# CLANS Tutorial

Version 2.0.1

## Table of Contents



## Overview

CLANS 2.0 is a Python-based program for clustering sequences in the 2D or 3D space, based on their sequence similarities. CLANS visualizes the dynamic clustering process and enables the user to interactively control it and explore the cluster map in various ways.

The program implements a version of the Fruchterman-Reingold force directed graph layout, which uses the similarity scores to iteratively calculate attractive and repulsive forces between all pairs of sequences and move them in space accordingly. The better the score, the higher the attractive force.

The program was originally designed to cluster protein sequences based on their all-against-all sequence similarities, obtained by BLASTP (HSP E-values). However, CLANS can be generalized and applied to cluster and visualize any weighted network.

The cluster map can be saved as an image or as a file, which can later be loaded again by the CLANS software or by other network-visualizing software.

CLANS can be used in two modes:

- **GUI-based visualization tool** (default mode), which gets a matrix of sequence similarities and displays them as a dynamic graph using the Fruchterman-Reingold force-directed layout.
  In addition to clustering the sequences in space, the visualization tool enables to explore the data in various ways, which include manual interaction with the graph (rotation, panning, zoom-in and out, selection of data-points), different views of the data, several selection options, grouping and coloring the data (or a subset of it) by different features.

- **Command-line tool** (executed using the '-nogui' flag), which can be used to obtain a matrix of sequence similarities by running all-against-all BLAST search. In addition, it can run the Fruchterman-Reingold force-directed graph layout for a defined number of iterations and save the results in a clans-formatted file which can later be loaded and presented in the visualization tool. This is recommended for large datasets (>5000 sequences, depending on the computer resources), in which the clustering process is slow and there is no advantage in visualizing it.

  The BLAST search is only available in the command-line mode and requires installation of Blast+ on the target computer.

## Installation

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

From Anaconda repository:

1. Create a clean conda environment: *conda create -n clans_2_0*
2. Activate the newly created environment: *conda activate clans_2_0*
3. Install the clans package from Anaconda repository by using the following command:

   *conda install -c inbalpaz clans -c defaults -c conda-forge*

From source using conda:

1. Download CLANS latest release from: https://github.com/inbalpaz/CLANS/releases .
2. Extract the tar.gz file into the desired working-directory.
3. Create a new conda environment using the 'clans_2_0.yml' file (located in the root directory of CLANS) using the following command:

   *conda env create -f clans_2_0.yml*
4. Activate the newly created environment: *conda activate clans_2_0*

## Running CLANS in command-line mode

The command-line mode is executed using the '-nogui' flag option. It 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). It is also recommended in cases of large datasets, where the clustering can be done in the background and the resulted clans map can later be loaded and explored using the graphical interface.

**Usage:**

From within the activated clans_2_0 conda environment, type:

*python -m clans -nogui -infile <fasta_file_path> -saveto <destination_file_path> [options]*
or
*python -m clans -nogui -load <network_file_path> -dorounds <number of iterations> -saveto <destination_file_path> [options]*

Mandatory arguments:

-nogui: Run in command-line mode (no graphical interface)

-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'/'mini-clans'/'delimited'>: The format of the network output file (default is 'clans' format).

--debug: Run in debug mode

**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 arguments:**

-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 'score' (0-1 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:**

Within the activated clans_2_0 conda environment, type:

*python -m clans [*-load <network file path>*] [options]*

When clans is executed without an input-file, the GUI is opened empty and an input-file can be loaded from the 'File' menu.

Optional arguments:

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

-input_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 'score' (0-1 values), the sequences with scores above the threshold are considered connected. The default in this case is 0.1.

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

--debug: Run in debug mode (prints debug output to the terminal)

## Input / Output file formats:

The minimal input for CLANS visualization tool is a network file, containing pairs of sequence identifiers and their 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 the CLANS web-utility in the MPI Bioinformatics Toolkit (https://toolkit.tuebingen.mpg.de/tools/clans) or saved in a previous session of the CLANS 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 ³ X,Y,Z £ 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 (scores) 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*
  *nodes_size=8*
  *nodes_color=0;0;0;255*
  *nodes_outline_color=0;0;0;255*
  *nodes_outline_width=0.5*
  *is_taxonomy_available=True*
  *found_taxa_number=415*
  *</param>*

- Groups block: a list of groups with their presentation definitions (size, color, font-size, etc.) and assigned sequences. In case of more than one defined grouping category, the list of groups should follow the category name.

  *<seqgroups>*
  *category=manual*
  *name=Proteobacteria*
  *size=10*
  *name_size=10*
  *color=255;0;0;255*
  *outline_color=0;0;0;255*
  *is_bold=True*
  *is_italic=False*
  *numbers=1;3;5;6;10;13;18;19;20;21;23;27;33;35;42;49;52;*
  *</seqgroups>*

- Metadata block: The values and color-range of the numeric parameters uploaded by the user in a previous session of the CLANS software (the values are separated by space).

  *<seqparams>*
  *Param=Param1*
  *min_color=255;255;0;255*
  *max_color=255;0;0;255*

*values=0.07 0.94 0.56 0.01 0.25 0.53*
*</seqparams>*

- Taxonomy block: The NCBI taxonomic classification that was obtained for the organisms in the dataset in a previous session of the CLANS software (taxonomic levels: Family, Order, Class, Phylum, Kingdom, Domain).

  *<taxonomy>*
  *tax_level=Domain*
  *name=Eukaryota*
  *numbers=1;2;3;4;5;6;7;8;9;10;11;12;13;14;15;16;17;18;19;20;21;22;23;24;25;26;27*
  *name=Bacteria*
  *numbers=43;44;46;48;50;53;55;87;89;90;91;92;93;95;96;97;100;115;116;117;129;436;437;439;440*
  *name=Archaea*
  *numbers=465;470;495;508;546;549;690;698;710;765;767;772;794;796;807*
  *name=Not assigned*
  *numbers=0;34;49;58;105;106;127;132;133;147;155;157;158;166;172;197;202;246;259;260*
  *</taxonomy>*

Tab-delimited format

This file type should contain 3-4 columns: a list of non-redundant sequence-pairs and their similarity scores (without a header). For example:

1. Using scores of type P-value (the lower the score, the stronger the connection):

   *Seq1 seq2 1e-10*

   *Seq1 seq3 1e-06*

2. Using scores between 0 and 1 (the higher the score, the stronger the connection):

   *Seq1 seq2 0.94 score*

   *Seq1 seq3 0.67 score*

   Mandatory columns:

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

   Optional column:

   - Type of score: ‘hsp’ for P-value and ’score’ for values between 0 and 1. When this column is omitted, the similarity score is considered as P-value type.

Save as image

The graph area of the current CLANS session can be saved as an image in one of the following formats: PNG, Tiff, Jpeg and EPS.

## 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: A file that was generated by any version of the CLANS program or by the MPI Bioinformatics Toolkit.
- Tab-delimited format.

- Save to file: Saves the current graph presentation into a file, either in 'CLANS' format or as tab-delimited file.
  - CLANS format options:
    - CLANS (version 2.x): Saves all the information required to restore the current session (including the sequences, coordinates, groups, metadata and running-parameters).
    - Legacy CLANS: compatible with older versions of the Java-CLANS software.
  - Tab-delimited format: it 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 session in one of the following formats: PNG, Tiff, Jpeg or EPS.
- Quit: Closes the application including all the open windows.

**Configure menu:**

- Layout parameters
  - 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).
- Data-points general settings: Sets the default size, color, outline-color and outline-width of the graph's data-points, when no other setting is defined. In case different values are set via the groups' definitions (each group can have a different setting), they will take over the default setting.
- Connections (edges) settings: Sets the color and the width of the connecting lines (edges).
  By default, the attraction values (scores) are divided into 5 bins (between 0 and 1) and the edges are colored accordingly in shades of gray (the higher the score-bin, the darker the color). It is possible, however to set a uniform color for all edges.
  The width of the edges can be set to a value between 1 and 5 (default is 1). By default, the width is uniform for all edges, but similarly to the color attribute, it can be set differently to each score-bin.

**Tools menu:**

- Group data by:
  - NCBI Taxonomy: This feature automatically divides the data into groups according to a taxonomic level of the user's choice. It can be applied when the input is a CLANS/FASTA formatted file and the sequences headers contain the organism names in one of the following formats:
    1. Inside square brackets, as obtained by the NCBI database.
    2. 'OS=organism_name' as obtained by the UniProt database.

    In other cases, the organism names cannot be extracted.
    The taxonomic information is taken from the NCBI Taxonomy database.

    **Usage:** The first time this feature is applied on a certain dataset, a taxonomic mapping of the organisms that are found in the input file against the Taxonomy database is performed. This process may take a while. Then, a dialog opens, allowing the user to select the taxonomic level to group according to in addition to other groups-related size and font parameters. Once the taxonomic information for a dataset was collected, the user can easily change the selection of the taxonomic level by opening this dialog again without having to wait (Tools -> Group data by -> Taxonomy). Once a certain taxonomic level was selected, it is added to the 'Group by' combo-box, which switches between different grouping options. The colors of the groups are generated automatically according to the number of groups within each taxonomic level. However, once the groups are generated, it is possible to edit them manually and change their parameters, such as color.

The sequences, for which taxonomic information could not be extracted, or a specific taxonomic level is missing, are grouped together as 'Not assigned'.

- Add custom grouping category: This option allows the user to upload a tab-delimited metadata file (see details below), with one or more grouping-categories, by which the data can be grouped. The new grouping category is added to the 'Group by' combo-box, which switches between different grouping options.
  The colors of the groups are generated automatically according to the number of groups within each category. However, once the groups are generated, it is possible to edit them manually and change their parameters, such as color via the 'Edit groups' dialog. If there are sequences, for which no group was assigned, they are grouped together as 'Not assigned'.

  Metadata file format: Tab-delimited file, containing at least two columns and a header:

  | ***Sequence ID*** | ***Grouping category*** |
  |---|---|
  | *0* | *group1* |
  | *1* | *group2* |

  **Sequence ID**: There are two options to provide a unique sequence ID:
  1. The sequence name as it is given in the input file. When the input was provided in CLANS/FASTA format, the sequence ID is the sequence header.
  2. The sequence serial number according to the order in the FASTA section of the clans file (starting by 0).

  Note that the number and identity of the sequences in the metadata file must match the sequences given in the input network file. Missing data can be filled as 'NA'.

**Note**: Once the grouping by either taxonomy or user-defined parameter action is performed, the 'Group by' combo-box is enabled and it is easily possible to switch between the newly added grouping options and the 'Manual definition' default grouping option (where groups are manually defined by selecting sequences/data-points).

- Color data by:
  - Sequence length: This feature colors the data-points according to the length of the sequences. The color-range can be set by clicking the colored 'change' buttons.
  - Add/Configure custom parameter: This option allows the user to upload a tab-delimited metadata file (see details below), with one or more numeric parameters, according to which the data can be colored. Once a metadata file is loaded, it is possible to set a different color-range for each parameter. An unlimited number of metadata files with different parameters can be uploaded and the information is added to the already existing user-defined parameters.

    Metadata file format: Tab-delimited file, containing at least two columns and a header:

    | ***Sequence ID*** | ***Param1*** | ***[Param2, Param3, …]*** |
    |---|---|---|
    | *0* | *30* | *0.22* |
    | *1* | *50* | *0.75* |

    **Sequence ID**: There are two options to provide a unique sequence ID:
    1. The sequence name as it is given in the input file. When the input was provided in CLANS/FASTA format, the sequence ID is the sequence header.
    2. The sequence serial number according to the order in the FASTA section of the clans file (starting by 0).

    **Parameter**: Int / float number.

    Note that the number and identity of the sequences in the metadata file must match the sequences given in the input network file. Missing data can be filled as 'NA'.

**Note**: Once the coloring by either sequence length or user-defined parameter action is performed, the 'Color by' combo-box is enabled and it is easily possible to switch between the newly added coloring options and the coloring by groups (or default color in case there are no defined groups).

## Interaction with the graph area

CLANS has three distinct interaction modes within the graph area:

**Rotate/Pan graph 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.

**Select data-points mode:**

Allows to manually select specific data-points from the graph, by using the mouse:

- To select a specific data-point, locate the mouse cursor on that 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.

The selection can be done in two modes:

- Sequences mode: Clicking on a data-point selects the specific sequence(s) within a small radius around the clicked point. (Since points may overlap each other, several sequences may be selected).
- Groups mode: Clicking on a point selects all the data-points that belong to the same group as the clicked point. When selecting an area by mouse-dragging, all the data-points that belong to groups within that area will be selected. This mode is available when the currently presented grouping-category contains at least one defined group.

**Move/Edit text mode:**

This mode allows to use the mouse in order to change the location of the group names text elements in the graph area or to edit them.

- Left button click on a group name + mouse drag: drags the text to the desired position in the scene.
- Double-click on a group name opens a dialog in which it is possible to edit the text (name, size, color, etc.). Please note that changing the color of the group name will also change the color of the data-points belong to that group.

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

## GUI Controls

**Clustering options:**

- Initialize: generates random coordinates to all the data-points.
- Start: starts the iterative clustering process. The number of iterations that were done is displayed below the graph area.
- Stop: 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 / Score 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 a value between 0 and1, sequence-pairs with similarity score above this threshold are considered connected.

**Interaction mode:**

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

- Selection mode: switches between the selection of specific data-points (sequences) and the selection of whole groups. In sequences selection mode, the points that are clicked or within the selected area, are being selected. In groups selection mode, when a point is clicked (or within the selected area), all the points that belong to the same group are being selected.
  This combo-box is only enabled in 'Select data-points' interaction mode and in case the currently presented grouping-category contains at least one defined group.

**Color by option:**

Determines by which type of parameter the data is colored (groups or some numeric parameter). By default, the data is colored by groups, defied in the first grouping-category (if there are no pre-defined groups, the data is colored in black and set to the default grouping category 'Manual definition').
The color-by combo-box is enabled when at least one numeric parameter was added by the user (using the 'Tools -> color data by' action or saved in the input CLANS file in a previous session). It is possible to switch between coloring the data by the defined groups (or default color when no group is defined) or by any of the numeric parameters, including sequence length. When each parameter has a different color-range setting, the data is colored according to the selected parameter's setting.

**View 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).
- Auto Z-index / By groups order: 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.
- Full / selected dataset view: switches between the full dataset presentation (default) to a presentation of the selected subset only. When the selected subset is displayed, any operation that is done is performed on the subset only. For example, it is possible to start the clustering considering the subset data-points only. When switching back to full dataset, 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 'selected subset' mode. The 'Selected subset' viewing mode is only enabled when there are at least 5 selected data-points.
- Hide singeltons: hides data-points which have no connections under the current p-value/score threshold.

**Display options:**

- Connections: when this button is checked, the connecting lines (edges) are presented in the graph**.**
  In order to display the edges "behind" the data-points, switch to a 2D view mode.
  By default, the edges are colored in 5 shades of gray according to the normalized attraction values (scores), which are divided into 5 bins. The higher the score, the darker the color. It is possible to change these colors and the width of the edges (for all the edges or per-bin) using the 'Connections (edges) settings' under the 'Configure' menu item.
- 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).

- Group names: When this button is checked, a list of the group names is presented on the left side of the graph area.
    - All / Selected: By default, when the 'Group names' button is checked, all the group names are presented. If there is at least one selected group, it is possible to display the selected group names only.
    - Init names positions: Clicking on this button brings the group names back to the top-left corner of the scene. It is useful after rotating the graph, zooming-in/out, etc.

**Selection options:**

CLANS enables the user to select a subset of data-points and perform various operations on this subset (for example: viewing and clustering the subset only, creating a new group from the selected subset, etc.). The selection can be done either manually, by changing the interaction mode to 'Select data-points' and mark specific points or an area of the graph. Or, it can be done by using the following GUI buttons. In all cases, the selected data-points are marked in the graph by having a bigger size and a magenta outline color.

- Select all: selects all the data-points.
- Clear: clears the selection.
- Select by text: this option enables to select sequences by searching text that appears in the sequence names/headers. Clicking on this button opens a find dialog, in which a search term can be entered. The results are displayed in the 'Select by text results' window (see detailed explanation in the 'Windows' section).
- Select by groups: this option is available when there is at least one defined group. It enables the user to select sequences according to their group(s) classification. Clicking on this button opens the 'Select by groups' window, presenting a list of all the defined grouping-categories and within each category, a list of all its groups. The user can select one or more groups from the desired categories and get the sequences that match the selected groups intersection (the intersection is done between the categories). Clicking on the 'Get sequences by groups intersection' button opens the 'Select by groups results' window, which presents a list of the sequences that meet the groups-intersection condition. These sequences can then be set/added to the selected subset (see detailed explanation in the 'Windows' section).

  Usage example: having two grouping-categories, taxonomic classification in the domain level and protein function. Using this option, it is possible to select only the sequences that meet the following condition: (Archaea OR Bacteria) AND Photosynthetic proteins.

- Edit selected sequences: This button is enabled only 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).

**Groups options:**

In order to help the user to navigate in the clustering map and examine whether the clustering meets other classifications, CLANS allows to define groups of sequences/data-points, which can be displayed in the graph by different colors. It is possible to define an unlimited number of grouping categories. The number of groups within each category is limited to 300. There are several ways to define groups:

1. **Manual definition** (default option): the groups can be manually defined by selecting data-points and adding them to a new or existing group or use the 'select by name' option.
2. **Input file**: The groups can be pre-defined in the CLANS input-file, using the <seqgroups> block.
3. **Taxonomy**: Use the Tools -> Group data by -> Taxonomy feature to automatically group the sequences according to the NCBI taxonomic classification. Each taxonomic level is a grouping category containing groups of sequences.
4. **User-defined parameter**: Upload a tab-delimited file with one or more parameter, by which the data is grouped.

Groups Controls:

- Group by: Determines by which grouping category the data is grouped. This combo-box is enabled once there is at least one additional grouping category other than 'Manual definition'. (The additional categories can be added by methods no. 2-4 described above). It is then possible to switch between all the defined grouping categories. When selecting a specific 'group by' category, all the groups-related controls refer to the selected category (for example: the 'edit groups' dialog will present the groups that belong to the selected group-by category).

- Edit categories: Opens a dialog presenting a list of all the added grouping-categories (not including the 'Manual definition' default category). When selecting a category, the following options are possible:
    - Edit category: Opens another dialog, in which it is possible to change the category's name and set the following parameters for all the groups in this category: data-points size, outline color, outline width, group-name text size, bold and italic states.
    - Move up: Moves this category one step up in the order of the categories list.
    - Move down: Moves this category one step down in the order of the categories list.
    - Delete category: Deletes this grouping-category.

- 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's name and set the following parameters for the specific group: the size, bold and italic states of the group-name, the size, color and outline-color of the data-points that belong to this group. 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 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 from group(s): Removes the selected data-points from their group(s) if they were assigned to any group. They get the default presentation.

## 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.

**Select by text results window:**

This window presents the sequences that their names or headers contain 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.

**Select by groups results window:**

This window presents the sequences that meet the groups intersection condition that was selected by the user in the 'Select by groups' previous window.
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.

Note that all the changes in the selected subset are updated in the graph as well.