CLANS-Python Tutorial

Version 1.0

Overview

CLANS-Python is a program for visualizing the relationship between proteins based on their all- against-all pairwise sequence similarities. The program implements a version of the Fruchterman-Reingold force directed graph layout algorithm to present the sequence similarities in a 2D or 3D graph.

CLANS-Python is composed of two components:

- A command-line tool (clans_cmd.py), which gets as an input a set of sequences in FASTA format, performs all-against-all BLAST search to obtain a matrix of sequence similarities.
- A GUI-based visualizing tool (clans.py), which gets a matrix of sequence similarities and displays them as a dynamic graph using the Fruchterman-Reingold force-directed layout.

The pairwise similarity scores (E-values in case of BLAST search) are used to calculate the attractive forces between each sequence pair. The better the score (lower E-value), the higher the attractive force. In addition, each sequence repulses every other sequence with a certain force (inversely proportional to their distance in space). Clustering is achieved by iteratively moving sequences according to the force vector resulting from all pairwise interactions (attraction and repulsion).

Installation

Using conda environment:

Requirements: Conda / Anaconda installed on the target computer. An OS-specific version of Anaconda can be downloaded from: https://www.anaconda.com/products/individual-d.

- Download the CLANS-Python package code.
- Create a conda environment using the 'CLANS_python_3_8.yml' file:

  conda env create --file CLANS-Python/CLANS_python_3_8.yml
- Activate this environment when running the CLANS-Python program:

  conda activate CLANS_python_3_8

BLAST search:

In order to run a BLAST search using clans_cmd.py, BLAST+ (version 2.6.0. or newer) must be installed on the computer.

Using the command-line tool

The command-line tool can be used to perform a BLAST search in order to create a matrix of sequence similarities and/or to perform a specific number of iterations of the force-directed graph layout calculation (which can be later loaded and displayed in the visualizing tool).

Usage:
python clans_cmd.py -infile <fasta_file_path> -saveto <destination_file_path> [options]
or
python clans_cmd.py -load <network_file_path> -dorounds <number of iterations> -saveto <destination_file_path> [options]

Mandatory arguments:

-infile <fasta file path>: a FASTA file input for BLAST search
or

-load <network file path>: Load an existing network file in CLANS or tab-delimited formats

-saveto <destination file path>: A destination path for saving the output file (in CLANS format, by default)

Optional arguments:

-h, --help: Show this help message and exit

-input_format <’clans’/’delimited’>: The format of the network input file (when using the -load option).
Accepted formats: ‘clans’ (default) or tab-delimited.

-output_format <’clans’/’delimited’>: The format of the network output file (default is ‘clans’ format).

BLAST-related arguments:

-eval <E-value threshold>: E-value threshold for extracting BLAST HSPs (default=1.0)

-matrix <scoring matrix>: Scoring matrix for BLAST search.
Options: 'BLOSUM62', 'BLOSUM45', 'BLOSUM80', 'PAM30', 'PAM70' (default: BLOSUM62)

Clustering-related argumenets:

-dorounds <number of iterations>: Number of clustering iterations to perform (default=0)

-pval <similarity threshold>: A threshold for the similarity score (default=0.0001).
In case the similarity scores type is ‘hsp’ (result of a BLAST search), sequences with scores below the threshold are considered connected.
In case the similarity scores type is ‘att’ (attraction values), the sequences with scores above the threshold are considered connected. The default in this case is 0.1.

-cluster2d: Perform the clustering in 2D instead of 3D (default: cluster in 3D).

--cooling <COOLING>: A multiplier for the 'maxmove' parameter. 0 < Cooling ≤ 1. By default, set to 1 which causes the graph to keep moving until the user stops it. When cooling<1, maxmove gradually converges to 0 and the graph points stop moving.

--maxmove <MAXMOVE>: The maximum distance a point is allowed to move per iteration (default=0.1).

--att_val <ATT_VAL>: A multiplier factor for the calculation of the attractive force between each two sequences (default=10.0).

--att_exp <ATT_EXP>: An integer number - determines how the attractive force scales with the distance between each two vertices in the graph. Default = 1, attraction increases linearly with the distance.

--rep_val <REP_VAL>: A multiplier factor for the calculation of the repulsive force between each two sequences (default=10.0).

--rep_exp <REP_EXP>: An integer number - determines how the repulsive force scales with the distance between each two vertices in the graph. Default = 1, repulsion decreases linearly with the distance.

--dampening <DAMPENING>: A value between 0 and 1, determines to what extent the movement vector of the last movement affects the current movement. The lower it is, the greater the last movement's influence (default=0.2).

--gravity <GRAVITY>: A minimal force that attracts each sequence towards the origin of the graph and prevents unconnected clusters/sequences from drifting apart indefinitely. It scales linearly with the distance from origin (default=1.0).

Using the GUI-based visualization tool

Open the GUI from the command-line

Usage:
python clans.py [-load <network file path>*] [options]*
When clans.py is run without an input file, the GUI is opened empty and loading a file is possible from the ‘File’ menu of the GUI.

Optional arguments:

-h, --help: Show this help message and exit

-load <network file path>: Load an existing network file in CLANS or tab-delimited formats.

-format <’clans’/’delimited’>: The format of the network input file (default is ‘clans’ format).

-dorounds <number of iterations>: Number of clustering iterations to perform (default=0)

-pval <similarity threshold>: A threshold for the similarity score (default=0.0001).
In case the similarity scores type is ‘hsp’ (result of a BLAST search), sequences with scores below the threshold are considered connected.
In case the similarity scores type is ‘att’ (attraction values), the sequences with scores above the threshold are considered connected. The default in this case is 0.1.

-cluster2d: Perform the clustering in 2D instead of 3D (default: cluster in 3D).

--cooling <COOLING>: A multiplier for the 'maxmove' parameter. 0 < Cooling ≤ 1. By default, set to 1 which causes the graph to keep moving until the user stops it. When cooling<1, maxmove gradually converges to 0 and the graph points stop moving.

--maxmove <MAXMOVE>: The maximum distance a point is allowed to move per iteration (default=0.1).

--att_val <ATT_VAL>: A multiplier factor for the calculation of the attractive force between each two sequences (default=10.0).

--att_exp <ATT_EXP>: An integer number - determines how the attractive force scales with the distance between each two vertices in the graph. Default = 1, attraction increases linearly with the distance.

--rep_val <REP_VAL>: A multiplier factor for the calculation of the repulsive force between each two sequences (default=10.0).

--rep_exp <REP_EXP>: An integer number - determines how the repulsive force scales with the distance between each two vertices in the graph. Default = 1, repulsion decreases linearly with the distance.

--dampening <DAMPENING>: A value between 0 and 1, determines to what extent the movement vector of the last movement affects the current movement. The lower it is, the greater the last movement's influence (default=0.2).

--gravity <GRAVITY>: A minimal force that attracts each sequence towards the origin of the graph and prevents unconnected clusters/sequences from drifting apart indefinitely. It scales linearly with the distance from origin (default=1.0).

Input / Output file formats:

The visualization tool requires as an input a file with sequences identifiers and pairwise similarity scores. The input file can be provided in one of two file formats: ‘CLANS’ format or tab-delimited format.

CLANS format

A file in 'CLANS' format (.clans), can be created by clans_cmd.py (as an output of the BLAST search), by the CLANS web-utility in the MPI Bioinformatics Toolkit (https://toolkit.tuebingen.mpg.de/tools/clans) or saved in a previous session of the visualization tool.

The 'CLANS' file format must contain the following blocks of information:

- The first line must be: *sequences=<number of sequences>*
- The sequences block: the original sequences in FASTA format (the order of the sequences is important and is further used to index the sequences, starting from 0).

  ```
  <seq>
  >seq0
  MSGRGKQGGKARAKAKTRSSRAGLQFPVGR
  >seq1
  LAAEVLELAGNAARDNKKTRIIPRHLQLAIRNDEELNKLLSGVT
  </seq>
  ```

- The coordinates block: the positions of the sequences in the 3D space.
  Every line contains the sequence index and a value for the X, Y, Z coordinates ($-1 \geq X,Y,Z \leq 1$).

  ```
  <pos>
  0 0.142 0.281 0.104
  1 0.298 0.631 0.913
  </pos>
  ```

- The similarity scores block: the E-values or attraction values for the pairwise sequence similarities.

  ```
  <hsp>
  0 1:5.1e-05
  0 4:1.1e-02
  0 5:6.8e-04
  </hsp>
  ```

  or

  ```
  <att>
  0 1 0.1
  0 4 0.2
  0 5 0.3
  </att>
  ```

The file may contain additional blocks:

- Parameters block: a list of all the parameters that were used in the calculation and presentation of the saved session.

  ```
  <param>
  rounds_done=264
  cluster2d=false
  pval=0.2
  attfactor=10.0
  attvalpow=1
  repfactor=10.0
  repvalpow=1
  cooling=1.0
  dampening=0.2
  maxmove=0.1
  minattract=1.0
  </param>
  ```

- Groups block: a list of groups with their definitions and assigned sequences.

  ```
  <seqgroups>
  name=Proteobacteria
  size=10
  ```

```
hide=0
color=255;0;0;255
numbers=1;3;5;6;10;13;18;19;20;21;23;27;33;35;42;49;52;
</seqgroups>
```

Tab-delimited format

This file type should contain a list of non-redundant sequence-pairs and their similarity scores.

Mandatory columns:

- Sequence1 ID / unique name
- Sequence2 ID / unique name
- Similarity score

Optional column:

- Type of score (‘hsp’ for P-value and ‘att’ for attraction value)

The Graphical User Interface (GUI) controls

Once loading a network file (from the command-line or by using File -> Load file from the GUI), the graph presentation of the sequences appears in the central part of the GUI and most of the button-controls become enabled.

Menus

File menu:

- Load file: Loads a network file, containing at least sequences names and pairwise similarity scores in one of the two accepted formats (see detailed explanations in the previous section):
    - CLANS format.
    - Tab-delimited format.
- Save to file: saves the current graph presentation into a file, either in ‘CLANS’ format or as tab-delimited file.
    - The ‘CLANS’ format can keep the data-points current coordinates, the groups definitions and all the running parameters. Thus, it is recommended for further visualization using the CLANS-Python tool.
    - The tab-delimited format is recommended only as an API with other tools, such as Cytoscape, as it keeps only the minimal information needed to present a network (pairs of sequence names and similarity scores).
- Save as image: saves the currently displayed graph session in PNG format.
- Quit: closes the application including all its open windows.

Configure menu:

- Layout
    - Fruchterman-Reingold: opens a dialog window, allowing the user to configure the parameters that are used while clustering the sequences using the Fruchterman-Reingold graph layout (a detailed explanation of all the parameters is found in the command-line tool section).

Interaction with the graph area

CLANS-Python has two distinct interaction modes within the graph area:

Move/Rotate/Pan mode:

- Rotating the graph is done by holding the left mouse button pressed + dragging the mouse.

- Panning the graph is done by shift + left mouse button + dragging the mouse.
- Moving the selected data-points to a different location in the graph is done by CTRL + left mouse button + dragging the mouse.

Manual selection mode:

- To select a specific data-point, locate the mouse cursor in this point and click the left mouse button. Another click on a selected point will deselect it.
- To select all the data-points within a rectangular area, hold the left mouse button pressed + drag the mouse.

Zoom in/out is done by scrolling the mouse down/up. It is in possible in both interaction modes.

Buttons

Clustering options:

- Initialize: generates random coordinates to all the data-points.
- Start clustering: starts the iterative clustering process. The number of iterations that were done is displayed below the graph area.
- Stop clustering: stops the clustering process. It be resumed from the same point by clicking 'Resume clustering'.
- Cluster in 3D / 2D combo-box: determines whether the clustering process is done in 3D or 2D (default is 3D, unless stated differently in the running parameters). Setting the clustering to be done in 2D will change the graph view to 2D as well.
- P-value / Attraction value threshold: this threshold determines at which similarity score sequences are considered as connected. When the similarity score is a p-value, all the sequence-pairs with lower score are considered connected. In case the similarity score is an 'attraction value', sequence-pairs with similarity score above this threshold are considered connected.

Interaction mode combo-box:

Switches between the two modes of interaction with the graph (see above). The default is 'Move/Rotate/Pan' mode.

Selection options:

CLANS-Python enables to select a subset of data-points and perform all kind of operations on this subset. The selected data-points are marked in the graph by having a bigger size and a magenta outline color.

- Data-points / Groups selection combo-box: switches between the two modes of selection (this is relevant in 'Manual selection' interaction mode and in case the data is divided into groups).
  In data-points selection mode, the points that are clicked or within the selected area, are being selected. In groups selection mode, when a point is clicked, all the points that belong to the same group are being selected.
- Select all: selects all the data-points.
- Clear: clears the selection.
- Select by name: this option enables to enter a search term and select data-points according to their sequence name or header. Clicking on this button opens a find dialog, in which a search term can be entered. The results are displayed in the 'Search results' window (see detailed explanation in the 'Windows' section).
- Edit selected: This button is enabled just in case there is at least one selected data-point. Clicking on it opens a window, presenting the selected sequences IDs and headers. The 'Selected subset' window can remain open and the display gets updated whenever there is a change in the selection subset (see detailed explanation in the 'Windows' section).

Viewing options:

- 3D / 2D view: switches the graph view (not the clustering) between 3D and 2D view.
  The 2D view can be useful for producing images, as the connecting lines are displayed 'behind' the data-points. Another option which is only available in 2D view is displaying and moving the group names (if any).

- Z-index combo-box: this option is only enabled if there are defined groups and the view is set to 2D. It allows the user to determine whether the Z-indexing of the data-points will be done automatically (the default) or according to the groups order (which can be set in the 'Manage groups' window).
  Note: Z-index by groups order is only allowed when the clustering process has stopped, as it slows the graphics significantly.

- Connections: when this button is checked, the connecting lines (edges) are presented in the graph.
- Selected names: when this button is checked, the names (headers) of the selected data-points are presented next to them (enabled when there is at least one selected point).
- Selected numbers: when this button is checked, the unique number-IDs of the selected data-points are presented next to them (enabled when there is at least one selected point).
  The selected names and numbers cannot be displayed together. When one button s checked, the other one is automatically unchecked.

- Selected subset view: when this button is checked, the view changes to present the selected subset only. Any operation that is done while it is checked, is performed on the subset only. For example, it is possible to start the clustering considering the subset data-points only. When unchecking the 'Selected subset view' button, the view changes back to the exact full-data presentation that was displayed before. If a clustering was performed on the subset, it is not automatically saved, but can rather be saved to a file while in 'subset view'. The 'Selected subset view' option is only enabled when there are at least 5 selected data-points.

Groups options:

In order to help the user to navigate in the clustering map and examine whether the clustering meets other classifications, CLANS-Python allows to define groups of data-points, which can be displayed in the graph by different colors. The groups can be pre-defined in the CLANS input-file (inside the <seqgroups> block) or in the GUI, by selecting data-points and adding them to a new or existing group.

- Edit groups: opens a dialog presenting a list of all the existing groups (in parentheses the number of group members). When selecting a group, the following options are possible:
  - Edit group: opens another dialog, in which it is possible to change the group name, the size and the color in which the group's data-points are presented in the graph. At the bottom of this window there is a list of the group's members, in which it is possible to select sequences and remove them from the group. Any change that is done in this dialog is updated in the graph.

  - Delete group: Remove this group's definition. The data-points that were assigned to this group are not assigned to any group anymore and get the default presentation (black color).

  - Move up: Move this group one step up in the order of the groups list.
  - Move down: Move this group one step down in the order of the groups list.
    The groups order is only relevant when viewing the graph in 2D and setting the Z-index to be done by the groups order.

- Add selected to group: this option is enabled when there is at least one selected data-point. It opens a dialog, allowing the user to add the selected data-points to a new group or to an existing group (if any). If the user chooses to create a new group, a new dialog is opened, allowing the user to define the group name, size and color parameters.

- Remove selected from group: removes the selected data-points from their group(s) is they were assigned to any group. They get the default presentation (black color).

- Show group names: When this button is checked, a list of all the group names is presented on the left side of the graph area. It is only available in 2D view.
- Move group names: This option becomes available when the 'Show group names' button is checked. When it is checked, the user can click on a group-name and drag it with the mouse to a desired location in the graph. This can useful to mark certain clusters / groups when making a figure from CLANS map. When this option is checked, the rotation/panning of the graph is not possible.

Windows

Selected subset window

This window presents the IDs and headers of the sequences in the selected subset. It is possible to mark one or multiple lines (sequences) by using the Shift / CTRL buttons or by clicking the left mouse button + dragging the mouse.

The following operations can be performed on the marked sequences:

- Highlight in graph: highlights the relevant data-points in the graph with turquoise color.
- Set as selected subset: leaves only the marked sequences as the selected subset (removing the other from the subset).
- Remove from subset: removes the marked sequences from the selected subset.
- Find in subset: opens a find dialog, in which a search term can be entered. If there are sequences in the subset that match the search term, they are marked in blue and the above operations can be done on them.

All the changes to the selected subset are updated in the graph as well.
The window can either remain open or be closed. The sequences display gets updated whenever there is a change to the selection subset.

Search results window:

This window presents the sequences that match the search term entered by the user.
It is possible to mark one or multiple lines (sequences) by using the Shift / CTRL buttons or by clicking the left mouse button + dragging the mouse. Clicking the 'Highlight all' button marks all the lines.

The following operations can be performed on the marked sequences:

- Add to selected subset: adds the marked sequences to the selected subset.
- Set as selected subset: sets the marked sequences as the selected subset. If there were other sequences in the subset, they are removed.
- New search: opens the find dialog again and allows the user to enter a new search term.

All the changes to the selected subset are updated in the graph as well.