CluSeek User Manual

Version 1.0.1/1

CluSeek is a bioinformatics data mining tool designed for the localization and analysis of colocalizing groups of genes, such as biosynthetic gene clusters, operons etc. within genomic sequences uploaded to the NCBI GenBank database. If you have one or several proteins and you wish to know where their coding sequences colocalize across different genomic sequences, and what other proteins may be associated with them, CluSeek is the tool for you!

CluSeek is a graphical application, currently available for Windows and Linux. MacOS is currently not supported and will require some improvisation not detailed in this user manual.

Table of Contents

1. Quickstart

Installation:

1. Create a folder for CluSeek (on your desktop for instance).
2. Download CluSeek from https://cluseek.com and extract the .zip archive to the folder for CluSeek.
   Note: CluSeek requires USEARCH for aligning amino acid sequences. CluSeek can download USEARCH by itself. Should that fail, you may download it manually here.
3. Open x86_redistributables file in the CluSeek folder and confirm its installation.
4. Open CluSeek in the CluSeek folder to begin using CluSeek!

Note: *A data folder is created automatically in the CluSeek folder during the first use and is used to store previously downloaded information. A new one is created automatically if a pre-existing one is not detected. If you wish to purge CluSeek's memory, simply delete the data folder.*

Python users - refer to chapter 2.2.

Colocalization:

1. Cluseek relies on BLASTP to identify the initial marker protein homologs, the colocalization of which is subsequently checked. CluSeek can accept (1) amino acid sequences as **FASTA** or **NCBI accession codes**, which are then used to queue a BLASTP search on NCBI servers, or (2) previously completed BLASTP searches saved in the **XML** format. You can read section 3.2 for details. If you are using FASTA as input, make sure to give a short recognizable name to each sequence. **Previously completed searches are saved in the data folder and can be re-used as input.**
   In our experience, 3-4 markers should suffice for most searches, but this of course highly depends on what you are trying to find out.
2. Launch CluSeek and input your chosen marker proteins into the input fields. Use the three dots on the right of each input field to select .fasta files with amino acid sequences, or .xml files with previous sessions. After pressing "Load Files", CluSeek will queue the necessary BLAST searches and download positional data for each BLAST result. As NCBI deprioritizes searches submitted programmatically, large searches may take hours to complete (small queries should still finish within 15-30 minutes, depending on server load). See section 4 for more.
3. In the "Filtering" tab, check the marker proteins that must be present in target regions (by default, all are checked), then press "Search!". You can select "View" below to review the results in more detail. See section 5.1 for more.

Neighborhood View:

4. If you wish to see other neighborhood proteins found in the regions of colocalization (target regions), click "Display Full Genetic Neighborhood" and then "Create". CluSeek will need several minutes to download the annotated genomic regions. See section 6.1 for more.
5. In the neighborhood view, each row corresponds to one target region, and each square corresponds to a protein coding sequence. You can left click on any protein to view information about the protein and its homologs, or right click for additional actions. See section 6.2 and later for details.

Important: Please keep in mind that the NCBI database is constantly undergoing changes, sequences are being renamed, retracted *et cetera*. It is therefore advisable to complete all necessary NCBI downloads on the same day. As of the 1.0 release, CluSeek stores downloaded information in the "data" folder which is automatically created in its working directory on first launch. You can delete this folder to start with a clean slate, or save it if you wish to preserve a given set of results.

2. Installation

CluSeek is currently available for either Windows or Linux. We'd like to provide a MacOS version eventually. If you are a Windows user, it is strongly recommended you try the portable version

If you'd still like to try installing CluSeek on MacOS, see the Python Package section below and read the caveats.

2.1 Portable version (Windows only) (Recommended)

If you are using a modern version of Windows (versions as low as 8 have been able to run CluSeek), installation should be relatively quick.

1. Download the latest version of the CluSeek portable executable
2. Download the USEARCH bioinformatics package from drive5.com
   It is likely you will also need to download and install Visual C redistributable DLL libraries necessary for USEARCH. If the installation file was not included in the CluSeek zip file, it may instead be obtained from https://aka.ms/vs/15/release/vc_redist.x86.exe
3. Place both USEARCH and CluSeek executables into your working folder and start CluSeek.

1. Download the latest version of the CluSeek portable executable
2. The archive should contain a Visual C redistributables installer, which will install DLL libraries required by USEARCH, a program that CluSeek relies on for sequence alignment.
   If the installer was not included, or you would prefer to download the installer directly from Microsoft, you may do so via https://aka.ms/vs/15/release/vc_redist.x86.exe.
3. You may now launch CluSeek and use it as normal.
   Note that before displaying neighborhoods, CluSeek will ask you where it can find USEARCH. The prompt includes an option to automatically download USEARCH if you do not have it on your computer – it is recommended to keep it into the same folder as CluSeek.
   Should the download fail, you may download USEARCH from drive5.com instead.

Note: *CluSeek will automatically create a data folder in its directory, in which it will store all information downloaded from NCBI servers. Deleting this folder is a simple way to force CluSeek to re-download previously obtained information.*

2.2 Python package (Windows & Linux)

Several notes to start:

It is recommended that only more experienced users should attempt this method of installation, as there are many more pitfalls, and it is impossible to cover all possible cases in this guide.

The barrier to using the MacOS version is not CluSeek itself, but its dependency, USEARCH. Only the 32 bit version of the software is freely available for personal use. MacOS versions beyond 10.15 do not support 32 bit applications. If you can overcome this hurdle, CluSeek should run just fine. The official USEARCH website has some suggestions on how to do that, mostly using virtualization, which you may want to try out.

Regardless, to run CluSeek as a Python package, the following steps should be taken:

1. Install Python 3.9 or older. CluSeek was originally developed for Python 3.9. If you already have a newer version of Python installed, it's likely CluSeek will work just fine, so long as all the necessary packages are available for this newer version.
2. Download USEARCH from drive5.com and, if necessary, install the Visual C redistributables required to run it.
3. Install the CluSeek package via pip using the command pip install CluSeek.
4. Navigate to your work directory and run CluSeek by entering CluSeek. This should launch the application. If your system has multiple python versions, you may need to call the correct python version with python -m CluSeek.

2.3 General notes

CluSeek uses a local database to store previously downloaded data. These are automatically created in the working directory in the folder "data". For the portable version, the working directory is the directory where the executable has been placed.
By deleting the data folder, you will force CluSeek to re-download all data anew. **Doing this is recommended when starting a new analysis.**

3. Marker proteins and BLAST configuration

This chapter contains general notes on the choice of marker proteins and configuration of NCBI BLAST to achieve the best results. **For a step-by-step guide on how to use CluSeek, see chapters 4, 5 and 6.**

3.1 Marker proteins and BLASTP

First, one must select several proteins the presence of which identifies the targeted colocalizing unit. These proteins and their sequence homologs, collectively referred to as **marker proteins**, will be used to find **target regions** which are likely to contain your desired **colocalizing unit**.

Generally, 2-4 proteins should be sufficient for this type of task, but more may be selected based on circumstances, for example if:

- You aren't sure about the composition of the colocalizing unit
- One or more proteins aren't always present in the target
- Your marker proteins are particularly common
- These proteins have known functional significance, and you want to have their presence or absence tracked from the start

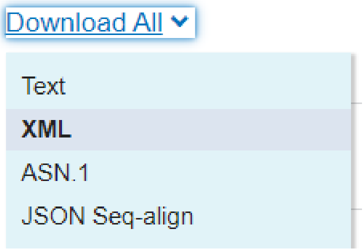
It is advisable to prioritize unique proteins with fewer sequence homologs found in the NCBI database -- mainly because the online BLAST interface provided by NCBI is limited to 5000 results.

3.2 Running BLASTP

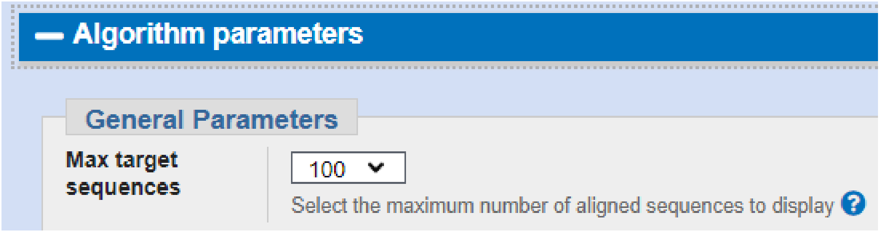
Once you have selected your initial set of proteins, it is necessary to obtain their homologs via NCBI's BLASTP.

There are two simple ways of running a BLASTP search:

1. Use NCBI's online BLAST interface at blast.ncbi.nlm.nih.gov and download the results as an XML file:



It is recommended to increase the the maximum number of returned sequences to 5000 first (it's all the way down):



Max target sequences 100

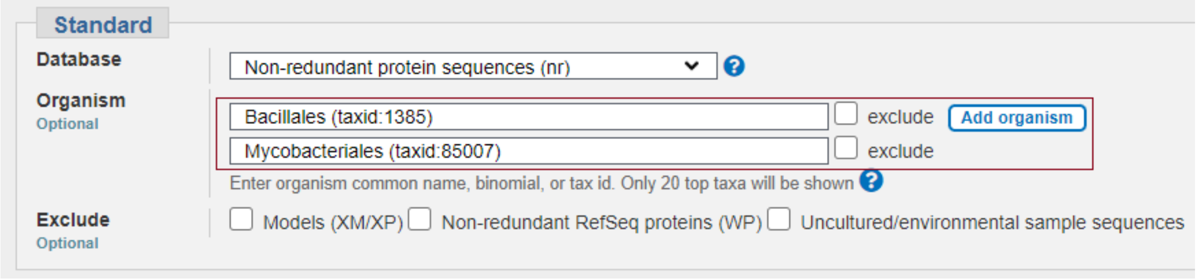
Select the maximum number of aligned sequences to display

2. Use CluSeek's built-in remote BLAST functionality by loading protein accessions or amino acid FASTA sequences, which will be automatically BLASTed when you select "Load Files"

Please note that while CluSeek does provide the option to increase the number of returned results above 5000, this is not recommended as it puts an increased strain on NCBI servers. The same applies to submitting large requests in excess of 50 proteins. Instead, consider limiting your search to a specific taxonomic group. NCBI recommends submitting large requests during weekends or low activity hours.

3.3 Restricting search to specific taxonomic group(s)

This is easiest done in the Standard section of the online BLAST interface:



Standard

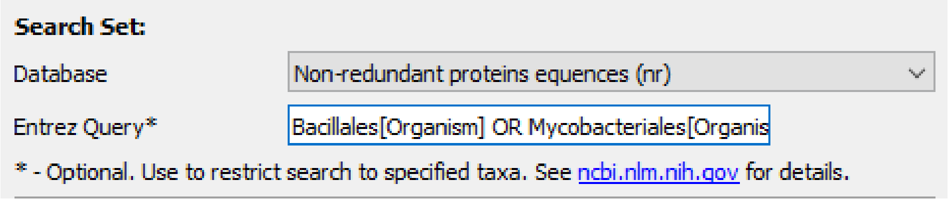
Database: Non-redundant protein sequences (nr)

Organism (Optional):
Bacillales (taxid:1385) □ exclude
Mycobacteriales (taxid:85007) □ exclude

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

Exclude (Optional): □ Models (XM/XP) □ Non-redundant RefSeq proteins (WP) □ Uncultured/environmental sample sequences

When using CluSeek's interface, this task is more complicated, as you have to construct your own Entrez query with the proper syntax:



Search Set:

Database: Non-redundant proteins equences (nr)

Entrez Query*: Bacillales[Organism] OR Mycobacteriales[Organis

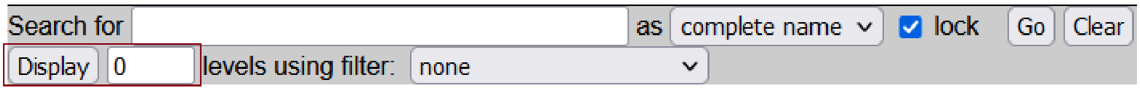
* - Optional. Use to restrict search to specified taxa. See ncbi.nlm.nih.gov for details.

To construct a valid Entrez query, you have to know the exact name of the taxon (eg. Bacillales), or its taxonomic identifier (eg. 1385) as CluSeek will not offer suggestions the way NCBI's interface does. A sample Entrez query may look like:

> (Bacillales[Organism] OR Mycobacteriales[Organism]) NOT (Bacillus[Organism] OR Corynebacterium[Organism])

The above query restricts the search to sequences belonging to members of the orders Bacillales and Mycobacteriales excluding the genera *Bacillus* and *Corynebacterium*. The [Organism] tag informs Entrez that the previous term is a taxon. You can find a longer explanation at ncbi.nlm.nih.gov.

To confirm the name of the taxon you are looking for, use NCBI's Taxonomy Browser to search for it in NCBI's taxonomic database.



Search for [ ] as complete name ☑ lock

Display 0 levels using filter: none

It is a good idea to display 0 levels, otherwise the browser will also display all subordinate taxa 3 levels down, which may take a while to load in some cases.

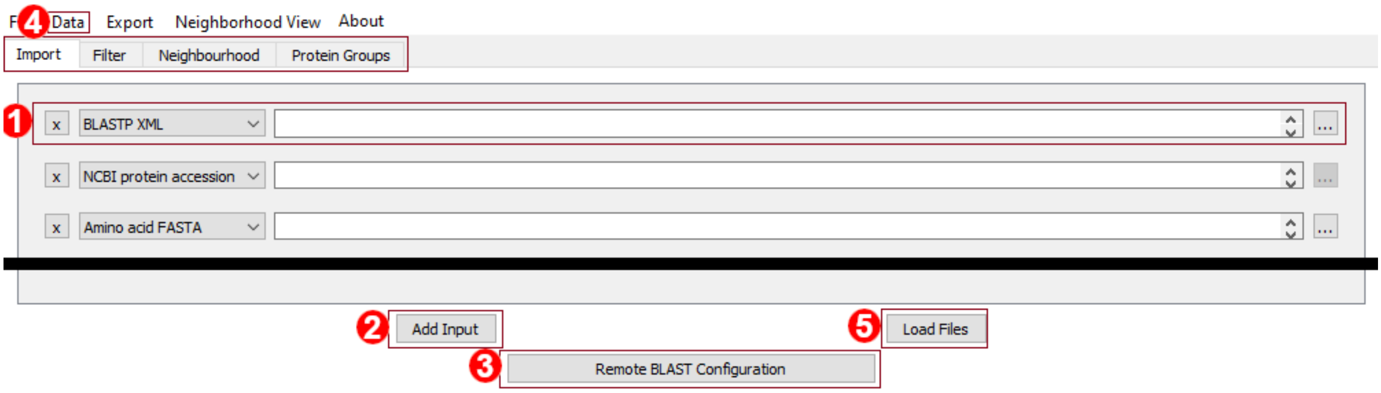
4. Importing data

This section covers the usage of the first (and simplest) screen. It will be the first screen you see once CluSeek has loaded.

Each CluSeek analysis begins by importing the marker proteins the colocalization of which you wish to study. CluSeek accepts input in the form of Amino acid FASTA, NCBI protein accession codes or BLASTP results XML files.

If a FASTA-formatted sequence or a protein accession code are used as input, CluSeek will automatically request NCBI servers to perform a remote BLASTP search to find all homologs of that sequence, which will take additional time to complete.

1. Each field corresponds to an input of a given type. Select which input you are trying to add. If adding XML or FASTA from a file, press the "..." button on the right to select the file. Note that as XML files are very large, only the path to the file will be loaded into the input box.
2. You can add new inputs by pressing "Add Input".
3. You can optionally configure the BLASTP search criteria using the "Remote Blast Configuration" button.
4. If you are performing a repeat analysis and want to ensure you are working with the same dataset as before, make sure you are only using XML files as inputs (as those are the only input type containing BLASTP results) and enable "Offline Mode". This will prevent CluSeek from downloading any missing information about proteins or sequences from NCBI (however, beware that remote BLAST searches will still be performed if non-XML inputs were provided).
5. Once all inputs have been selected and BLASTP configured, select "Load Files". CluSeek will now queue any necessary BLASTP searches, and download information about the genomic position of each protein BLASTP returned. This process may take some time.



Note: If you are entering FASTA sequences, make sure to name them suitably, as the header will be the only way of identifying them later.

Completed BLASTP searches are automatically saved in the data folder.

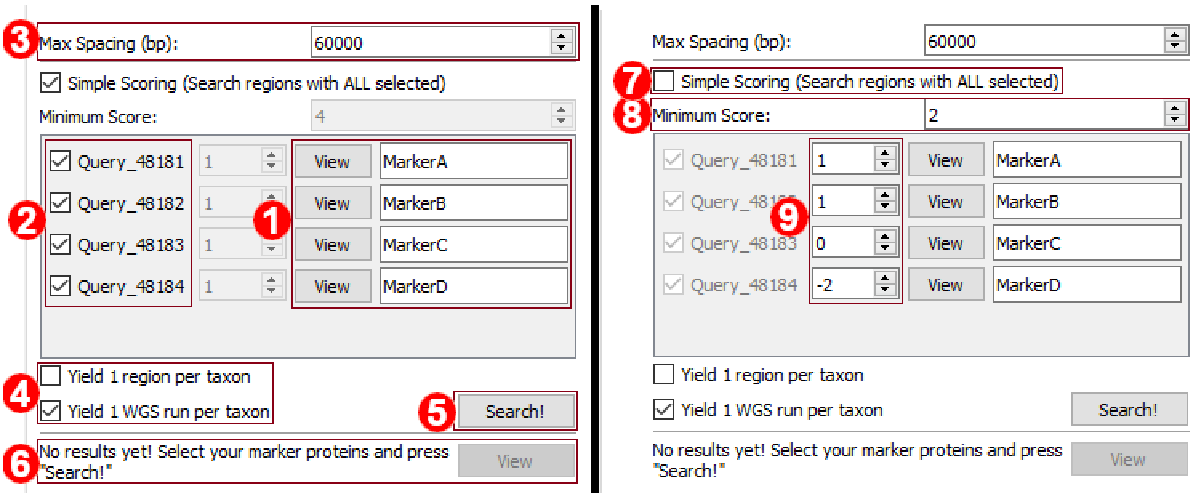
5. Colocalization and filtering

This section covers the second CluSeek tab, Filtering.

5.1 Marker protein list

The left hand side of the Filtering tab contains a list of all loaded marker proteins, as well as the settings for the colocalization algorithm itself. As before, each row corresponds to one marker protein.

1. First, rename your marker proteins to short identifiers that you want to be used from now on. Shorthand labels under 4 characters are recommended. The default text corresponds to either the BLASTed protein's description, or the header of its FASTA sequence. You may also press "View" to view the BLAST results graph for quality control.
2. Select which marker proteins you want to colocalize. The text displayed in this column is the sequence's unique identifier -- either a NCBI accession code, or an arbitrary sequence identifier assigned by BLAST.
3. Then select the maximum size of the region in which the selected marker proteins may be encoded to be identified as co-localized. Please note that if two regions which both individually pass the search criteria happen to overlap, they are merged into a single region which may exceed the maximum size. As of the current version, this behavior cannot be disabled.
4. Select one of the options to restrict the results to show either one result per NCBI Taxon (typically strain), or several results per taxon, so long as the contigs these results were found on belong to the same whole genome sequencing run.
5. Finally, press "Search!" to find regions in which proteins colocalize based on your criteria.
6. Below, a brief summary of the search results will appear. Pressing the "View" button next to the results summary will display a summary table of the results.
7. For a more customizable search, you may disable simple scoring.
8. When simple scoring is disabled, CluSeek searches for target regions whose score is at least equal to the minimum score value.
9. The presence of one or more instances of a marker protein is awarded with the set score value.You use set negative score values to exclude a given marker protein. Please note that if you set the maximum region size to be too small, there may be excluded proteins just outside the limits of found target regions. Therefore, if using negative scores, it is best to set greater maximum target region size.
   In the example below, a genomic region can only reach the threshold score value of 2 if it has both MarkerA and Marker B, and lacks MarkerD



Max Spacing (bp): 60000

☑ Simple Scoring (Search regions with ALL selected)

Minimum Score: 4

☑ Query_48181	1	View	MarkerA
☑ Query_48182	1	View	MarkerB
☑ Query_48183	1	View	MarkerC
☑ Query_48184	1	View	MarkerD

□ Yield 1 region per taxon

☑ Yield 1 WGS run per taxon

Search!

No results yet! Select your marker proteins and press "Search!" View

Max Spacing (bp): 60000

□ Simple Scoring (Search regions with ALL selected)

Minimum Score: 2

☑ Query_48181	1	View	MarkerA
☑ Query_481	1	View	MarkerB
☑ Query_48183	0	View	MarkerC
☑ Query_48184	-2	View	MarkerD

□ Yield 1 region per taxon

☑ Yield 1 WGS run per taxon

Search!

No results yet! Select your marker proteins and press "Search!" View

Note: *Un-checking a marker protein (or setting its score to 0) is not equivalent to not loading a marker protein at all. The presence of other loaded marker proteins is checked by the search algorithm, and they are incorporated into target regions, increasing their size.*

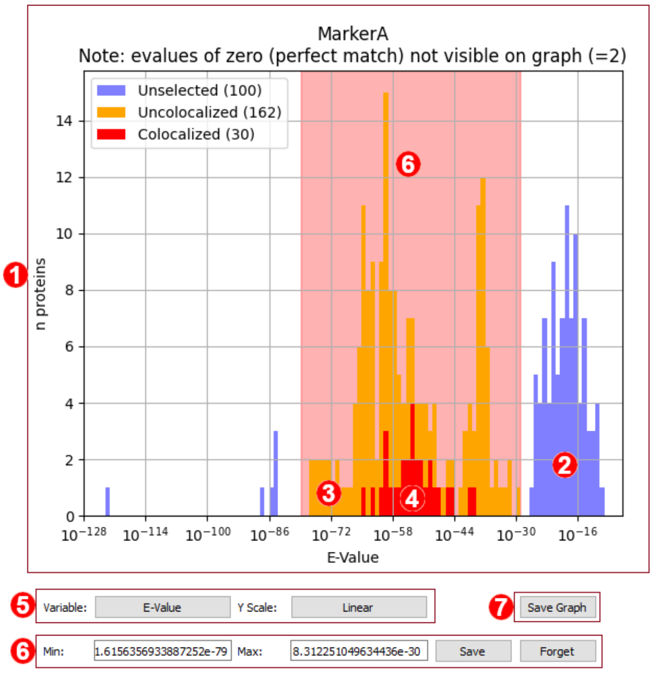
5.2 BLAST results graph

The BLAST results graph summarizes a single BLAST search, and shows which of the BLAST results were co-localized (found in target regions fitting the search criteria). There is a separate graph for each marker protein, each of which can be accessed from the marker protein list.

Please note that due to the dereplication of identical proteins in the GenBank nr database, one protein returned by BLAST may correspond to multiple coding sequences across multiple sequences or even genomes. In other words, while jarring, **the numbers of colocalized proteins or found target regions will likely not correspond. This does not necessarily indicate an error on part of CluSeek.**

1. The X axis displays an alignment statistic such as E-Value or the percentage of identical amino-acids between the query and hit sequence. The Y axis displays how many proteins (alignments) had a given value of the displayed statistic. Please note that E-values of 0 (which indicate an identical sequence) cannot be displayed on the logarithmic graph and are made note of above the graph .
2. Blue indicates proteins the colocalization of which was not checked. At first, all proteins are displayed as blue until you initiate the first search.
3. Orange indicates markers which were not colocalized with other marker proteins.
4. Red indicates proteins which are colocalized with other selected marker proteins.
5. To switch between variables and logarithmic/linear Y scale, simply click the relevant button to cycle through the options.
6. Although it typically isn't required, you may limit your search to a smaller subset of the viewed homologs if necessary. Either type the desired minimum and maximum values into the corresponding fields at the bottom OR hold click and drag across the graph to select the desired area, THEN press "Save" to save your selection. You can restrict the data based on multiple variables simultaneously. Pressing "Forget" removes the restriction for the current variable.

7. The graph in its current form can be saved using the "Save Graph" button. Unfortunately, it is not currently possible to scale the graph to a bigger resolution.



Variable: E-Value Y Scale: Linear

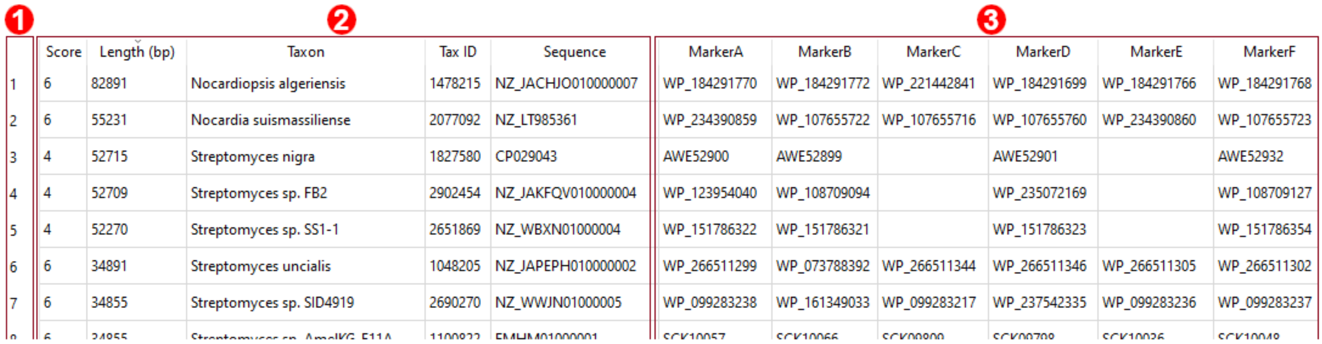
Min: 1.6156356933887252e-79 Max: 8.312251049634436e-30

5.3 Results summary table

The summary table shows colocalized markers identified by the CluSeek, as well as additional information about each one.

The table may be sorted by clicking once or twice on the header of each respective column. Width of each column and row can be adjusted individually by the user.

1. The left hand column simply contains the number of each row.
2. The following columns contain information about each target region: The score earned (only varies if advanced scoring mode is used in prior steps), the length of the target region, the scientific name and taxonomic ID of host species/strain, and the accession code of the nucleotide sequence in which the target region is found. Internal length denotes the number of base pairs between the first and last coding sequence. A negative internal length denotes an overlap.
3. The final columns indicate the presence/absence of a given marker protein, along with the accession codes of each protein. If multiple marker proteins of the same type are present in the sequence, all accession codes are displayed. The results summary table may be exported in the Data submenu at the top of the application's window.



	Score	Length (bp)	Taxon	Tax ID	Sequence	MarkerA	MarkerB	MarkerC	MarkerD	MarkerE	MarkerF
1	6	82891	Nocardiopsis algeriensis	1478215	NZ_JACHJO010000007	WP_184291770	WP_184291772	WP_221442841	WP_184291699	WP_184291766	WP_184291768
2	6	55231	Nocardia suismassiliense	2077092	NZ_LT985361	WP_234390859	WP_107655722	WP_107655716	WP_107655760	WP_234390860	WP_107655723
3	4	52715	Streptomyces nigra	1827580	CP029043	AWE52900	AWE52899		AWE52901		AWE52932
4	4	52709	Streptomyces sp. FB2	2902454	NZ_JAKFQV010000004	WP_123954040	WP_108709094		WP_235072169		WP_108709127
5	4	52270	Streptomyces sp. SS1-1	2651869	NZ_WBXN01000004	WP_151786322	WP_151786321		WP_151786323		WP_151786354
6	6	34891	Streptomyces uncialis	1048205	NZ_JAPEPH010000002	WP_266511299	WP_073788392	WP_266511344	WP_266511346	WP_266511305	WP_266511302
7	6	34855	Streptomyces sp. SID4919	2690270	NZ_WWJN01000005	WP_099283238	WP_161349033	WP_099283217	WP_237542335	WP_099283236	WP_099283237
8	6	34855	Streptomyces sp. AmelKG-E11A	1100822	FMHM01000001	SCK10057	SCK10066	SCK09800	SCK09708	SCK10036	SCK10048

6. Neighborhood view

The neighborhood view lets you view and compare the previously-found target regions by displaying **all** proteins encoded within these regions, not just your marker proteins. CluSeek downloads the relevant NCBI entries, and then performs a grouping of all proteins found in all target regions, dividing them into groups of similar proteins, allowing you to easily see similarities.

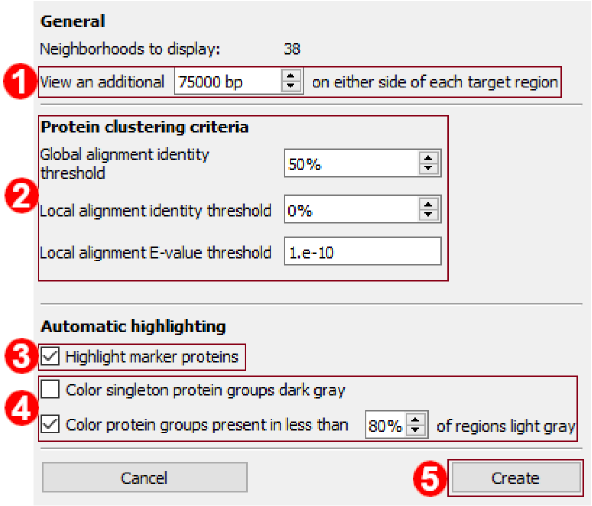
You can access the neighborhood view by pressing "Display Full Genetic Neighborhood" in the bottom right corner of the "Filter" tab.

6.1 Initial Configuration and Grouping

Before viewing all target regions, the neighborhood view must be configured. **For general use, it is safe to use the default settings and move on.**

1. First, you select how large an area around the target region you wish to view. If you are displaying a large number of target regions, it is recommended to set this to a lower value. Keep in mind, that the final viewed region size is increased by twice this number (if you add 75 000 bp on either side of a target region, the total increase is 150 000 bp)
2. Next, you may configure the protein grouping criteria. **You should not need to modify these settings, unless you believe there are issues with how proteins are being grouped.**
   The most common issue is that marker proteins of the same type (usually distantly related homologs) end up in multiple protein groups. If that is the case, you may try lowering the local alignment E-value threshold. For a more detailed explanation, see below.
3. You may select several automatic coloring options next. It is recommended to leave these options on.
   "Highlight marker proteins" will rename and give color labels to all protein groups which contain at least one marker protein. This means that proteins which weren't returned by the initial BLAST search (e.g. if the limit of 5000 homologs returned by BLAST is reached), but still have some sequence homology to a protein that was, will end up labeled this way. This may be relevant when BLAST search reached the limit of 5000 homologs

4. **"Color singleton protein groups dark gray"** does exactly what it advertises. If a protein group is only found in one gene cluster, it is painted dark gray.
   The next option is similar, but it selects proteins that are found in fewer than the specified percentage of gene clusters).
5. Finally, press "Create" to create the neighborhood view. The process may take some time as CluSeek may need to download sequence data from NCBI servers.



General

Neighborhoods to display: 38

View an additional 75000 bp on either side of each target region

Protein clustering criteria

Global alignment identity threshold 50%

Local alignment identity threshold 0%

Local alignment E-value threshold 1.e-10

Automatic highlighting

☑ Highlight marker proteins

□ Color singleton protein groups dark gray

☑ Color protein groups present in less than 80% of regions light gray

***Note**: The protein grouping process is performed in two steps: A global alignment-based grouping followed by local alignment-based grouping. **Global alignment-based grouping** creates homogenous clusters of highly similar proteins, while **local alignment-based grouping** can find far more distant homologies, but does not guarantee that all members of the same protein group will be similar within the whole amino acid sequence to one another. This can be a major issue with multi-domain proteins such as non-ribosomal peptide synthetases.*

6.2 Neighborhood view

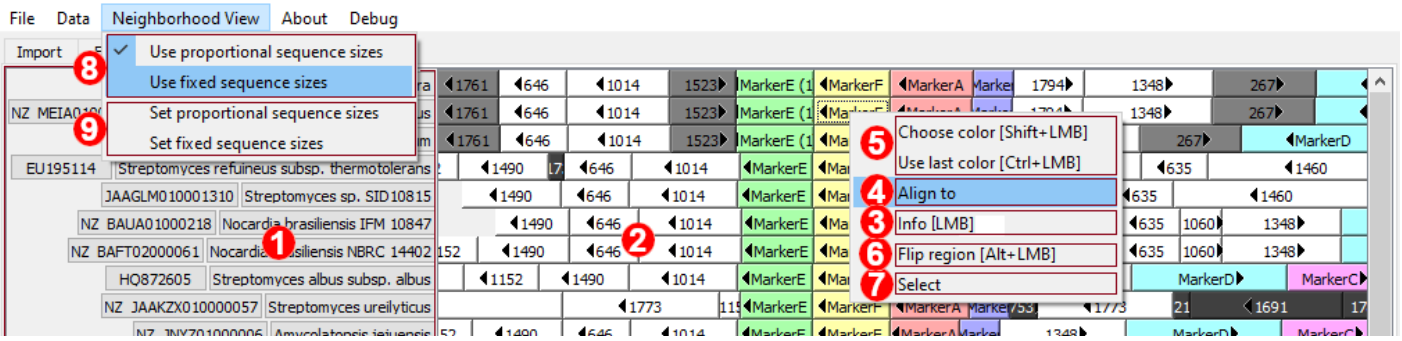
This section will cover the initial basic steps and explanations.

1. The left hand column contains the accession code of the sequence containing the target region, and name of the strain. Left click on a given name or identifier to display detailed information. See the information window. **If you do not see the opened detail information window, it is probably** overlayed by the main CluSeek window or minimized in the CluSeek icon in the bottom bar. Use the option “Window Pinned” in the upper left corner of the information window to ensure the **window is visible**.
2. The right side space contains a view of each target region. Individual blocks represent protein coding sequences of individual proteins. Left click on a given protein to display detailed information about the protein and its protein group. See the information window.

   All proteins are grouped into **protein groups** based on sequence similarity. The protein group a given protein belongs to is denoted by the text on the protein (typically a number) and its color.

Right clicking on a protein will open a contextual menu. The actions on this menu are executed with respect to the protein (group) which was right clicked. For details see steps 3 - 7 below.

3. Selecting the "Info" option, or simply left-clicking any protein will open the information window with additional information about the protein itself, as well as its protein group. Taxa and nucleotide sequences in the left hand column as well.
4. Selecting "Align to" will shift all displayed regions so that the selected encoded protein group becomes aligned in the center of the screen. If the protein group being aligned to is not present in a region, that region will move to the left. This option is recommended immediately at the beginning of the gene neighborhood analysis, because without it the gene clusters are not aligned and it is difficult to compare them.
5. You may color all proteins belonging to the same protein group to highlight similarities. If you attempt to color a protein group with the same color it already has, it will revert to white. (If this is unclear, Ctrl+Left click the same protein twice to see what happens). In combination with the indicated shortcuts, this can be used to quickly highlight and unhighlight a protein group you are interested in, to see where else it is found.
6. You may also flip an entire region to reverse its orientation. There is a known bug where CluSeek does not automatically re-align the region properly, i.e when the flip option is selected, you have to repeat the “Align to” option.



7. This option is used to select a protein group for later use. This is discussed further in the next section.
8. Left-click on the Neighborhood View button in the upper left corner of the window and select proportional or fixed view of the genes.
9. In the same way, you can set the size of the proportional or fixed view. Click on this option and a new small window will appear. If you prefer a larger gene size, set a smaller number inside the window.

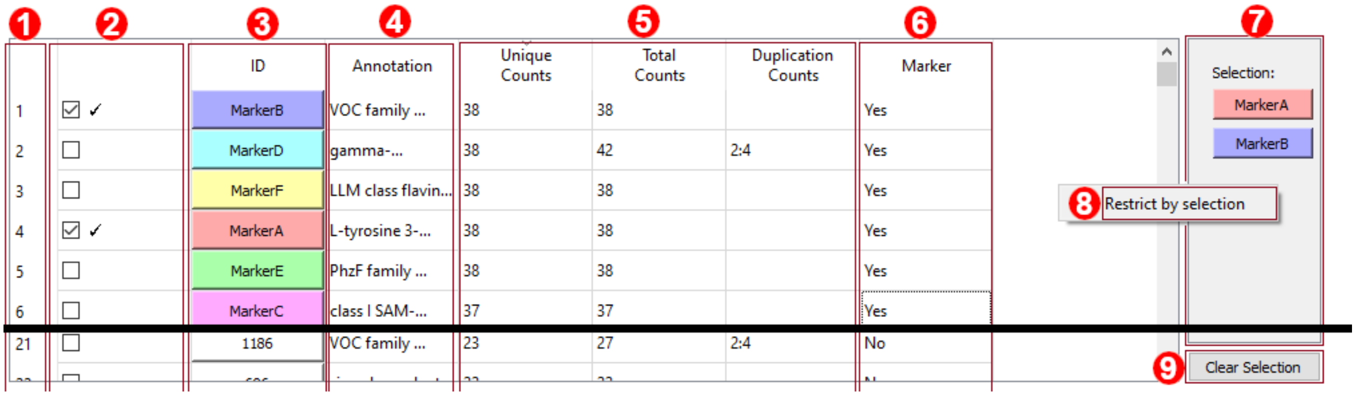
6.3 Protein groups table

To open it, left-click on the Protein Groups button in the upper left corner of the window. This tab provides a summary of all protein groups identified by the CluSeek. **You can sort by the contents of each column by clicking once (or twice) at the respective header, or change the width of a column by grabbing the edge of its header.**

1. The first column simply numbers each row from top.
2. The second column is used for selecting a protein group for later use. It is analogous to the "Select" option described in point 7 of the previous section
3. The third column displays the protein group button. Although no specific protein is depicted, you can interact with it (left-click, right-click) to view detailed information about the protein group, or to change its color/name. **If you do not see the opened detail information window, it is probably overlayed by the main CluSeek window or minimized in the CluSeek icon in the bottom bar. Use the option “Window Pinned” in the upper left corner of the information window to ensure the window is visible.**

   If you immediately do not see your markers in the third column, left-click at the header of the fifth column (Unique Counts). This can display your markers at the top of the column.
4. The fourth column shows the most common annotation (in GenBank) among member proteins of the protein group. You can view all annotations by clicking on the button in column 3. If you entered a custom annotation in the information window, it will be shown instead.
5. The next three columns depict (1) the number of regions in which this protein group is represented at least once (=Unique Counts), (2) The total number of proteins that are in this protein group (=Total Counts) and finally (3) The average number of duplications per region. The format is [number of duplications]:[Number of regions with this number of duplications]. If you look at the figure below, 2:4 means that four regions have two copies of MarkerD.
6. The last column simply indicates whether a given protein group contains at least one of the original marker proteins yielded by BLAST. As the clustering is independent of the initial search, it is possible for marker proteins to fall into separate protein groups. Sorting by this column will help you find them.
7. The section on the right contains protein groups selected either through the contextual menu (Point 7 of section 6.3) or by checking them in column 2 (see above).

8. Currently, there is only one action that can be performed with your selected protein groups: If you right click anywhere in the Protein Groups table, you may choose to "Restrict by selection", which will hide all gene clusters in the Neighborhood view that do not contain at least one member of EACH of your selected protein groups.
9. You can unselect ALL the currently selected proteins through the button below, or individually by left clicking them.



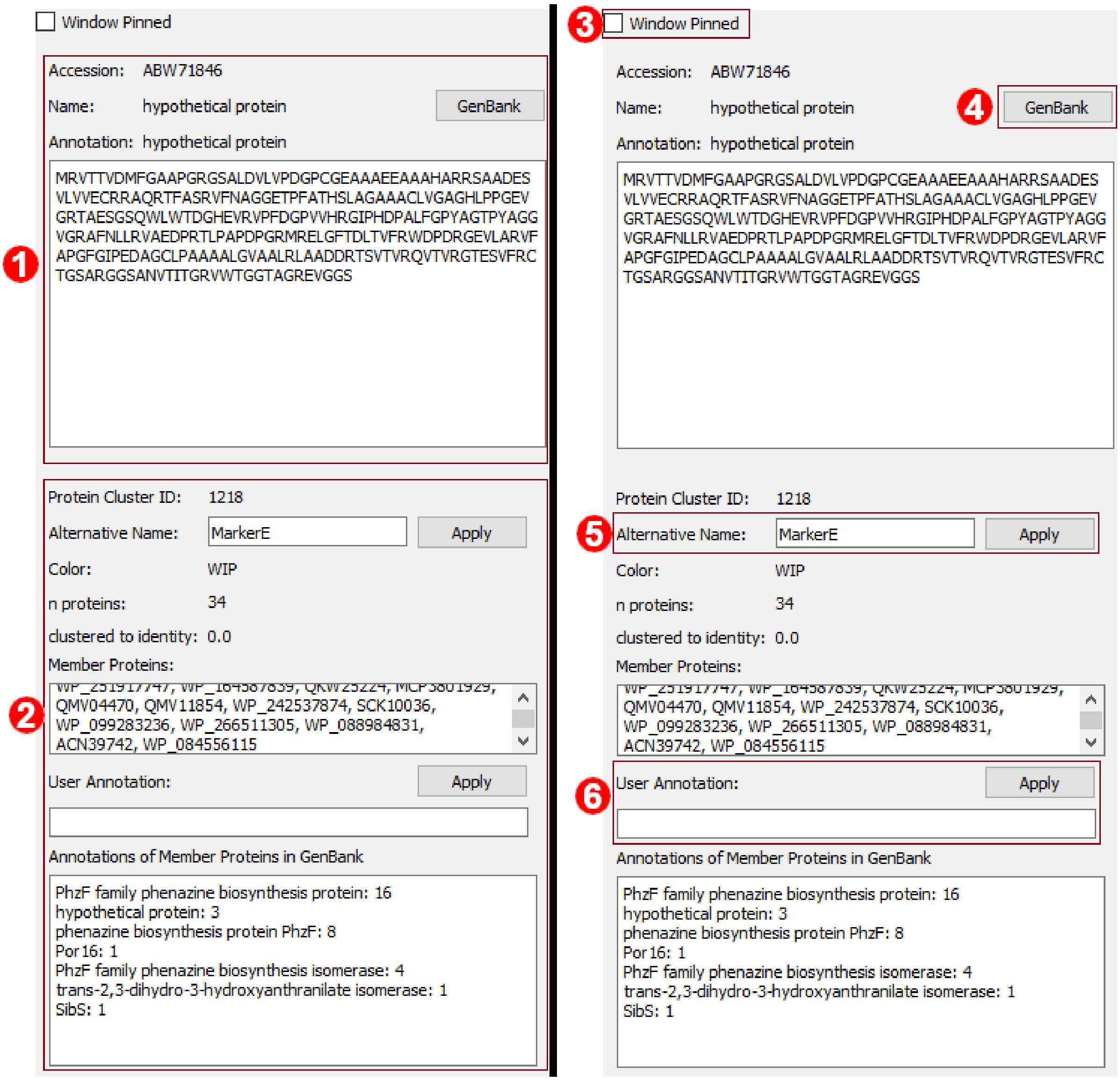
		ID	Annotation	Unique Counts	Total Counts	Duplication Counts	Marker
1	☑ ✓	MarkerB	VOC family ...	38	38		Yes
2	☐	MarkerD	gamma-...	38	42	2:4	Yes
3	☐	MarkerF	LLM class flavin...	38	38		Yes
4	☑ ✓	MarkerA	L-tyrosine 3-...	38	38		Yes
5	☐	MarkerE	PhzF family ...	38	38		Yes
6	☐	MarkerC	class I SAM-...	37	37		Yes
21	☐	1186	VOC family ...	23	27	2:4	No

Note: **While local alignment-based clustering can find more distant homologies, clusters constructed via local alignment can be heterogeneous as amino acid sequences are not aligned along their entire length. This is particularly the case for modular proteins such as non-ribosomal peptide synthetases or polyketide synthetases.**

6.4 Information window

The information window, as the name implies, displays all available information about a given element. Typically, multiple types of entries are associated with a given element. For example, if you click on a protein, you will be able to see...

1. Information about the protein itself,
2. but also information about the protein group it belongs to. You can click on proteins, sequences as well as taxa to display detailed information.
3. You can check "Window Pinned" to ensure the information window always stays on top, regardless of if you click away.
4. For many entries, you can press the "GenBank" button to open the corresponding entry in the NCBI database.
5. Protein groups may be renamed by filling in the "Alternative name" field and pressing "Apply".
6. Similarly, you may enter a custom annotation by filling in the "User annotation" field and pressing "Apply".



□ Window Pinned

Accession: ABW71846

Name: hypothetical protein

Annotation: hypothetical protein

```
MRVTTVDMFGAAPGRGSALDVLVPDGPCGEAAAEEAAAHARRSAADES
VLVVECRRAQRTFASRVFNAGGETPFATHSLAGAAACLVGAGHLPPGEV
GRTAESGSQWLWTDGHEVRVPFDGPVVHRGIPHDPALFGPYAGTPYAGG
VGRAFNLLRVAEDPRTLPAPDPGRMRELGFTDLTVFRWDPDRGEVLARVF
APGFGIPEDAGCLPAAAALGVAALRLAADDRTSVTVRQVTVRGTESVFRC
TGSARGGSANVTITGRVWTGGTAGREVGGS
```

Protein Cluster ID: 1218

Alternative Name: MarkerE

Color: WIP

n proteins: 34

clustered to identity: 0.0

Member Proteins:

WP_251917747, WP_184587839, QKW25224, MCP3801929,
QMV04470, QMV11854, WP_242537874, SCK10036,
WP_099283236, WP_266511305, WP_088984831,
ACN39742, WP_084556115

User Annotation:

Annotations of Member Proteins in GenBank

PhzF family phenazine biosynthesis protein: 16
hypothetical protein: 3
phenazine biosynthesis protein PhzF: 8
Por16: 1
PhzF family phenazine biosynthesis isomerase: 4
trans-2,3-dihydro-3-hydroxyanthranilate isomerase: 1
SibS: 1

□ Window Pinned

Accession: ABW71846

Name: hypothetical protein

Annotation: hypothetical protein

MRVTTVDMFGAAPGRGSALDVLVPDGPCGEAAAEEAAAHARRSAADES
VLVVECRRAQRTFASRVFNAGGETPFATHSLAGAAACLVGAGHLPPGEV
GRTAESGSQWLWTDGHEVRVPFDGPVVHRGIPHDPALFGPYAGTPYAGG
VGRAFNLLRVAEDPRTLPAPDPGRMRELGFTDLTVFRWDPDRGEVLARVF
APGFGIPEDAGCLPAAAALGVAALRLAADDRTSVTVRQVTVRGTESVFRC
TGSARGGSANVTITGRVWTGGTAGREVGGS

Protein Cluster ID: 1218

Alternative Name: MarkerE

Color: WIP

n proteins: 34

clustered to identity: 0.0

Member Proteins:

WP_251917747, WP_184587839, QKW25224, MCP3801929,
QMV04470, QMV11854, WP_242537874, SCK10036,
WP_099283236, WP_266511305, WP_088984831,
ACN39742, WP_084556115

User Annotation:

Annotations of Member Proteins in GenBank

PhzF family phenazine biosynthesis protein: 16
hypothetical protein: 3
phenazine biosynthesis protein PhzF: 8
Por16: 1
PhzF family phenazine biosynthesis isomerase: 4
trans-2,3-dihydro-3-hydroxyanthranilate isomerase: 1
SibS: 1