

CIAAlign - Clean and Interpret Alignments

Charlotte Tumescheit, Dr. Katherine Brown, Dr. Andrew E. Firth

CIAAlign is a command line tool which performs various functions to clean 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 alignment markup showing exactly how the alignment has changed and what has been removed by which function.

This allows the user to:

- Remove sources of noise from their MSA
- Remove insertions which are not present in the majority of sequences
- Remove sequences below a threshold number of bases or amino acids
- Crop poorly aligned sequence ends
- 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 using CIAAlign cleaning functions 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
- Unalign the alignment
- Replace U's by T's

Requirements

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

Installation

The easiest way to install CIAAlign is using pip3:

```
pip3 install cialign
```

The current release of CIAAlign can also be downloaded directly using this link,

If you download the package directly, you will also need to add the CIAAlign directory to your PATH environment variable as described here

Usage

Basic Usage

```
CIAAlign --infile INFILE --outfile_stem STEM --inifile my_config.ini
```

Parameters

Parameters can be specified in the command line or in a config file using the naming system below.

A template config file is provided in CIAAlign/templates/ini_template.txt - edit this file and provide the path to the `--inifile` argument. If this argument is not provided command line arguments and defaults will be used.

Parameters passed in the command line will take precedence over config file parameters, which take precedence over defaults.

Command help can be accessed by typing `CIAAlign --help`

Parameter	Description	Default
<code>--infile</code>	Path to input alignment file in FASTA format	None
<code>--inifile</code>	Path to config file	None
<code>--outfile_stem</code>	Prefix for output files, including the path to the output directory	CIAAlign
<code>--silent</code>	Do not print progress to the screen	False
<code>--all</code>	Use all available functions with default parameters	False
<code>--help</code>	Show all available parameters with an explanation	None
<code>--version</code>	Show the version	None

Beside these main parameters, the use of every function and corresponding thresholds can be specified by the user by adding parameters to the command line or by setting them in the configuration file. Available functions and their parameters will be specified in the following section.

CIAAlign always produces a log file, specifying which functions have been run with which parameters and what has been removed. It also outputs a file that only specifies what has been removed with the original column positions and the sequence names.

Output files:

- `OUTFILE_STEM_log.txt` - general log file
- `OUTFILE_STEM_removed.txt` - removed columns positions and sequence names text file

Cleaning an MSA

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

The “cleaned” alignment after all steps have been performed will be saved as `OUTFILE_STEM_cleaned.fasta`

Parameter	Description	Default Value	Min	Max
<code>--remove_divergent</code>	Remove sequences with $\leq N$ proportion of positions at which the most common base / amino acid in the alignment is present	False	NA	NA
<code>--remove_divergent_minperc</code>	Minimum proportion of positions which should be identical to the most common base / amino acid in order to be preserved	0.65	0	1

Parameter	Description	Default Value	Min	Max
<code>--remove_insertions</code>	Remove insertions found in $\leq 50\%$ of sequences from the alignment	False	NA	NA
<code>- -insertion_min_size</code>	Only remove insertions \geq this number of residues	3	1	n_columns
<code>- -insertion_max_size</code>	Only remove insertions \leq this number of residues	200	1	1000
<code>- -insertion_min_flank</code>	Minimum number of bases on either side of an insertion to classify it as an insertion	5	0	n_columns/2
<code>--crop_ends</code>	Crop the ends of sequences if they are poorly aligned	False	NA	NA
<code>- -crop_ends_mingap_perc</code>	Minimum proportion of the sequence length (excluding gaps) that is the threshold for change in gap numbers.	0.05	0	0.5
<code>- -crop_ends_redefine_perc</code>	Proportion of the sequence length (excluding gaps) that is being checked for change in gap numbers to redefine start/end.	0.1	0	0.5
<code>--remove_short</code>	Remove sequences $\leq N$ bases / amino acids from the alignment	False	NA	NA
<code>- -remove_min_length</code>	Sequences are removed if they are shorter than this minimum length, excluding gaps.	50	0	n_columns
<code>--keep_gaponly</code>	Keep gap only columns in the alignment	False	NA	NA

Note: if the sequences are short (e.g. < 100), a low `crop_ends_mingap_perc` (e.g. 0.01) will result in a change of gap numbers that is too low (e.g. 0). If this happens, the change in gap numbers will be set to 2 and a warning will be printed.

Generating a Consensus Sequence

This step generates a consensus sequence based on the cleaned alignment. If no cleaning functions are performed, the consensus will be based on the input alignment. For the “majority” based consensus sequences, where the two most frequent characters are equally common a random character is selected.

Output files:

- `OUTFILE_STEM_consensus.fasta` - the consensus sequence only
- `OUTFILE_STEM_with_consensus.fasta` - the cleaned alignment plus the consensus

Parameter	Description	Default
<code>--make_consensus</code>	Make a consensus sequence based on the cleaned alignment	False
<code>- -consensus_type</code>	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
<code>- -consensus_keep_gaps</code>	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
<code>- -consensus_name</code>	Name to use for the consensus sequence in the output fasta file	consensus

Unaligning the Alignment

This function simply removes the gaps from the input or output alignment and creates an unaligned file of the sequences.

Output files:

- `OUTFILE_STEM_unaligned_input.fasta` - unaligned sequences of input alignment
- `OUTFILE_STEM_unaligned_output.fasta` - unaligned sequences of output alignment

Parameter	Description	Default
<code>--unalign_input</code>	Generates a copy of the input alignment with no gaps	False
<code>--unalign_output</code>	Generates a copy of the output alignment with no gaps	False

Replacing U's by T's

This function replaces the U nucleotides by T nucleotides without disturbing the sequence names.

Output files:

- `OUTFILE_STEM_T_input.fasta` - input alignment with T's instead of U's
- `OUTFILE_STEM_T_output.fasta` - output alignment with T's instead of U's

Parameter	Description	Default
<code>--replace_input</code>	Generates a copy of the input alignment with T's instead of U's	False
<code>--replace_output</code>	Generates a copy of the output alignment with T's instead of U's	False

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 cleaned output alignment
- `OUTFILE_STEM_markup.png` (or `svg`, `tiff`, `jpg`) - the input alignment with deleted rows and columns marked

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

Sequence logos

These functions draw sequence logos representing your output (cleaned) alignment. If no cleaning 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
<code>--make_sequence_logo</code>	Draw a sequence logo	False
<code>-sequence_logo_type</code>	Type of sequence logo - bar/text/both	bar
<code>-sequence_logo_dpi</code>	DPI for sequence logo	300
<code>-sequence_logo_font</code>	Font (see NB below) for bases / amino acids in a text based sequence logo	monospace
<code>-sequence_logo_nt_per_row</code>	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
<code>-sequence_logo_filetype</code>	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 output alignment coverage

Parameter	Description	Default
<code>--plot_coverage_input</code>	Plot the coverage of the input MSA	False
<code>--plot_coverage_output</code>	Plot the coverage of the output MSA	False
<code>-plot_coverage_dpi</code>	DPI for coverage plot	300
<code>-plot_coverage_height</code>	Height for coverage plot (inches)	3
<code>-plot_coverage_width</code>	Width for coverage plot (inches)	5
<code>-plot_coverage_colour</code>	Colour for coverage plot (hex code or name)	#007bf5
<code>-plot_coverage_filetype</code>	File type for coverage plot (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.

Output file:

- OUTFILE_STEM_input_similarity.tsv - similarity matrix for the input file
- OUTFILE_STEM_output_similarity.tsv - similarity matrix for the output file

Parameter	Description	Default
<code>--make_similarity_matrix_input</code>	Make a similarity matrix for the input alignment	False
<code>--make_similarity_matrix_output</code>	Make a similarity matrix for the output alignment	False
<code>- -make_simmatrix_keepgaps</code>	0 - exclude positions which are gaps in either or both sequences from similarity calculations, 1 - exclude positions which are gaps in both sequences, 2 - include all positions	0
<code>- -make_simmatrix_dp</code>	Number of decimal places to display in the similarity matrix output file	4
<code>- -make_simmatrix_minoverlap</code>	Minimum overlap between two sequences to have non-zero similarity in the similarity matrix	1