CluSeek 1.3.4 QuickStart Guide

Installation

1. Download CluSeek and extract the .zip file from our website
2. Run the x86_redistributables* installer included
3. That's it. Run CluSeek and start a project.

As this version uses regular project files, no elaborate folder management is required – just make sure to save before you exit.

** - These are libraries distributed for free by Microsoft, which are necessary for the running of Diamond, the sequence aligner used by CluSeek. They only need to be installed once. We provide the installer in the download archive for your convenience, but you may also obtain them directly from Microsoft if you wish.*

Running your analysis

CluSeek searches for sets of proteins whose coding sequences are co-localized in the genome.

1. Choose marker proteins

Choose a set of proteins in the FASTA format, whose colocalizations you wish to find. For example, these could be members of an operon, or they could encode a specific biosynthetic pathway in a secondary metabolite biosynthetic cluster.

A proper FASTA-formatted sequence should span exactly two lines with the first being a header preceded by the > symbol:

```
>sequence_header
amino_acid_sequence
```

For example:

```
>Apd1
MSSLEARRTDRTDLPLPAAGDWEYGGYPYGLEPLTLPLASPGSSAAHRRSDGSPPPWPGTWRT
PSPEFPANAAVDLTDPLGVDRLFWFRW
```

Every sequence MUST have a header that starts with >

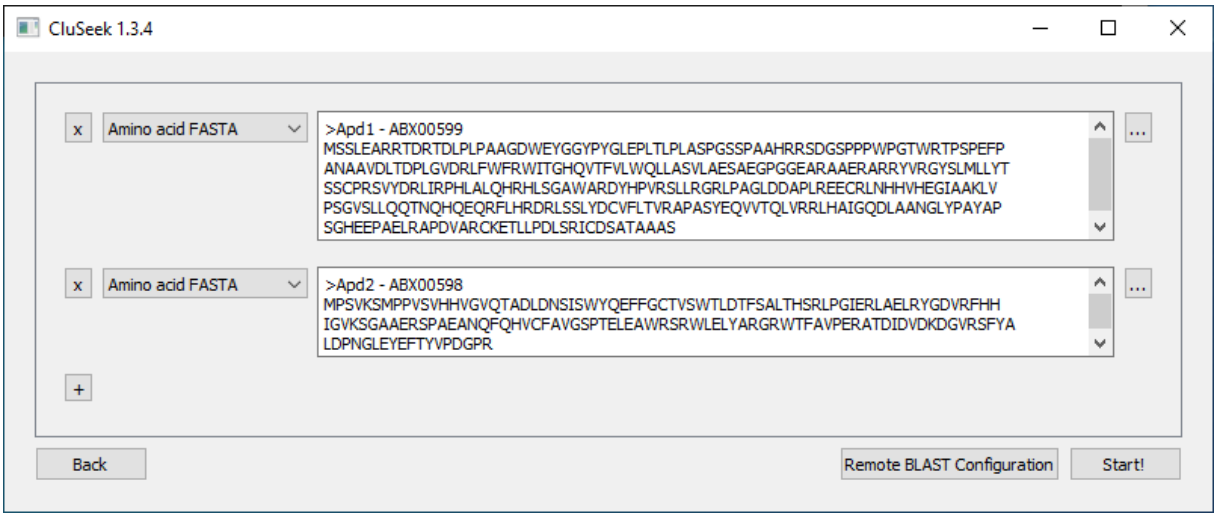
Some databases provide FASTA files that have line breaks in the middle of the sequence for formatting reasons. Currently, **CluSeek will complain if you try to input them, but they should work fine** as BLAST tolerates such inputs. They will look like this:

```
>Apd1
MSSLEARRTDRTDLPLPAAGD
WEYGGYPYGLEPLTLPLASPG
SSAAHRRSDGSPPPWPGTWRT
PSPEFPANAAVDLTDPLGVDR
LFWFRW
```

Note how the lines stop before the edge of the page. This is because there are additional line breaks have been inserted.

2. Input your marker sequences

This step should be self-explanatory. You can select other input types, for example you can run your own BLASTp search on the NCBI website, download it as an XML file, and input it here.

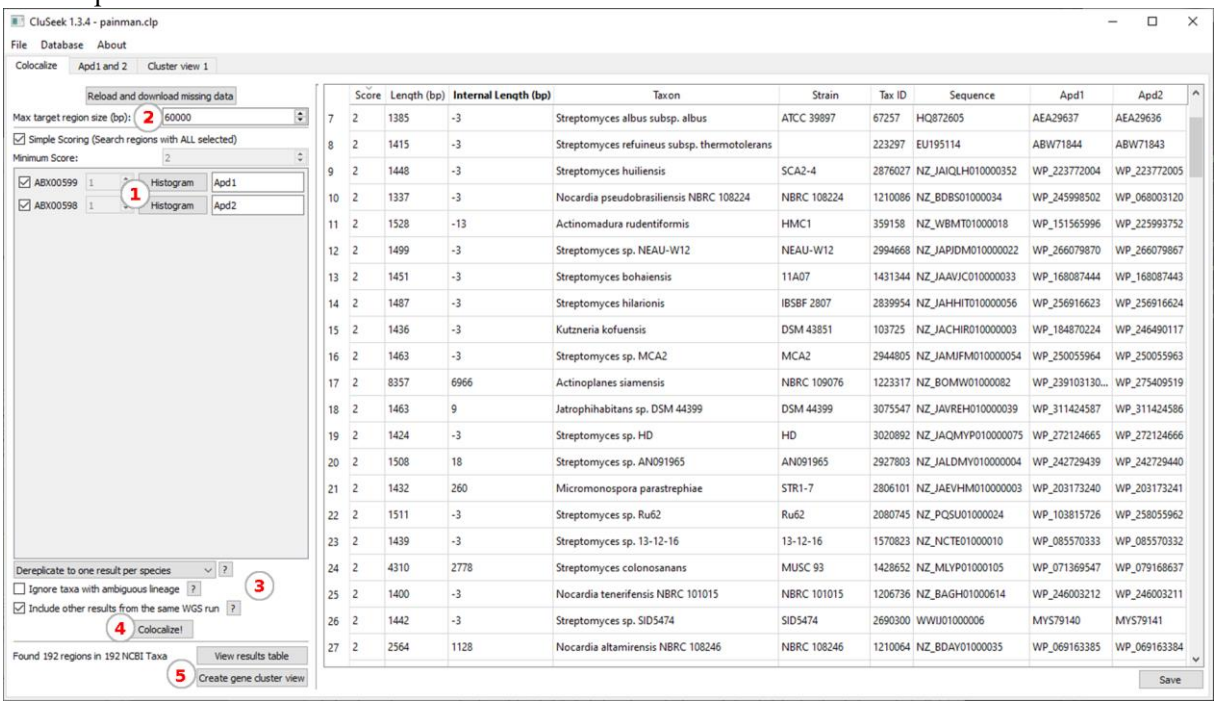


You can press the „+“ button to add additional inputs. You can press the „...“ button to load FASTA or XML files instead of inputting values manually.

When you are ready, press the „Start!“ Button. The processing requires CluSeek to query NCBI servers, so you need to be connected to the internet. Depending on how busy NCBI servers are, the process may take as little as 10 minutes or upwards of an hour. We’ve found the time of day can have quite a noticeable effect.

3. Colocalization

Now that your inputs have been loaded, you can start looking for co-localized combinations of your marker proteins.

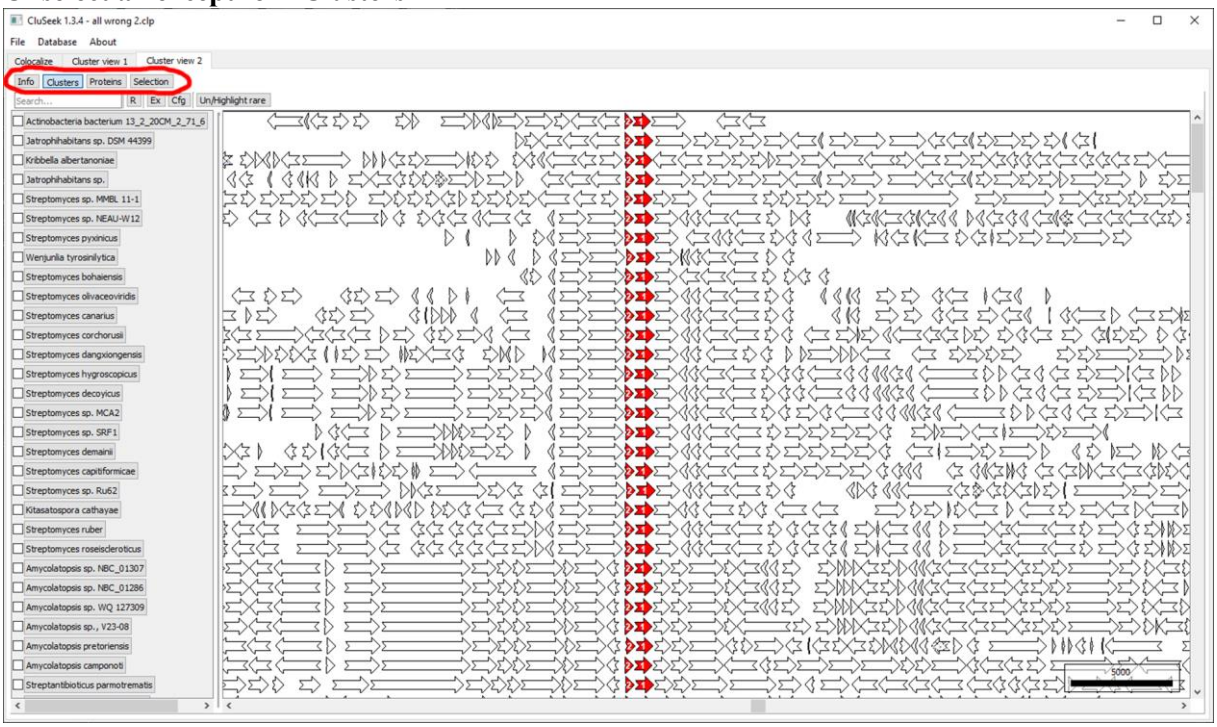


	Score	Length (bp)	Internal Length (bp)	Taxon	Strain	Tax ID	Sequence	Apd1	Apd2
7	2	1385	-3	Streptomyces albus subsp. albus	ATCC 39897	67257	HQ872605	AEA29637	AEA29636
8	2	1415	-3	Streptomyces refuineus subsp. thermotolerans		223297	EU195114	ABW71844	ABW71843
9	2	1448	-3	Streptomyces huiliensis	SCA2-4	2876027	NZ_JAIQLH010000352	WP_223772004	WP_223772005
10	2	1337	-3	Nocardia pseudobrasiliensis NBRC 108224	NBRC 108224	1210086	NZ_BDBS01000034	WP_245998502	WP_068003120
11	2	1528	-13	Actinomadura rudentiformis	HMC1	359158	NZ_WBMT01000018	WP_151565996	WP_225993752
12	2	1499	-3	Streptomyces sp. NEAU-W12	NEAU-W12	2994668	NZ_JAPJDM010000022	WP_266079870	WP_266079867
13	2	1451	-3	Streptomyces bohaiensis	11A07	1431344	NZ_JAAVJC010000033	WP_168087444	WP_168087443
14	2	1487	-3	Streptomyces hilarionis	IBSBF 2807	2839954	NZ_JAHHIT010000056	WP_256916623	WP_256916624
15	2	1436	-3	Kutzneria kofuensis	DSM 43851	103725	NZ_JACHIR010000003	WP_184870224	WP_246490117
16	2	1463	-3	Streptomyces sp. MCA2	MCA2	2944805	NZ_JAMJFM010000054	WP_250055964	WP_250055963
17	2	8357	6966	Actinoplanes siamensis	NBRC 109076	1223317	NZ_BOMW01000082	WP_239103130...	WP_275409519
18	2	1463	9	Jatrophihabitans sp. DSM 44399	DSM 44399	3075547	NZ_JAVREH010000039	WP_311424587	WP_311424586
19	2	1424	-3	Streptomyces sp. HD	HD	3020892	NZ_JAQMYP010000075	WP_272124665	WP_272124666
20	2	1508	18	Streptomyces sp. AN091965	AN091965	2927803	NZ_JALDMY010000004	WP_242729439	WP_242729440
21	2	1432	260	Micromonospora parastrephiae	STR1-7	2806101	NZ_JAEVHM010000003	WP_203173240	WP_203173241
22	2	1511	-3	Streptomyces sp. Ru62	Ru62	2080745	NZ_PQSU01000024	WP_103815726	WP_258055962
23	2	1439	-3	Streptomyces sp. 13-12-16	13-12-16	1570823	NZ_NCTE01000010	WP_085570333	WP_085570332
24	2	4310	2778	Streptomyces colonosanans	MUSC 93	1428652	NZ_MLYP01000105	WP_071369547	WP_079168637
25	2	1400	-3	Nocardia tenerifensis NBRC 101015	NBRC 101015	1206736	NZ_BAGH01000614	WP_246003212	WP_246003211
26	2	1442	-3	Streptomyces sp. SID5474	SID5474	2690300	WWIJ01000006	MYS79140	MYS79141
27	2	2564	1128	Nocardia altamirensis NBRC 108246	NBRC 108246	1210064	NZ_BDAY01000035	WP_069163385	WP_069163384

1. Here, you can see each of your inputs. Only inputs whose checkmark is checked will be co-localized, meaning that you can try different combinations without starting a new analysis. I recommend re-naming them to a short 5-letter identifier for later.
2. You can also set the maximum size of the region in which they can be co-localized at the top.
3. Depending on the type of analysis, you can select filtering and dereplication options.
    a. The first option will remove all but one result for a given taxonomic level. For instance, when dereplicating to one result per species, even if that species has dozens of strains that have been sequenced hundreds of times, you will only see one result.
    b. **Ignore taxa with ambiguous lineage** – When dereplicating by taxonomy (see above), there are some cases when only parts of the lineage are known. By default, if CluSeek is not able to identify the relevant taxonomy of a result, it will include it, as it cannot be sure if this is a novel sequence or not. **Checking this box will make CluSeek discard results with ambiguous classification instead.**
    c. **Include other results from the same WGS run** – When there are multiple results in the same sequencing run (indicating that this duplication has a biological basis), CluSeek will include all of them even while dereplicating the results (see a).
    For example, even while set to show only one result per species, if a given species has two of your target gene clusters in its genome, both will be shown.
4. Once everything is set, press co-localize. In the bottom left corner of the window, you will see how many results were found in total, and on the right (accessible via “View results table”) you will see a table of all results with additional information.
5. If you wish to see the cluster view, select “Create gene cluster view”. This process may take some time (our experience shows about 1 second per result), as CluSeek will have to download the genomic sequence of each result.
    We don’t recommend visualizing more than several hundred gene clusters – but this is primarily a hardware limitation. If you are confident in your hardware, you can always try.

4. Gene cluster view

When you first see your new cluster view**, do not panic**. In the top left, you will see four buttons. **Unselect all except for "Clusters"**.



This will leave you with the main gene cluster view. Each row corresponds to a single sequence. To the left, you see the names of the relevant species and sequences. Each arrow corresponds to a protein coding sequence. Red arrows are your marker proteins.

If you want to neatly align your view like the one in the picture, **right click one of the red arrows and select "Align to"**. In the right click menu, you can also change its color to something more your preference. Note that all proteins are grouped by similarity, **therefore changing the color of one protein will change the color of all similar proteins.** Unfortunately, protein grouping has some issues in the current version, and so you may see very large, heterogenous protein groups. We are working on this issue and hopefully a fixed version will be available very soon.

That is all for the basic pipeline. You can explore other parts of the user interface. Briefly:

- If you select the **Info** button again, you will be able to see information about any protein you left click, particularly its annotations and the annotations of other proteins in its group.
  You can also re-name protein groups here.
- If you select **Proteins**, you will see an overview table of all protein groups, their annotation and frequencies in your dataset.
- If you select **Selection**, you will be able to further filter your results based on the presence or absence of various protein groups.