

CIAAlign

CIAAlign is a command line tool which performs various functions to parse and analyse a multiple sequence alignment (MSA).

The tool is designed to be highly customisable, allowing users to specify exactly which functions to run and which settings to use. It is also transparent, generating a clear log file and diagram showing exactly how the alignment has changed.

This allows the user to:

- Remove sources of noise from their MSA
- Crop of poorly aligned sequence ends
- Remove of insertions which are not present in the majority of sequences
- Remove of sequences below a threshold number of bases or amino acids
- Remove columns containing only gaps
- Remove sequences above a threshold level percentage of divergence from the majority
- Generate consensus sequences
- Visualise alignments
- Generate image files showing the alignment before and after parsing and showing which columns and rows have been removed
- Draw sequence logos
- Visualise coverage at each position in the alignment
- Analyse alignment statistics
- Generate a similarity matrix showing the percentage identity between each sequence pair

Requirements

- python \geq 3.6
- matplotlib \geq 2.1.1
- numpy \geq 1.16.3
- scipy \geq 1.3.0

Installation

? conda

The current release of CIALign can be downloaded directly using this link
Add the CIALign directory to your PATH environment variable as described here

Usage

Basic Usage

```
CIALign --infile INFILE --outfile_stem STEM --infile cialign.ini
```

Parameters

Parameter	Description	Default
<code>-infile</code>	Path to input alignment FASTA file	None
<code>-infile</code>	Path to ini file	None
<code>-outfile_stem</code>	Prefix for output files, including the path to the output directory	CIALign
<code>-silent</code>	Do not print progress to the screen	False

Functions

Specify which functions to run by adding the following optional arguments to the command

Parsing an MSA

Each of these steps will be performed sequentially in the order specified in the table below.

The parsed alignment after all steps have been performed will be saved as **OUTFILE_STEM_parsed.fasta**

Parameter	Description	Default Value
<code>-crop_ends</code>	Crop the ends of sequences if they are poorly aligned	False
<code>-crop_ends_mingap</code>	minimum gap size to consider when classifying a sequence as poorly aligned	10

Parameter	Description	Default Value
-remove_badlyaligned	Remove sequences with $\leq N$ proportion of positions at which the most common base / amino acid in the alignment is present	False
<i>-remove_badlyaligned_minperc</i>	Minimum proportion of positions which should be identical to the most common base / amino acid in order to be preserved	0.9
-remove_insertions	Remove insertions found in $\leq 50\%$ of sequences from the alignment	False
<i>-insertion_min_size</i>	Only remove insertions \geq this number of residues	3
<i>-insertion_max_size</i>	Only remove insertions \leq this number of residues	300
<i>-insertion_min_flank</i>	Minimum number of bases on either side of an insertion to classify it as an insertion	5
-remove_short	Remove sequences $\leq N$ bases / amino acids from the alignment	False
<i>-remove_minlength</i>	Minimum number of non-gap residues in a sequence to be preserved	50
-remove_gaponly	Remove gap only columns from the alignment	True

Generating a Consensus Sequence

This step generates a consensus sequence based on the parsed alignment. If no parsing functions are performed, the consensus will be based on the input alignment.

Output files:

- **OUTFILE_STEM_consensus.fasta** - the consensus sequence only
- **OUTFILE_STEM_with_consensus.fasta** - the parsed alignment plus the consensus

Parameter	Description	Default
-make_consensus	Make a consensus sequence based on the parsed alignment	False
<i>-consensus_type</i>	Type of consensus sequence to make - can be majority, to use the most common character at each position in the consensus, even if this is a gap, or majority_nongap, to use the most common non-gap character at each position	majority
<i>-consensus_keepgaps</i>	If there are gaps in the consensus (if majority_nongap is used as consensus_type), should these be included in the consensus (True) or should this position in the consensus be deleted (False)	False
<i>-consensus_name</i>	Name to use for the consensus sequence in the output fasta file	consensus

Visualising Alignments

Each of these functions produces some kind of visualisation of your alignment.

Mini Alignments

These functions produce “mini alignments” - images showing a small representation of your whole alignment, so that gaps and poorly aligned regions are clearly visible.

Output files:

- **OUTFILE_STEM_input.png (or svg, tiff, jpg)** - the input alignment
- **OUTFILE_STEM_output.png (or svg, tiff, jpg)** - the parsed output alignment
- **OUTFILE_STEM_markup.png (or svg, tiff, jpg)** - the input alignment with deleted rows and columns marked

Parameter	Description	Default
-plot_input	Draws a mini alignment for the input FASTA file	False

Parameter	Description	Default
-plot_output	Draws a mini alignment for the output FASTA file	False
-plot_markup	Draws the input alignment but with the columns and rows which have been removed by each function marked	False
<i>-plot_dpi</i>	DPI for mini alignments	300
<i>-plot_format</i>	Image format for mini alignments - can be png, svg, tiff or jpg	png
<i>-plot_width</i>	Mini alignment width in inches	5
<i>-plot_height</i>	Mini alignment height in inches	3

Sequence logos

These functions draw sequence logos representing your output (parsed) alignment. If no parsing functions are specified, the logo will be based on your input alignment.

Output_files:

- **OUTFILE_STEM_logo_bar.png (or svg, tiff, jpg)** - the alignment represented as a bar chart
- **OUTFILE_STEM_logo_text.png (or svg, tiff, jpg)** - the alignment represented as a standard sequence logo using text

Parameter	Description	Default
-make_sequence_logo	Draw a sequence logo	False
<i>-sequence_logo_type</i>	Can be bar, to draw the logo as a bar chart, text, to draw a standard sequence logo using text, or both, to draw both	bar
<i>-sequence_logo_dpi</i>	DPI for sequence logo	300
<i>-sequence_logo_font</i>	font (see NB below) for bases / amino acids in a text based sequence logo	monospace
<i>-sequence_logo_nt_per_row</i>	number of bases / amino acids to show per row in the sequence logo, where the logo is too large to show on a single line	50

Parameter	Description	Default
<i>-sequence_logo_filetype</i>	Image file type to use for the sequence logo - can be png, svg, tiff or jpg	png

NB: to see available fonts on your system, run `CIAAlign -list_fonts_only` and view `CIAAlign_fonts.png`

Coverage Plots

This function plots the number of non-gap residues at each position in the alignment.

Output file:

- **OUTFILE_STEM_input_coverage.png (or svg, tiff, jpg)** - image showing the input alignment coverage
- **OUTFILE_STEM_output_coverage.png (or svg, tiff, jpg)** - image showing the parsed alignment coverage

Parameter	Description	Default
-plot_coverage_input	Plot the coverage of the multiple sequence alignment	False
-plot_coverage_output	Plot the coverage of the multiple sequence alignment	False
<i>-plot_coverage_dpi</i>	DPI for coverage plot	300
<i>-plot_coverage_height</i>	Height for coverage plot (inches)	3
<i>-plot_coverage_width</i>	Width for coverage plot (inches)	5
<i>-plot_coverage_colour</i>	Colour for coverage plot (hex code or name)	#007bf5
<i>-plot_coverage_filetype</i>	File type for coverage plot (can be png, svg, tiff, jpg)	png

Analysing Alignment Statistics

These functions provide additional analyses you may wish to perform on your alignment.

Similarity Matrices

Generates a matrix showing the proportion of identical bases / amino acids between each pair of sequences in the MSA.

Parameter	Description	Default
– make_similarity_matrix_input	make a similarity matrix for the input alignment	False
– make_similarity_matrix_output	make a similarity matrix for the output alignment	False
<i>–make_simmatrix_keepgaps</i>	Include positions with gaps in either or both sequences in the similarity calculation	False
<i>–make_simmatrix_dp</i>	Number of decimal places to display in the similarity matrix output file	4
<i>–make_simmatrix_minoverlap</i>	Minimum overlap between two sequences to have non-zero similarity in the similarity matrix	1