

AMOEBAE command-line interface documentation

Lael D. Barlow

Version of January 5, 2022

Contents

1 Command reference

This documentation refers to the command-line interface for the "amoebae" python script (in the main repository directory), not the command-line interface for the AMOEBAE Snake-make workflow. Rules in the Snakemake workflow definition file (Snakefile) run the "amoebae" script with various parameters appropriate for relevant steps in the workflow. Documentation for each amoebae command and the various options may be accessed from the command line via the "-h" options. The following command reference information is the output of running amoebae (and each command) with the "-h" option.

1.1 amoebae

usage: amoebae <command> [<args>]

Commands for setting up data structure:

mkdatadir	Make a directory with subdirectories and CSV files for storing sequence data, etc.
-----------	--

Commands for similarity searching:

setup_hmddb	Construct an HMM database (with hmmpress).
add_to_dbs	Format and add a file to a formatted directory.
list_dbs	Print a list of all usable database files in the database directory as defined in the settings file.
add_to_queries	Add a query file to a formatted directory.
list_queries	Print a list of all usable query files in the query directory as defined in the settings file.
get_redun_hits	Run searches with queries to find redundant hits in databases (for interpreting results).
setup_fwd_srch	Make directory in which to perform forward searches.
run_fwd_srch	Perform searches with given queries into given dbs.
sum_fwd_srch	Append information about forward searches to csv summary file (this is used to organize reverse searches).
setup_rev_srch	Make a directory in which to perform reverse searches.
run_rev_srch	Perform searches with given forward search hits into given db.
sum_rev_srch	Append information about reverse searches to csv summary file.
interp_srchs	Interpret search results based on summary.
find_redun_seqs	Identify sequences likely encoded on redundant loci predicted for the same species.
plot	Plot search results.

Miscellaneous commands:

csv_to_fasta	Generate a fasta file from sequences detailed in a spreadsheet of similarity search results.
check_depend	Check that all the dependencies are properly installed and useable.
check_imports	Check that all the import statements used in the AMOEBAE

repository run without error.
regen_genome_info Write a new genome info spreadsheet file using filenames
from the Genomes directory.

positional arguments:

command Specify one of the functionalities of amoebae.

optional arguments:

-h, --help show this help message and exit

Copyright 2018 Lael D. Barlow Licensed under the Apache License, Version 2.0 (the "License"); you may not use this file except in compliance with the License. You may obtain a copy of the License at <http://www.apache.org/licenses/LICENSE-2.0> Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an "AS IS" BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.

1.2 amoebae mkdatadir

usage: amoebae [-h] new_dir_path

Make a directory with subdirectories and CSV files for storing sequence data, etc.

positional arguments:

new_dir_path Specify the full file path that you want the new directory to have.

optional arguments:

-h, --help show this help message and exit

1.3 amoebae setup_hmmdb

usage: amoebae [-h] indirpath

Construct an HMM database (with hmmpress). This is for later sorting of given sequences into categories based on which HMM the score highest against.

positional arguments:

indirpath Path to directory containing amino acid sequence alignment file(s) to be constructed into an HMM database using hmmpress from the HMMer3 software package.

optional arguments:

-h, --help show this help message and exit

1.4 amoebae add_to_dbs

```
usage: amoebae [-h] [--split_char SPLIT_CHAR] [--split_pos SPLIT_POS]
               [--skip_header_reformat] [--auto_extract_accs]
               new_file main_data_dir
```

Format and add a file to a formatted directory.

positional arguments:

new_file	Can be a fasta file (prot or nucl) or HMM databases, generated using the hmmpress program in the HMMer software package. Or a GFF3 annotation file.
main_data_dir	Path to main data directory (with Genomes, Queries, and Models subdirectories).

optional arguments:

-h, --help	show this help message and exit
--split_char SPLIT_CHAR	Character to split the header string on for extracting the accession. (default:)
--split_pos SPLIT_POS	Position that the accession will be in after splitting. (default: 0)
--skip_header_reformat	Skip reformatting of header lines in input fasta file. (default: False)
--auto_extract_accs	Automatically identify accessions/IDs in sequence headers (overrides split_char and split_pos options above). (default: False)

1.5 amoebae list_dbs

```
usage: amoebae [-h] main_data_dir
```

Print a list of all usable query files in the query directory as defined in a given AMOEBAE data directory.

positional arguments:

main_data_dir	Path to main data directory (with Genomes, Queries, and Models subdirectories).
---------------	---

optional arguments:

-h, --help show this help message and exit

1.6 amoebae add_to_queries

usage: amoebae [-h] query_file main_data_dir

Add a query file to a formatted directory. This command adds a given sequence file to the directory with the path that you have specified in the settings.py file, and appends a corresponding line to the CSV file that you specified (e.g., '0_query_info.csv') to indicate the query title, etc.

positional arguments:

query_file	Path to a sequence file in FASTA format that can be used as a similarity search query file. Or path to a directory containing only files for addition to the queries. Note: By default, the portion of the input filename preceding the first underscore character will be recorded as the "query title", the remaining substring preceding the second underscore character will be recorded as the taxon (e.g., "Hsapiens"), and the rest of the filename preceding the filename extension will be recorded as the sequence ID. So the filename might look like this: "QUERYTITLE_HSAPIENS_SEQUENCEID.fa". However, the relevant information can be revised in the "Queries/0_query_info.csv" file afterward if necessary.
main_data_dir	Path to main data directory (with Genomes, Queries, and Models subdirectories).

optional arguments:

-h, --help show this help message and exit

1.7 amoebae list_queries

usage: amoebae [-h] main_data_dir

Print a list of all usable query files in the query directory as defined in a given AMOEBAE data directory.

positional arguments:

main_data_dir	Path to main data directory (with Genomes, Queries, and Models subdirectories).
---------------	---

optional arguments:

-h, --help show this help message and exit

1.8 amoebae get_redun_hits

```
usage: amoebae [-h] [--query_name QUERY_NAME]
               [--query_list_file QUERY_LIST_FILE] [--db_name DB_NAME]
               [--db_list_file DB_LIST_FILE] [--query_title QUERY_TITLE]
               [--blast_report_evalue_cutoff BLAST_REPORT_EVALUE_CUTOFF]
               [--blast_max_target_seqs BLAST_MAX_TARGET_SEQS]
               [--hmmmer_report_evalue_cutoff HMMER_REPORT_EVALUE_CUTOFF]
               [--hmmmer_report_score_cutoff HMMER_REPORT_SCORE_CUTOFF]
               [--max_number_of_hits_to_summarize MAX_NUMBER_OF_HITS_TO_SUMMARIZE]
               [--num_threads_similarity_searching NUM_THREADS_SIMILARITY_SEARCHING]
               ]
               [--predict_redun_hit_selection] [--csv_file CSV_FILE]
               out_dir_path main_data_dir
```

Run searches with queries to find redundant hits in databases (for interpreting results).

positional arguments:

```
out_dir_path      Path to directory to write search results to.
main_data_dir     Path to main data directory (with Genomes, Queries,
                  and Models subdirectories).
```

optional arguments:

```
-h, --help          show this help message and exit
--query_name QUERY_NAME
                    Query filename to use (not full path). (default: None)
--query_list_file QUERY_LIST_FILE
                    Path to file containing a list of query files to use,
                    if no query_name is specified (or all queries by
                    default). (default: None)
--db_name DB_NAME   Name of database file in the database directory in
                    which to do searches (not full path). (default: None)
--db_list_file DB_LIST_FILE
                    Path to file containing a list of database files to
                    use (if no db_name specified). (default: None)
--query_title QUERY_TITLE
                    Name to be assigned to hits in databases that may be
                    considered redundant with a search query to which the
                    same title is assigned, otherwise it is taken from the
                    query info spreadsheet specified in the settings.py
                    file ('query_info_csv'). (default: None)
--blast_report_evalue_cutoff BLAST_REPORT_EVALUE_CUTOFF
                    Maximum E-value for reporting BLAST hits. (default:
                    0.05)
--blast_max_target_seqs BLAST_MAX_TARGET_SEQS
                    Maximum BLAST target sequences to consider. (default:
                    500)
--hmmmer_report_evalue_cutoff HMMER_REPORT_EVALUE_CUTOFF
                    Maximum E-value for reporting HMMer hits. (default:
                    0.05)
```

```

--hmmer_report_score_cutoff HMMER_REPORT_SCORE_CUTOFF
    Minimum sequence score for reporting HMMer hits.
    (default: 5)
--max_number_of_hits_to_summarize MAX_NUMBER_OF_HITS_TO_SUMMARIZE
    Absolute maximum number of hits (BLAST, HMMer, etc) to
    summarize in the output spreadsheet. This is important
    when working with sequences with WD40 domains, for
    example. (default: 50)
--num_threads_similarity_searching NUM_THREADS_SIMILARITY_SEARCHING
    Number of threads to use for running searches.
    (default: 4)
--predict_redun_hit_selection
    Write a copy of the output spreadsheet with '+' in
    rows for hits that may be specific to each query
    title, due to not being retrieved as top hits by
    queries associated with different query titles.
    (default: False)
--csv_file CSV_FILE Path to spreadsheet to append summary of result to for
    manual annotation. (default: None)

```

Recommendation: For most analyses, use the `--query_name` option and the `--db_name` option, and run the `get_redun_hits` command for each query separately. Otherwise, there will be redundant information in the output spreadsheet(s).

1.9 amoebae setup_fwd_srch

usage: amoebae [-h] [--outdir OUTDIR] srch_dir query_list_file db_list_file

Make a directory in which to write output files from similarity searches.

positional arguments:

```

srch_dir          Path to directory that will contain output directory as a
                  subdirectory.
query_list_file   Path to file with list of queries to search with.
db_list_file      Path to file with list of databases to search with.

```

optional arguments:

```

-h, --help        show this help message and exit
--outdir OUTDIR   Path to directory to put search results into (so that this
                  step can be piped together with other commands). (default:
                  None)

```

Note: Use the bash script to run forward searches on a remote server.

1.10 amoebae run_fwd_srch

```
usage: amoebae [-h] [--blast_report_evalue_cutoff BLAST_REPORT_EVALUE_CUTOFF]
               [--blast_max_target_seqs BLAST_MAX_TARGET_SEQS]
               [--hmmmer_report_evalue_cutoff HMMER_REPORT_EVALUE_CUTOFF]
               [--hmmmer_report_score_cutoff HMMER_REPORT_SCORE_CUTOFF]
               [--num_threads_similarity_searching NUM_THREADS_SIMILARITY_SEARCHING]
               fwd_srch_dir main_data_dir
```

Perform searches with original queries into subject databases.

positional arguments:

fwd_srch_dir	Path to directory that will contain forward search output files.
main_data_dir	Path to main data directory (with Genomes, Queries, and Models subdirectories).

optional arguments:

-h, --help	show this help message and exit
--blast_report_evalue_cutoff BLAST_REPORT_EVALUE_CUTOFF	Maximum E-value for reporting BLAST hits. (default: 0.05)
--blast_max_target_seqs BLAST_MAX_TARGET_SEQS	Maximum BLAST target sequences to consider. (default: 500)
--hmmmer_report_evalue_cutoff HMMER_REPORT_EVALUE_CUTOFF	Maximum E-value for reporting HMMer hits. (default: 0.05)
--hmmmer_report_score_cutoff HMMER_REPORT_SCORE_CUTOFF	Minimum sequence score for reporting HMMer hits. (default: 5)
--num_threads_similarity_searching NUM_THREADS_SIMILARITY_SEARCHING	Number of threads to use for running searches. (default: 4)

1.11 amoebae sum_fwd_srch

```
usage: amoebae [-h] [--csv_file CSV_FILE] [--max_evalue MAX_EVALUE]
               [--max_gap_between_tblastn_hsps MAX_GAP_BETWEEN_TBLASTN_HSPS]
               [--do_not_use_exonerate]
               [--exonerate_score_threshold EXONERATE_SCORE_THRESHOLD]
               [--max_hits_to_sum MAX_HITS_TO_SUM]
               [--max_length_diff MAX_LENGTH_DIFF]
               fwd_srch_out csv_out_path main_data_dir
```

Append information about forward searches to csv summary file (this is used to

organize reverse searches). For TBLASTN searches (protein queries, nucleotide target sequences), HSPs are clustered into groups that are close enough within the target sequence to potentially represent exons from the same coding sequence. The nucleotide subsequences in which these clusters of HSPs are found are then analyzed using exonerate to identify and translate potential exons, in "protein2genome" mode, because exonerate, unlike TBLASTN, attempts to identify exon boundaries, yielding translations that are less likely to include translations of non-coding regions outside exons (which might include apparent stop codons).

positional arguments:

fwd_srch_out	Path to directory where forward search results were written.
csv_out_path	Path to output summary spreadsheet (CSV) file.
main_data_dir	Path to main data directory (with Genomes, Queries, and Models subdirectories).

optional arguments:

-h, --help	show this help message and exit
--csv_file CSV_FILE	Path to summary spreadsheet (CSV) file, which already contains search summaries. If such a file is specified, then the output CSV file will contain the columns from this CSV file with additional columns summarizing additional forward search results. (default: None)
--max_value MAX_VALUE	Maximum E-value threshold for reporting forward search hits. (default: 0.0005)
--max_gap_between_tblastn_hsps MAX_GAP_BETWEEN_TBLASTN_HSPS	Maximum number of nucleotide bases between TBLASTN HSPs to be considered part of the same gene locus. This is important, because it will be assumed that HSP separated by more than this number of nucleotide bases are not part of the same gene or TBLASTN "hit". (default: 10000)
--do_not_use_exonerate	Override the default use of exonerate to identify coding sequences and translations, and just use TBLASTN instead. This option is provided because concatenated TBLASTN HSPs may be more inclusive of sequences within the target sequence, and the results of TBLASTN and exonerate may need to be compared. Also, note that HSPs identified by TBLASTN but for which exonerate yields no alignments will be ignored if exonerate is used. (default: False)
--exonerate_score_threshold EXONERATE_SCORE_THRESHOLD	Set score threshold to be applied when running exonerate on nucleotide sequences identified by TBLASTN. The default for setting of exonerate is 100,

but a lower score is set as default here, because otherwise exonerate cannot identify some of the sequences identified by TBLASTN. This option is only relevant if using exonerate. (default: 10)

`--max_hits_to_sum MAX_HITS_TO_SUM`
Maximum number of forward search hits to list in the summary spreadsheet. If zero, then reverse searches will be performed for all hits. (default: 0)

`--max_length_diff MAX_LENGTH_DIFF`
Maximum number of amino acid residues length difference allowed between the original query and the forward hit sequence. If -1, then a maximum length cutoff will not be applied. (default: -1)

1.12 amoebae setup_rev_srch

usage: amoebae [-h] [--outdir OUTDIR] [--aasubseq] [--nafullseq]
srch_dir csv_file databases main_data_dir

Make directory in which to write results of reverse searches.

positional arguments:

<code>srch_dir</code>	Path to directory that will contain output directory as a subdirectory.
<code>csv_file</code>	Path to summary spreadsheet (CSV) file, which contains a summary of forward search(es).
<code>databases</code>	Database filename (in database directory) or path to file with list of database filenames. Note that filenames are needed, not file paths.
<code>main_data_dir</code>	Path to main data directory (with Genomes, Queries, and Models subdirectories).

optional arguments:

<code>-h, --help</code>	show this help message and exit
<code>--outdir OUTDIR</code>	Path to directory to put search results into (so that this step can be piped together with other commands). (default: None)
<code>--aasubseq</code>	Use only the portion of each (amino acid) forward hit sequence that aligns to the original query used (top HSP subject sequence). This is default for nucleotide hits. (default: False)
<code>--nafullseq</code>	Use the full (nucleic acid) forward hit sequence. This is default for amino acid hits. (default: False)

1.13 amoebae run_rev_srch

```
usage: amoebae [-h] [--blast_report_evalue_cutoff BLAST_REPORT_EVALUE_CUTOFF]
              [--blast_max_target_seqs BLAST_MAX_TARGET_SEQS]
              [--hmmmer_report_evalue_cutoff HMMER_REPORT_EVALUE_CUTOFF]
              [--hmmmer_report_score_cutoff HMMER_REPORT_SCORE_CUTOFF]
              [--num_threads_similarity_searching NUM_THREADS_SIMILARITY_SEARCHING]
              ]
              rev_srch_dir main_data_dir
```

Perform searches with forward search hit sequences as queries into the original query databases.

positional arguments:

```
  rev_srch_dir      Path to directory that will contain output of
                    searches.
  main_data_dir     Path to main data directory (with Genomes, Queries,
                    and Models subdirectories).
```

optional arguments:

```
  -h, --help          show this help message and exit
  --blast_report_evalue_cutoff BLAST_REPORT_EVALUE_CUTOFF
                    Maximum E-value for reporting BLAST hits. (default:
                    0.05)
  --blast_max_target_seqs BLAST_MAX_TARGET_SEQS
                    Maximum BLAST target sequences to consider. (default:
                    500)
  --hmmmer_report_evalue_cutoff HMMER_REPORT_EVALUE_CUTOFF
                    Maximum E-value for reporting HMMer hits. (default:
                    0.05)
  --hmmmer_report_score_cutoff HMMER_REPORT_SCORE_CUTOFF
                    Minimum sequence score for reporting HMMer hits.
                    (default: 5)
  --num_threads_similarity_searching NUM_THREADS_SIMILARITY_SEARCHING
                    Number of threads to use for running searches.
                    (default: 4)
```

1.14 amoebae sum_rev_srch

```
usage: amoebae [-h] [--redun_hit_csv REDUN_HIT_CSV]
              [--min_evaldiff MIN_EVALDIFF] [--aasubseq] [--nafullseq]
              [--max_rev_srchs MAX_REV_SRCHS]
              fwd_srch_csv rev_srch_out csv_out_path main_data_dir
```

Append information about reverse searches to csv summary file. Use information from redundant hit csv file to interpret results.

positional arguments:

```
  fwd_srch_csv      Path to summary spreadsheet (CSV) file, which contains
```

forward search summaries and also may already contain reverse search summaries.

rev_srch_out Path to directory where reverse search results were written.

csv_out_path Path to output spreadsheet (CSV) file with reverse search results appended to previous results.

main_data_dir Path to main data directory (with Genomes, Queries, and Models subdirectories).

optional arguments:

-h, --help show this help message and exit

--redun_hit_csv REDUN_HIT_CSV
Path to spreadsheet (CSV) file, which specifies which hits are redundant positive hits for a given query (query title) in a given database. If this is not provided, then it is assumed that any and all reverse search hits are equivalent to/redundant with the original query. (default: None)

--min_evaldiff MIN_EVALDIFF
Minimum difference in E-value order of magnitude between top reverse search hit and first reverse search hit that is not redundant with the original query. (default: 5)

--aasubseq Use only the portion of each (amino acid) forward hit sequence that aligns to the original query used (top HSP subject sequence). This is default for nucleotide hits. Must be selected if selected when the setup_rev_srch command was run. (default: False)

--nafullseq Use the full (nucleic acid) forward hit sequence. This is default for amino acid hits. Must be selected if selected when the setup_rev_srch command was run. (default: False)

--max_rev_srchs MAX_REV_SRCHS
Maximum number of forward search hits to perform reverse searches for per query database. If zero, then reverse searches will be performed for all hits. (default: 0)

1.15 amoebae interp_srchs

usage: amoebae [-h] [--fwd_only] [--fwd_evalue_cutoff FWD_EVALUATE_CUTOFF]
 [--rev_evalue_cutoff REV_EVALUATE_CUTOFF]
 [--hmmmer_cutoff HMMER_CUTOFF] [--no_overlapping_hits]
 [--out_csv_path OUT_CSV_PATH]
 csv_file

Interpret search results based on final summary, which provides a basis for

further analyses of positive hits.

positional arguments:

csv_file Path to spreadsheet with forward and reverse search results.

optional arguments:

-h, --help show this help message and exit
--fwd_only Interpret forward searches based on score (HMMer) cutoff. (default: False)
--fwd_evalue_cutoff FWD_EVALUE_CUTOFF Specify an (more stringent) E-value cutoff for forward search results. (default: None)
--rev_evalue_cutoff REV_EVALUE_CUTOFF Specify an (more stringent) E-value cutoff for reverse search results. (default: None)
--hmmer_cutoff HMMER_CUTOFF Specify a score that hits must exceed to be included. (default: 20)
--no_overlapping_hits If more than one query (query title) retrieves a given sequence as a positive hit based on the search criteria, make the sequence a negative hit for all queries (query titles), except for the one that retrieved the sequence with the lowest (strongest) E-value. Warning: Do not use this option if you are searching sequences that include genomic sequences that may include more than one genomic locus per sequence. False-negative results could occur in this case, because different queries for non-orthologous genes could retrieve subsequences in the same subject sequence. (default: False)
--out_csv_path OUT_CSV_PATH Optionally specify an output file path, so that this command can be piped together with others. (default: None)

1.16 amoebae find__redun__seqs

usage: amoebae [-h] [--out_csv_path OUT_CSV_PATH]
 [--remove_tblastn_hits_at_annotated_loci]
 [--just_look_for_genes_in_gff3] [--ignore_gff3]
 [--allow_internal_stops ALLOW_INTERNAL_STOPS]
 [--min_length MIN_LENGTH]
 [--min_percent_length MIN_PERCENT_LENGTH]
 [--min_percent_query_cover MIN_PERCENT_QUERY_COVER]
 [--overlap_required] [--max_percent_ident MAX_PERCENT_IDENT]

```

[--min_alig_res_in_overlap MIN_ALIG_RES_IN_OVERLAP]
[--min_ident_res_in_overlap MIN_IDENT_RES_IN_OVERLAP]
[--min_sim_res_in_overlap MIN_SIM_RES_IN_OVERLAP]
[--min_ident_span_len MIN_IDENT_SPAN_LEN]
[--min_sim_span_len MIN_SIM_SPAN_LEN]
[--min_percent_ident_in_overlap MIN_PERCENT_IDENT_IN_OVERLAP]
[--min_percent_sim_in_overlap MIN_PERCENT_SIM_IN_OVERLAP]
[--min_percent_overlap MIN_PERCENT_OVERLAP]
[--plot_hit_exclusion] [--add_ali_col]
csv_file main_data_dir

```

Identify hit sequences likely encoded by the same gene loci in the genome of a given species, or otherwise not representing paralogous genes. Criteria are applied in this order: 1. Peptide hits with the same ID as higher-ranking hits for the same query (query title) are excluded. 2. Nucleotide hits for the same loci as peptide sequence hits are excluded. 3. Sequences with internal stop codons are excluded, as these are potentially pseudogenes. 4. Sequences are excluded if they do not meet several minimum length criteria: Absolute minimum length (in amino acids) and percent query cover. 5. Sequences are excluded if they do not overlap to a specified degree with all included higher-ranking hits for the same query (query title) in sequence data for the same species/genome. This is determined by algorithmically comparing pairs of sequences aligned to a reference alignment of homologues, and several minimum measures of alignment overlap may be specified. 6. Secondary hit sequences are excluded if they do not meet a specified maximum percent identity threshold. Highly identical sequences may result from false segmental duplications in the genome assembly, may represent alleles, etc. Note: Applying these criteria requires a column to be manually added to the input csv file prior to running with the header "Alignment for sequence comparison" and filled with the appropriate alignment name to use (one for each query title, as listed in the "Query title" column). Alternatively, you can run this command with the --add_ali_col option to automatically identify appropriate alignments among your aligned FASTA queries used for running HMMer searches. If no alignment (.afaa) file can be found, then the first single sequence query file (.faa) that appears in the summary CSV file will be used instead.

positional arguments:

csv_file	Path to spreadsheet with interpreted search results outputted by the interp_srchs command.
main_data_dir	Path to main data directory (with Genomes, Queries, and Models subdirectories).

optional arguments:

-h, --help	show this help message and exit
--out_csv_path OUT_CSV_PATH	Optionally specify an output file path, so that this command can be piped together with others. (default: None)
--remove_tblastn_hits_at_annotated_loci	

Ignore tblastn hits that overlap with any previously annotated loci. The rationale for this would be that the corresponding protein sequences should have been retrieved if the tblastn hit were a true positive anyway. If this option is not specified, then sequences will still be excluded if they specifically correspond to the same loci as do higher-ranking hits. (default: False)

`--just_look_for_genes_in_gff3` When looking for records in GFF3 annotation files that overlap with subsequences identified by similarity searching (TBLASTN), ignore records that are not explicitly "gene" (for example, "CDS", "mRNA", and "exon"). This option should probably not be selected, because in some GFF3 annotation files do not include "gene" records, but do include predicted coding sequences for genes. (default: False)

`--ignore_gff3` Disregard any information regarding redundancy of identified nucleotide sequences with identified protein sequences that may be found in GFF3 annotation files. (default: False)

`--allow_internal_stops ALLOW_INTERNAL_STOPS` Include sequences that have internal stop codons (anywhere other than the N-terminal position). (default: True)

`--min_length MIN_LENGTH` Absolute minimum length (in AA) of a hit sequence to be considered a potential distinct paralogue. (default: 55)

`--min_percent_length MIN_PERCENT_LENGTH` Minimum length (in AA) of a hit sequence as a percentage of query length for the hit to be considered a potential distinct paralogue. (default: 15)

`--min_percent_query_cover MIN_PERCENT_QUERY_COVER` Minimum number of residues aligning with the original query as a percentage of the original query sequence length. (default: 0)

`--overlap_required` True if hits must overlap with a higher-ranking hit to be considered potential unique paralogues. (default: False)

`--max_percent_ident MAX_PERCENT_IDENT` Maximum percent identity (among aligning residues) for evaluating whether two sequences are redundant or not (secondary hits showing a percent identity with a higher-ranking hit exceeding this value will be excluded). (default: 98.0)

`--min_alig_res_in_overlap MIN_ALIG_RES_IN_OVERLAP` Minimum number of residues which must align for two

sequences to be considered as potentially distinct hits. This is only relevant if the `overlap_required` option is specified. (default: 20)

`--min_ident_res_in_overlap MIN_IDENT_RES_IN_OVERLAP`
 Minimum number of aligning residues which must be identical for two sequences to be considered as potentially distinct hits. This is only relevant if the `overlap_required` option is specified. (default: 10)

`--min_sim_res_in_overlap MIN_SIM_RES_IN_OVERLAP`
 Minimum number of aligning residues which must be similar for two sequences to be considered as potentially distinct hits. This is only relevant if the `overlap_required` option is specified. (default: 15)

`--min_ident_span_len MIN_IDENT_SPAN_LEN`
 Minimum number of aligning residues which are identical that must exist in at least one continuous span for two sequences to be considered as potentially distinct hits (not counting positions where both sequences have gaps). This is only relevant if the `overlap_required` option is specified. (default: 0)

`--min_sim_span_len MIN_SIM_SPAN_LEN`
 Minimum number of aligning residues which are similar (or identical) that must exist in at least one continuous span for two sequences to be considered as potentially distinct hits (not counting positions where both sequences have gaps). This is only relevant if the `overlap_required` option is specified. (default: 0)

`--min_percent_ident_in_overlap MIN_PERCENT_IDENT_IN_OVERLAP`
 Minimum percent identity between the two sequences of interest in the alignment. This is only relevant if the `overlap_required` option is specified. (default: 0)

`--min_percent_sim_in_overlap MIN_PERCENT_SIM_IN_OVERLAP`
 Minimum percent similarity (including identity) between the two sequences of interest in the alignment. This is only relevant if the `overlap_required` option is specified. (default: 0)

`--min_percent_overlap MIN_PERCENT_OVERLAP`
 Minimum number of aligning residues between the two sequences of interest as a percentage of the length of the second sequence (the last sequence in the alignment), not including gaps, for the two sequences to be considered as potentially distinct hits. This is only relevant if the `overlap_required` option is specified. (default: 0)

`--plot_hit_exclusion` Plot number of hits excluded by the various criteria applied. (default: False)

`--add_ali_col` Add a column to the csv file listing which alignment file in the queries directory to use for comparing sequences. Aligned FASTA queries are selected that match the query titles of the original queries used to retrieve each of the relevant hits listed in the csv file. No other options need to be specified in this case. (default: False)

1.17 amoebae plot

usage: amoebae [-h] [--csv_file2 CSV_FILE2] [--complex_info COMPLEX_INFO]
 [--row_order ROW_ORDER] [--out_pdf OUT_PDF]
 csv_file

Plot results of similarity search and sequence classification analyses. The outputs are PDF files.

positional arguments:

`csv_file` Path to a spreadsheet with the relevant results to be plotted. This can be either a CSV file output of the `sum_rev_srch` command or from the `find_redun_seqs` command. If the output of the `sum_rev_srch` command is used, however, redundant hits will be counted (e.g., BLASTP and TBLASTN hits corresponding to the same or highly identical genomic loci).

optional arguments:

`-h, --help` show this help message and exit
`--csv_file2 CSV_FILE2` Path to a second spreadsheet with relevant results to be compared to the first and plotted. (default: None)
`--complex_info COMPLEX_INFO` Path to file that specifies which query titles represent components of which protein complexes (or otherwise grouped proteins). (default: None)
`--row_order ROW_ORDER` Path to file that specifies the order in which data for each species will be displayed. (default: None)
`--out_pdf OUT_PDF` Path to output pdf file. (default: None)

1.18 amoebae csv_to_fasta

usage: amoebae [-h] [--output_dir OUTPUT_DIR] [--abbrev] [--paralogue_names]
 [--only_descr] [--subseq] [--all_hits] [--split_by_query_title]
 [--split_by_top_rev_srch_hit SPLIT_BY_TOP_REV_SRCH_HIT]

`[--split_to_query_fastas]`
`csv_file`

Extract sequences described in a spreadsheet output by AMOEBAE, and write to a file in FASTA format.

positional arguments:

`csv_file` Path to csv file listing sequences.

optional arguments:

`-h, --help` show this help message and exit

`--output_dir OUTPUT_DIR`

Path for output directory to contain FASTA files.
(default: None)

`--abbrev` Add species name instead of sequence description from fasta header. Applicable when the output file is to be used for alignment and phylogenetic analysis.
(default: False)

`--parologue_names` Use species name, query title, and parologue number instead of sequence description from fasta header. Applicable when the output file is to be used for alignment and phylogenetic analysis. Does not work if the abbrev option is specified. (default: False)

`--only_descr` Use the description but not the ID as the new fasta sequence header. Does not work if the abbrev option is specified. (default: False)

`--subseq` Write subsequences that aligned to forward search query, instead of the full sequences. (default: False)

`--all_hits` Write all forward hits listed in the input csv file.
(default: False)

`--split_by_query_title` Write sequences to files according to the query title of the query which retrieved them in a similarity search. (default: False)

`--split_by_top_rev_srch_hit SPLIT_BY_TOP_REV_SRCH_HIT`

Write sequences to files according to the top hit that they retrieve in a reverse search, for each sequence that meets the reverse search criteria. (Provide the reverse search identifier, eg, "rev_srch_20180924122402-1") (default: None)

`--split_to_query_fastas`

Write sequences to separate files with filenames that can be easily parsed for loading the the files as queries using the `add_to_queries` command. (default: False)

1.19 amoebae check_depend

usage: amoebae [-h]

Check that all the dependencies (other than python modules) are properly installed and useable.

optional arguments:

-h, --help show this help message and exit

1.20 amoebae check_imports

usage: amoebae [-h]

Check that all the import statements used in the AMOEBAE repository run without error.

optional arguments:

-h, --help show this help message and exit

1.21 amoebae regen_genome_info

usage: amoebae [-h] data_dir_path

Write a new genome info spreadsheet (O_genome_info.csv) file using filenames from the Genomes directory.

positional arguments:

data_dir_path Specify the full path to an existing AMOEBAE data directory, which contains a 'Genomes' subdirectory. The new genome info file will be added to this subdirectory.

optional arguments:

-h, --help show this help message and exit