup of removed areas
- **Interpret**
  - Consensus sequence, Similarity Matrix, Sequence logos and Coverage plot
- **User intervention**
  - Many adjustable parameters, user can adjust which cleaning functions are used
- **Clarity**
  - Clear documentation of what has been removed and why

CIAlign is developed in Python 3.

Download: From Github or via pip

Usage: Start from terminal with inifile and/or command line parameter

Input: Aligned MSA in fasta format

Output: MSA in fasta format with certain areas removed, log files, plus additional files depending on which functions are being used

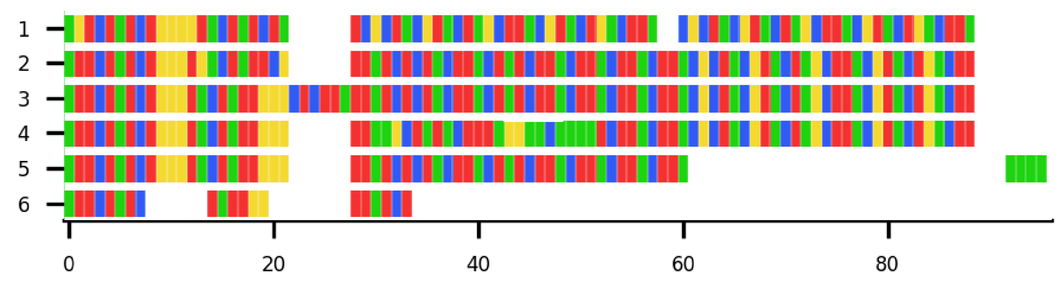
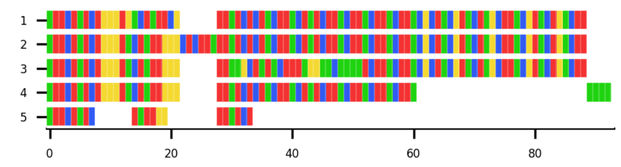
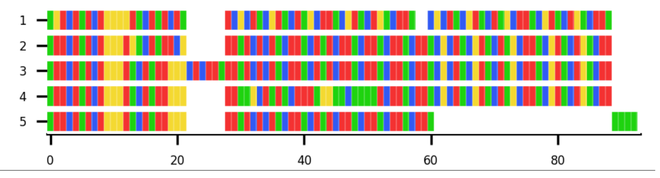
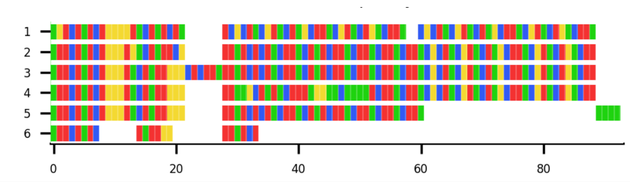
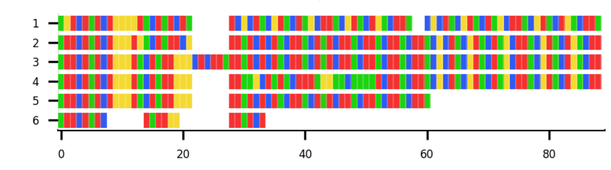
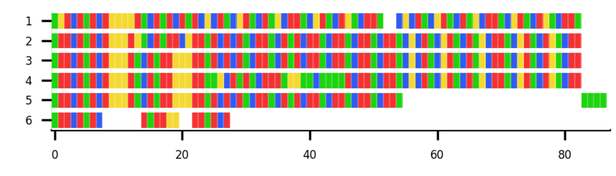


Figure 4: Input nucleotide MSA, see table for effect of each cleaning function

Function	Options	Output
Remove divergent	• set threshold for divergence	
Remove short	• min length	
Remove gap-only columns	• on or off	
Crop ends	• min proportion of sequence length for gaps to non-gap ratio threshold	
Remove insertions	• insertion min size • insertion max size • flanking regions min size	

232 COI sequences from different species, identified using BLAST against the NCBI Transcriptome Shotgut Assembly database

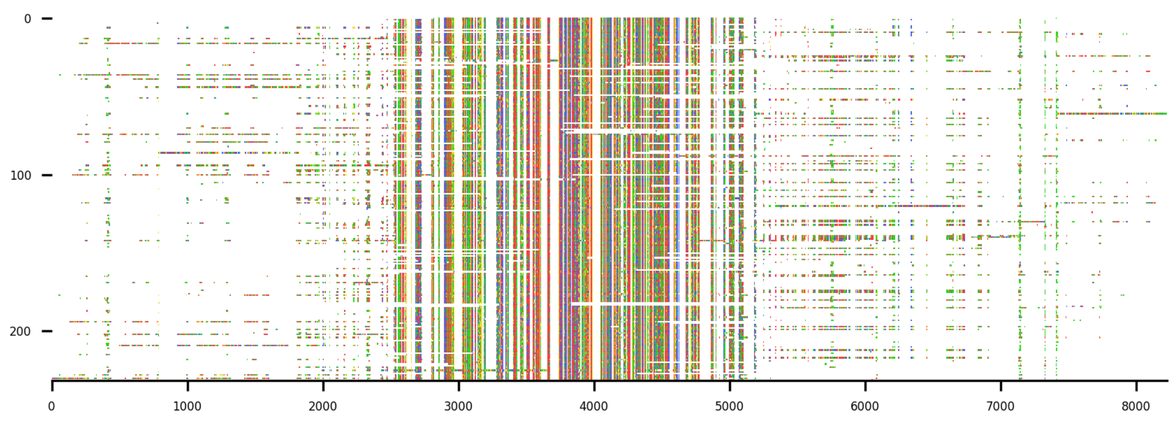


Figure 5: Input nucleotide MSA

Figure 6: Output MSA after using CIAlign using all functions with default settings

References

[1] T D Schneider and R M Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research*, 18:6097–6100, 1990.

European Research Council
Established by the European Commission

If you'd like to try out our tool, have a look at https://github.com/KatyBrown/CIAlign. Here you can also find the documentation. We'd love to hear your feedback! We are also happy to help you install and use CIAlign!