CLANS Tutorial

Overview

CLANS is a Java 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 can get as an input a set of sequences in FASTA format, perform an all-against-all BLAST search to obtain a matrix of sequence similarities and display it as a dynamic graph using the Fruchterman-Reingold layout. Alternatively, a matrix with precomputed "attraction" values can be provided.

The E-values of the BLAST HSPs are used to calculate the attractive forces between each sequence pair. The lower (better) the 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).

Usage

Prerequisites for running CLANS

- Java Runtime Environment (JRE) should be installed on the computer.

Input file

A file in special ‘CLANS’ format (.clans file) that was created either by the CLANS web-utility in the MPI Bioinformatics Toolkit (for the initial run) or by a previous session of the CLANS tool (CLANS saved-file).

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 (0<X,Y,Z<1).
  <pos>
  0 0.142 0.281 0.104
  1 0.298 0.631 0.913
  </pos>
- The BLAST HSPs block: the E-values for the pairwise sequence similarities.
  <hsp>
  0 1:5.1e-05
  0 4:1.1e-02
  0 5:6.8e-04
  </hsp>

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>
  maxmove=0.1
  pval=1.0
  </param>
- The rotation matrix block: the rotation matrix for the current clustering.
  <rotmtx>
  1.0;0.0;0.0;
  0.0;1.0;0.0;
  0.0;0.0;1.0;
  </rotmtx>

BLAST search

The stage of all-against-all BLAST search to obtain the matrix of pairwise sequence similarities in 'CLANS' special format (.clans file) can be done using the CLANS web-utility in the MPI Bioinformatics Toolkit. It accepts sequences in FASTA format (up to 10,000 sequences) and produces a file in CLANS format that can be later loaded and visualized in the CLANS tool. The scoring matrix and the E-value threshold for the BLAST search can be set by this utility.

Opening the CLANS graphical user interface (GUI)

The CLANS executable (clans.jar) is a Java JAR file. It can be executed in two ways:

1. From the command-line.
2. By launching the GUI and loading a pre-calculated 'CLANS' or matrix file.

1. Executing CLANS from the command line

The command to execute CLANS from the command line is:

java [-Xmx4G] -jar clans.jar [-load <clans file>]

Note: clans.jar must have executing permissions (*chmod +x clans.jar* will grant such permissions).

Optional parameters:

- -load <path of clans file>: Opens the CLANS GUI and loads the sequences information into it using a 'CLANS' formatted file. When -load is omitted, the CLANS GUI is started empty and a CLANS input-file can then be loaded using the 'Load Run' menu item.
- -Xmx<number>m / -Xmx<number>G: An optional Java parameter, which specifies the maximum memory allocation pool (in Mb or in Gb) for the Java virtual machine (JVM) and can be omitted or increased if needed. This parameter may be useful when loading very large files (>50Mb). It can then be set to -Xmx8G and even more (depending on the memory capabilities of the computer).

2. Launching the GUI directly from the file browser

When Java Runtime Environment is installed on the computer, double-clicking the clans.jar executable opens CLANS graphical interface (without a pre-loaded file). Please note that when launching the program in this way, it is not possible to set the memory allocation.

Clustering the sequences using CLANS graphical interface

If the GUI was opened without a file, the first stage is to load a file in 'CLANS' format: Menu -> File -> Load Run.

1. **Loading a 'CLANS' file that was created by the CLANS web-utility in the MPI Bioinformatics Toolkit:**
   In this case, this is the initial visualization of the sequences. The sequences are presented as dots in the 3D space and their positions are randomly determined. The clustering process will start by clicking the Start run button.

2. **Loading a saved-file that was created by a previous session of CLANS:**
   In this case, the sequences are visualized exactly as they were saved in the previous session. The clustering process may be continued from the same point by clicking the Start run / Resume button.

Button options:

- Initialize: randomize the sequence positions in space prior to the clustering process.
- Start run / Stop / Resume: start a new clustering process (at the first time or after initializing) / Stop an existing clustering process at a certain iteration / Resume the clustering from the same point.
- Select / Move: by default, the Move option is toggled and enables to rotate the 3D-space using the mouse.
  When the Select option is toggled (the button is colored in blue), it is possible to select specific sequences by marking a specific area on the 3D graph using the mouse. The selected sequences are then marked in red and different operations can be performed on them.
- Show selected: display the names of the selected sequences in a different window. By default, displays all the sequences.
- Select All / Clear Selection: toggle between selecting all the sequences and clearing the current selection.
- Zoom on selected / Show all: enable to zoom in to the selected sequences / Zoom out again to display all the sequences.
- Use P-values better than <P-value threshold>: consider only sequence pairs with E-values better (=lower) than specified (in the next text field) as connected. Changing the P-value threshold may influence the clustering (it affects the calculation of the attractive forces) as well as the connections display (when checking the 'show connections' option). The P-value can be given as an integer, a float number or in exponential notation (for example: 2.0e-10). Please note that after changing the P-value threshold you must hit the 'return' key to submit the change to the program.
  * The text field right to the P-value threshold (cannot be changed) presents the highest (=worst) E-value among the BLAST HSPs.

Checkboxes:

- show names: display the sequence names on the graph.
- show numbers: display the sequence indices on the graph.
- show connections: display the edges connecting the sequences in the graph. The color of the edge reflects the pairwise sequence similarity (darker grey for higher similarity and lighter for lower similarity) and can be set via Menu -> Draw -> Change color (dot connections). Please consider that displaying the connections during the clustering process slows the graphical visualization, especially when there is a large number of sequences.

Selected Sequences window (Show selected):

When selecting sequences and pressing the Show selected button, a window is opened and presents the names of the selected sequences (by their original order). At the bottom of the

window there are buttons enabling to perform different operations on the list of sequences. Please note that when pressing the Show selected button when no sequence is selected, the window presents all the sequences in the CLANS file.

- Find: open a search window, in which a text (exact or not) can be entered and searched among the sequences. When clicking the OK button (note that the Return key will not execute the operation), the sequences in which the search term is found are then marked in blue. Clicking the OK button again completes the operation and presents only the marked sequences in the Selected Sequences window (here too, only the OK button executes the operation and not the Return key). The operation can be cancelled by clicking the Back button.
- Show all names / show selected names: toggle between showing only the selected sequences and showing the names of all the sequences in the CLANS file.
- Save to file: save to file the current displayed sequences (selected / all / sequences containing a certain search term).
- Back: go one operation back within the Selected Sequences window.
- Clear: clear the last operations memory and retain only the last selection. After Clear it is not possible to display the initial selection again.
- Close: close the Selected Sequences window and retain the last selection made.

* Please note that defining a new selection in the Selected Sequences window (using Find), affects the selections in the main GUI window and changes it accordingly.

Menu options

File

Commonly-used options:

- Load Run: open a CLANS file (that was created by either the MPI Toolkit web-utility or saved by a previous run of CLANS) and displays the sequences as dots in the 3D-space according to the parameters and positions saved in the CLANS file.
- Save Run: save the current display (including the positions of the sequences and other defined parameters) in a 'CLANS' formatted file.

Advanced options:

- Save attraction values to file: save a list of the pairwise attraction values that meet the P-value threshold, to a file in the following format:
  seq1_index seq2_index attraction_value.
  (Attraction value = -log(E-value), divided by the highest value).
- Save 2d graph data: save the sequences names together with their (X,Y) coordinates in the following format:
  seq_index seq_name X Y
- Print view: print or saves the current graphical presentation in PDF format.

Misc

Commonly-used options:

- Extract selected sequences: save to file the currently selected sequences in FASTA format.
- Hide singletons: remove sequences that are not connected to any other sequence from the graph.
- Cluster in 2D: when this option is checked, the clustering is performed in two dimensions instead of three. Clustering in 2D is recommended when generating figures out of CLANS graph.

Advanced options:

- Use selected subset: display only the selected sequences (similar to Zoom-In).
- Use parent group: undo the last 'Use selected subset' operation.
- Set rotation values: set the values for the current rotation matrix (9 values, separated by commas).
- Rescale attraction values: when this option is checked, the attraction values are normalized according to the current P-value threshold, where 0 is the current lowest attraction value and 1 is the highest attraction value. Please note that rescaling lowers the attraction values and thus makes the clusters less condensed. (After checking/unchecking this option, put the cursor on the P-value threshold text-field and hit the 'Return' key, for the change to take action).
- Only draw every Nth round: set a new value (instead of 1) for the iterations interval for drawing the new positions of the dots in the graph. This makes the drawing less smooth but speeds up the clustering process.

Draw

Commonly-used options:

- Set dot size: set the size of the dots representing the sequences in the graph (the default and minimum is 2).
- Set selected circle size: set the size of the circle highlighting the selected sequences (dots) and the sequence groups.
- Center graph: set the current view on the center of the graph.
- Antialiasing: when this option is checked, spatial anti-aliasing is enabled (the graphics is nicer but slower, especially if the connecting lines are displayed). It is recommended to turn-on the anti-aliasing in the end of the clustering process, for the purpose of generating an image of CLANS map (it smooths the lines and shapes of the graph).
- Stereo: when this option is checked, a stereo image is displayed. The stereo image is composed of two identical 2D graphs, with a small angle between them, that are viewed separately by the left and right eyes of the viewer, to give the perception of 3D depth.
- Change stereo angle (0-360): change the angle between the left and right views of the stereo image (the default angle is 4).

Advanced options:

- Change Font: set the font used in the graph display (affects the text that appears inside the graph area, for example: the sequences names).
- Change color (dot connections): open a window, where the colors of the edges (dot connections) can be set according to their P-values (by default, the color gradient is from light grey for the worst P-values to black for the best P-values).
    - Set the P-value thresholds for the bins: it is possible to set a value for each bin separately or define the minimal and maximal values and use the 'Value gradient' button to equally distribute the values between the bins.
    - Set the colors of the bins: it is possible to set a color for each bin separately or define the colors of the minimal and maximal values and use the 'Color gradient' button to create a gradient from these colors.
- Change color (Foreground): set the foreground color of the text displayed in the graph area (the default is black).
- Change color (Background): set the color for the background of the graph area.
- Change color (Selected): set the color of the ovals highlighting the selected sequences.
- Change color (BLAST hits numbers): change the color of the HSP sequence numbers, presented when choosing Window -> Show BLAST hits for sequence and the option Draw -> Show hsp sequence numbers is checked.

- Change color (BLAST hits circles): change the color of the circles highlighting the sequences having BLAST hits for a selected sequence (when choosing Window -> Show BLAST hits for sequence).
- Color dots by sequence length: when this option is checked, the dots representing the sequences are colored according to their length (yellow=shortest, blue=longest, gradient=in-between). (After checking/unchecking this option, put the cursor on the graph area and left-click the mouse for the change to take action).
- Color by edge "frustration": when this option is checked, the edges in the graph are colored according to whether they are longer(red) or shorter(blue) than they should be according to the attraction values in the matrix.
- Show origin: when this option is checked, the origin (0,0,0) is marked in red on the graph area. (After checking/unchecking this option, put the cursor on the graph area and left-click the mouse for the change to take action).
- Show info: when this option is checked (it is checked by default), information about the current clustering is displayed (the edges coloring, maximum X,Y coordinates, current rotation matrix).
- Show HSP sequence numbers: this option is related to the 'Show blast hits for sequence' option (from the 'Windows' menu item). When it is checked, and the BLAST hits for a certain sequence are highlighted, the hits sequence numbers are also presented on the graph (next to the points representing them).
- Zoom: set a zoom factor for the view (default: 100%; fits all vertices to the screen).

Windows

Commonly-used options:

- Show options window: open a pop-up window which enables to set several parameters related to the clustering algorithm. For any change of parameter to take action, clicking the 'return' key or the 'maxmove' button is needed.
    - Cooling: a multiplier for the 'maxmove' parameter (see below), can be set between 0 and 1. When cooling=1 (the default), maxmove does not converge to 0 and the dots keep moving infinitely. When cooling<1, maxmove converges to 0 and the graph reaches a state where the dots stop moving at all (the rate of this convergence depends on the value of the cooling parameter. The closer it is to 1, the slower the convergence = more iterations are needed to reach convergence).
    - Current cooling: display (cannot be set) the changing "temperature" of the system during the clustering process. When the cooling parameter is set to 1, the current cooling (temperature) remains 1 as well and does not change. But when the cooling parameter < 1, the system "cools down" during the clustering process until it reaches convergence.
    - Maxmove: the maximum distance a point is allowed to move per round (the default is 0.1).
      It makes sense to increase the maxmove parameter when decreasing the cooling to allow bigger movements of the points in each round, since the number of rounds is limited.
    - Attract value: a multiplier factor for the calculation of the attractive force between each two sequences (default=10).
    - Attract exponent (int): determine how the attractive force scales with the distance between each two vertices in the graph (default=1, attraction increases linearly with the distance).
    - Repulse value: a multiplier factor for the calculation of the repulsive force between each two sequences (default=10).
    - Repulse exponent: determine how the repulsive force scales with the distance between each two vertices in the graph (default=1, repulsion decreases linearly with the distance).

- Dampening: a value between 0 and 1, determines to what extent the movement vector of the last movement affects the current movement (default=0.2). The higher the dampening parameter, the higher the previous movement influence. When it is set to 0, there is no influence.
- Min. attraction: a minimal force that attracts each sequence towards the origin of the graph (also called “gravity”, default=1). This gravity force keeps unconnected clusters/sequences from drifting apart indefinitely. It scales linearly with the distance.
- Cluster for rounds: set the number of iterations to be performed in the clustering process whenever the Start Run/Resume button is pressed (default=-1). When it is set to -1, the number of rounds is infinite and will be stopped only by the user.

- Selected: open the ‘Selected Sequences’ window (same as using the ‘Show selected’ button), which displays the names of the selected sequences. By default, it presents the names of all the sequences. If the selection is changed while the ‘Selected Sequences’ window is already open, there is a need to press the ‘Show selected’ button again to update the presentation in the window.
- Edit Groups: The Edit Groups window enables to set/edit different attributes for sequence groups (pre-defined or selected interactively) and to perform different operations on the groups. When saving the current run as a ‘CLANS’ file, the groups and their settings are saved in a separate block, marked by the <seqgroups> tag.

  Sequence groups can be defined or added in several ways:
  - Manually in the CLANS file, in the following format:
    ```
    <seqgroups>
    name=group1
    type=1
    size=6
    hide=0
    color=153;0;51;255
    numbers=435;436;437;438;439;440;
    </seqgroups>
    ```
  - By searching for clusters (Windows -> Find clusters) and defining each cluster as a sequence group using the Add each as separate sequence group button.
  - From the Edit Groups window, by using the Add selected button.

  The last option is the one commonly used. To form a group in this way, sequences must be selected and converted to a group using the Add selected button. Upon clicking that button, a pop-up window allows to name the group. The group then appears with the given name in the Edit Groups window, followed in brackets by the number of sequences in it. In order to visualize the group in the CLANS map, the checkbox Draw groups must be clicked on.

  By default, a new group is shown as red colored circles of size 4.

  The following attributes can be changed for each group after selecting it using the mouse (blue highlight):
  - Shape (field to the right of the group names): the circles can be changed to other shapes by clicking in this field.
  - Size (buttons above and below the Shape field): the size of the shapes (default=5) can be changed.
  - Hide/Show button: whether this group is displayed in the graph or not.
  - Change name button: change the name of group.
  - Change color button: change the color of the circles (or other shapes) marking a group.

If the Color group names option is checked, the group names will also be colored in the same color.

Other operations that can be done using the buttons:

- Set as selected: set the sequences belong to the currently selected group as selected in the graph.
- Move up / Move down: move the selected group one step up or down in the groups list.
- Delete: remove the selected group(s) from the groups list.

Advanced options:

- P-value plot: open a window showing the distribution of P-values (or attraction values) for the current dataset.
- Show blast hits for sequence: first, a window with all the sequences names is opened and it is possible to select one sequence. Then, a window showing the distribution of HSPs throughout the selected sequence is opened. Clicking with the mouse inside this new window first highlights the selected sequence. Moving the red slider throughout the distribution, highlights the related HSPs in the graph (in the color defined by Draw -> Change color (BLAST hits circles)). If the option Draw -> Show HSP sequence numbers is checked, the HSP sequence numbers are presented in the graph as well (in the color defined by Draw -> Change color (BLAST hits numbers)).

- Find Clusters: determine which clusters exist in the dataset. The clustering can be done by one of three methods:
  - N-linkage clustering (the default method): a simple clustering according to a minimal number of connections between the sequences, defined by the user (the default is 1). Two sequences are defined as connected if their sequence similarity E-value is better than the current P-value threshold. Thus, changing the P-value threshold parameter may influence the clustering.
  - Convex clustering: all the sequences, which their average sequence-cluster attractive force is better than X*SD (standard deviation) of the average attraction for the dataset, are grouped together (X can be set by the user, default is 0.5). This method is much slower than the N-linkage clustering.
  - Network-based clustering: each sequence forms a node of the input layer for a network. These nodes emit the number of the cluster the sequence belongs to (at the beginning: number of clusters=number of sequences). The "weight" of each value is proportional to the -log(P-value) of the blast hit. The second layer integrates all these inputs and emits the cluster number with the highest sum of entries for each sequence. This value is then fed back as the new "cluster assignment" for the sequence to the input layer. The above steps are repeated until no cluster assignment changes (generally 5 to 6 rounds).

  It is possible to combine each of the above methods with a jackknife test. The user can set the number of replicates to perform (default=100) and the amount of data to disregard in each replicate (default=0.1). Two confidence values are calculated:
  - For each cluster - how often each cluster appears exactly the way it is in the replicates (the values are displayed in the Clusters window).
  - For each sequence - how often each sequence is assigned to the same cluster. The confidence values are displayed in the graph (black=low-confidence, red=high confidence).

  When the clustering is done, a window showing the clusters that were found is opened. Clicking on each cluster in this window, highlights the sequences that belong to the selected cluster in the 3D graph. Selecting a cluster and then clicking the Add to sequence groups button, defines the sequences belong to this cluster as a group that can be edited by selecting the Window -> Edit Groups option. Selecting all the sequences and clicking the Add each as separate sequence group button, defines

the sequences that belong to each cluster as a different group that can be edited in the Edit Groups window. Once the clusters are defined as separate sequence groups, this information will be written to a CLANS saved-file (inside a <seqgroups> block), including the numbers of the sequences composing each group (=cluster) and each group's attributes.

- Get sequence with hits from/to selected: this option should be selected after selecting sequences from the graph. Then, a window composed of two parts is opened. On the left side, displayed the names of the selected sequences. On the right side, the names of the sequences, having BLAST hits from/to the selected sequences. Clicking the OK button highlights the sequences from the right side of the window in the 3D graph. Selecting sequences (from both sides) and clicking the Set as selected button, sets these sequences as selected in the graph.

- Show selected sequences as text (copy/pastable): open a window displaying the selected sequences in FASTA format.

- Rotation: the rotation window enables to set a rotation angle for a continuous or discrete rotation around the X or Y axes.
    - X: set a rotation angle for a rotation around the Y axis.
    - Y: set a rotation angle for a rotation around the X axis.
    - Time (min. ms): define the time (in milliseconds) between each rotation, when checking the continuous rotate option (the lower the value, the faster the rotation).

Generating an image of CLANS map

Once the clustering process has reached its desired state, you would probably want to save the obtained CLANS map as a high-resolution image. Since CLANS has no built-in 'export to image format' function, it is required to make a screenshot. The following steps may help you generating a high-resolution image out of your CLANS map:

1. For the purpose of generating an image, it is recommended to perform the clustering in 2D instead of 3D (Misc -> Cluster in 2D).
2. Once you are satisfied with the clustering, set all the visual features (size and colors of dots and lines, groups features, etc.) and position the graph as desired.
3. Turn the anti-aliasing feature on (Draw -> Antialiasing) to make the graph look smoother.
4. Before you make a screenshot, adjust the dimensions of the CLANS window as desired and make sure that your display resolution is optimal (the resolution of a screenshot image depends on the display resolution).
5. Capture a screenshot of the CLANS map (in Windows: click on the CLANS window and then use ALT + PrtScn to copy it to the clipboard. In Mac: use command + shift + 4 and then drag the mouse to select the screen area to capture. The image is then saved on the desktop and can be copied to the clipboard and pasted to another program).
6. Paste the image from the clipboard to an image manipulation program like Photoshop or GIMP, and edit it as desired.