NAME

swarm — find clusters of nearly-identical nucleotide amplicons

SYNOPSIS

swarm [*options*] *filename*

DESCRIPTION

Environmental or clinical molecular studies generate large volumes of amplicons (e.g., 16S or 18S SSU-rRNA sequences) that need to be clustered into molecular operational taxonomic units (OTUs). Common clustering methods are based on greedy, input-order dependent algorithms, with arbitrary selection of global cluster size and cluster centroids. To address that problem, we developed **swarm**, a fast and robust method that recursively groups amplicons with *d* or less differences (i.e. substitutions, insertions or deletions). **swarm** produces natural and stable clusters centered on local peaks of abundance, mostly free from input-order dependency induced by centroid selection.

Exact clustering is impractical on large data sets when using a naïve all-vs-all approach (more precisely a 2-combination without repetitions), as it implies unrealistic numbers of pairwise comparisons. **swarm** is based on a maximum number of differences *d* between two amplicons, and focuses only on very close local relationships. For *d* = 1, the default value, **swarm** uses an algorithm of linear complexity that generates all possible single mutations and performs exact-string matching by comparing hash-values. For *d* = 2 or greater, **swarm** uses an algorithm of quadratic complexity that performs pairwise string comparisons. An efficient *k*-mer-based filtering and an astute use of comparisons results obtained during the clustering process allows **swarm** to avoid most of the amplicon comparisons needed in a naïve approach. To speed up the remaining amplicon comparisons, **swarm** implements an extremely fast Needleman-Wunsch algorithm making use of the Streaming SIMD Extensions (SSE2) of modern x86-64 CPUs. If SSE2 instructions are not available, **swarm** exits with an error message.

swarm reads the named input *filename*, a fasta file of nucleotide amplicons. The amplicon identifier is defined as the string comprised between the ">" symbol and the first space or the end of the line, whichever comes first. As **swarm** outputs lists of amplicon identifiers, amplicon identifiers must be unique to avoid ambiguity; swarm exits with an error message if identifiers are not unique. Amplicon identifiers must end with a "_" followed by a positive integer representing the amplicon copy number (or abundance annotation; usearch/vsearch users can use the option -z to change that behavior). Abundance annotations play a crucial role in the clustering process, and swarm exits with an error message if that information is not available. The amplicon sequence is defined as a string of [acgt] or [acgu] symbols (case insensitive), starting after the end of the identifier line and ending before the next identifier line or the file end; **swarm** exits with an error message if any other symbol is present.

General options

-h, **--help** display this help and exit successfully.

-t, **--threads** *positive integer*
: number of computation threads to use. Values between 1 and 256 are accepted, but we recommend to use a number of threads lesser or equal to the number of available CPU cores. Default number of threads is 1.

-v, **--version**
: output version information and exit successfully.

-- delimit the option list. Later arguments, if any, are treated as operands even if they begin with "-". For example, "swarm -- -file.fasta" reads from the file "-file.fasta".

Clustering options

-d, **--differences** *zero or positive integer*
: maximum number of differences allowed between two amplicons, meaning that two amplicons will be grouped if they have *integer* (or less) differences. This is **swarm**'s most important parameter. The number of differences is calculated as the number of mismatches (substitutions, insertions or deletions) between the two amplicons once the optimal pairwise global alignment has been found (see "pairwise alignment advanced options" to influence that step). Any *integer*

between 0 and 256 can be used, but high *d* values will decrease the taxonomical resolution of **swarm** results. Commonly used *d* values are 1, 2 or 3, rarely higher. When using *d* = 0, **swarm** will output results corresponding to a strict dereplication of the dataset, i.e. merging identical amplicons. Warning, whatever the *d* value, **swarm** requires fasta entries to present abundance values. Default number of differences *d* is 1.

-n, **--no-otu-breaking**

deactivate the built-in OTU refinement (not recommended). Amplicon abundance values are used to identify transitions among in-contact OTUs and to separate them, yielding higher-resolution clustering results. That option prevents that separation, and in practice, allows the creation of a link between amplicons A and B, even if the abundance of B is higher than the abundance of A.

Fastidious options

-b, **--boundary** *positive integer*

when using the option --fastidious (-f), define the minimum mass of a *large* OTU. By default, an OTU with a mass of 3 or more is considered large. Conversely, an OTU is small if it has a mass of less than 3, meaning that it is composed of either one amplicon of abundance 2, or two amplicons of abundance 1. Any positive value greater than 1 can be specified. Using higher boundary values will speed up the second pass, but also reduce the taxonomical resolution of **swarm** results. Default mass of a large OTU is 3.

-c, **--ceiling** *positive integer*

when using the option --fastidious (-f), define **swarm**'s maximum memory footprint (in megabytes). **swarm** will adjust the --bloom-bits (-y) value of the Bloom filter to fit within the specified amount of memory.

-f, **--fastidious**

when working with *d* = 1, perform a second clustering pass to reduce the number of small OTUs (recommended option). During the first clustering pass, an intermediate amplicon can be missing for purely stochastic reasons, interrupting the aggregation process. The fastidious option will create virtual amplicons, allowing to graft small OTUs upon bigger ones. By default, an OTU is small if it has a mass of 2 or less (see the --boundary option to modify that value). To speed things up, **swarm** uses a Bloom filter to store intermediate results. Warning, the second clustering pass can be 2 to 3 times slower than the first pass and requires much more memory to store the virtual amplicons in Bloom filters. See the options --bloom-bits (-y) or --ceiling (-c) to control the memory footprint of the Bloom filter. The fastidious option modifies clustering results: the output files produced by the options --log (-l), --output-file (-o), --mothur (-r), --uclust-file, and --seeds (-w) are updated to reflect these modifications; the file --statistics-file (-s) is partially updated (columns 6 and 7 are not updated); the output file --internal-structure (-i) is not updated.

-y, **--bloom-bits** *positive integer*

when using the option --fastidious (-f), define the size (in bits) of each entry in the Bloom filter. That option allows to balance the efficiency (i.e. speed) and the memory footprint of the Bloom filter. Large values will make the Bloom filter more efficient but will require more memory. Any value between 2 and 64 can be used. Default value is 16. See the --ceiling (-c) option for an alternative way to control the memory footprint.

Input/output options

-a, **--append-abundance** *positive integer*

set abundance value to use when some or all amplicons in the input file lack abundance values. Warning, it is not recommended to use **swarm** on datasets where abundance values are all identical. We provide that option as a courtesy to advanced users, please use it carefully. **swarm** exits with an error message if abundance values are missing and if this option is not used.

-i, **--internal-structure** *filename*

output all pairs of nearly-identical amplicons to *filename* using a five-columns tab-delimited format:

1. amplicon A label.
2. amplicon B label.
3. number of differences between amplicons A and B (*positive integer*).
4. OTU number (*positive integer*). OTUs are numbered in their order of delineation, starting from 1. All pairs of amplicons belonging to the same OTU will receive the same number.
5. number of steps from the OTU seed to amplicon B (*positive integer*).

-l, **--log** *filename*
output all messages to *filename* instead of *standard error*, with the exception of error messages of course. That option is useful in situations where writing to *standard error* is problematic (for example, with certain job schedulers).

-o, **--output-file** *filename*
output clustering results to *filename*. Results consist of a list of OTUs, one OTU per line. An OTU is a list of amplicon identifiers separated by spaces. That output format can be modified by the option --mothur (-r). Default is to write to standard output.

-r, **--mothur**
output clustering results in a format compatible with Mothur. That option modifies **swarm**'s default output format.

-s, **--statistics-file** *filename*
output statistics to *filename*. The file is a tab-separated table with one OTU per row and seven columns of information:

1. number of unique amplicons in the OTU,
2. total abundance of amplicons in the OTU,
3. identifier of the initial seed,
4. initial seed abundance,
5. number of amplicons with an abundance of 1 in the OTU,
6. maximum number of iterations before the OTU reached its natural limit),
7. theoretical maximum radius of the OTU (i.e., number of cummulated differences between the seed and the furthermost amplicon in the OTU). The actual maximum radius of the OTU is often much smaller.

-u, **--uclust-file** *filename*
output clustering results in uclust-like file format to the specified file. That option does not modify **swarm**'s default output format.

-w, **--seeds** *filename*
output OTU representatives to *filename* in fasta format. The abundance value of each OTU representative is the sum of the abundances of all the amplicons in the OTU.

-z, **--usearch-abundance**
accept amplicon abundance values in usearch/vsearch's style (>label;size=*integer*[;]). That option influences the abundance annotation style used in output files.

Pairwise alignment advanced options
when using $d > 1$, **swarm** recognizes advanced command-line options modifying the pairwise global alignment scoring parameters:

-m, **--match-reward** *positive integer*
Default reward for a nucleotide match is 5.

-p, **--mismatch-penalty** *positive integer*
Default penalty for a nucleotide mismatch is 4.

-g, **--gap-opening-penalty** *positive integer*
Default gap opening penalty is 12.

-e, **--gap-extension-penalty** *positive integer*
Default gap extension penalty is 4.

As **swarm** focuses on close relationships (e.g., d = 2 or 3), clustering results are resilient to pairwise alignment model parameters modifications. When clustering using a higher d value, modifying model parameters has a stronger impact.

EXAMPLES

Clusterize the data set *myfile.fasta* into OTUs with the finest resolution possible (1 difference, built-in breaking, fastidious option) using 4 computation threads. OTUs are written to the file *myfile.swarms*, and OTU representatives are written to *myfile.representatives.fasta*.

swarm -t 4 -f -w *myfile.representatives.fasta* < *myfile.fasta* > *myfile.swarms*

AUTHORS

Concept by Frédéric Mahé, implementation by Torbjørn Rognes.

CITATION

Mahé F, Rognes T, Quince C, de Vargas C, Dunthorn M. (2014) Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ* 2:e593 <https://doi.org/10.7717/peerj.593>

Mahé F, Rognes T, Quince C, de Vargas C, Dunthorn M. (2015) Swarm v2: highly-scalable and high-resolution amplicon clustering. *PeerJ* 3:e1420 <https://doi.org/10.7717/peerj.1420>

REPORTING BUGS

Submit suggestions and bug-reports at <https://github.com/torognes/swarm/issues>, send a pull request on <https://github.com/torognes/swarm>, or compose a friendly or curmudgeonly e-mail to Frédéric Mahé <mahe@rhrk.uni-kl.de> and Torbjørn Rognes <torognes@ifi.uio.no>.

AVAILABILITY

Source code and binaries are available at <https://github.com/torognes/swarm>

COPYRIGHT

SEE ALSO

swipe, an extremely fast Smith-Waterman database search tool by Torbjørn Rognes (available from <https://github.com/torognes/swipe>).

vsearch, an open-source re-implementation of the classic uclust clustering method (by Robert C. Edgar), along with other amplicon filtering and searching tools. **vsearch** is implemented by Torbjørn Rognes and documented by Frédéric Mahé, and is available at <https://github.com/torognes/vsearch>.

VERSION HISTORY

New features and important modifications of **swarm** (short lived or minor bug releases are not mentioned):

v2.1.11 released January 16, 2017
Version 2.1.11 fixes two bugs related to the SIMD implementation of alignment that might result in incorrect alignments and scores. The bug only applies when d>1.

v2.1.10 released December 22, 2016
Version 2.1.10 fixes two bugs related to gap penalties of alignments. The first bug may lead to wrong aligments and similarity percentages reported in UCLUST (.uc) files. The second bug makes Swarm use a slightly higher gap extension penalty than specified. The default gap extension penalty used have actually been 4.5 instead of 4.

v2.1.9 released July 6, 2016
Version 2.1.9 fixes errors when compiling with GCC version 6.

v2.1.8 released March 11, 2016
Version 2.1.8 fixes a rare bug triggered when clustering extremely short undereplicated sequences. Also, alignment parameters are not shown when $d = 1$.

v2.1.7 released February 24, 2016
Version 2.1.7 fixes a bug in the output of seeds with the -w option when $d > 1$ that was not properly fixed in version 2.1.6. It also handles ascii character n°13 (CR) in FASTA files better. Swarm will now exit with status 0 if the -h or the -v option is specified. The help text and some error messages have been improved.

v2.1.6 released December 14, 2015
Version 2.1.6 fixes problems with older compilers that do not have the x86intrin.h header file. It also fixes a bug in the output of seeds with the -w option when $d > 1$.

v2.1.5 released September 8, 2015
Version 2.1.5 fixes minor bugs.

v2.1.4 released September 4, 2015
Version 2.1.4 fixes minor bugs in the swarm algorithm used for $d = 1$.

v2.1.3 released August 28, 2015
Version 2.1.3 adds checks of numeric option arguments.

v2.1.1 released March 31, 2015
Version 2.1.1 fixes a bug with the fastidious option that caused it to ignore some connections between large and small OTUs.

v2.1.0 released March 24, 2015
Version 2.1.0 marks the first official release of swarm v2.

v2.0.7 released March 18, 2015
Version 2.0.7 writes abundance information in usearch style when using options -w (--seeds) in combination with -z (--usearch-abundance).

v2.0.6 released March 13, 2015
Version 2.0.6 fixes a minor bug.

v2.0.5 released March 13, 2015
Version 2.0.5 improves the implementation of the fastidious option and adds options to control memory usage of the Bloom filter (-y and -c). In addition, an option (-w) allows to output OTU representatives sequences with updated abundances (sum of all abundances inside each OTU). This version also enables **swarm** to run with $d = 0$.

v2.0.4 released March 6, 2015
Version 2.0.4 includes a fully parallelised implementation of the fastidious option.

v2.0.3 released March 4, 2015
Version 2.0.3 includes a working implementation of the fastidious option, but only the initial clustering is parallelized.

v2.0.2 released February 26, 2015
Version 2.0.2 fixes SSSE3 problems.

v2.0.1 released February 26, 2015
Version 2.0.1 is a development version that contains a partial implementation of the fastidious option, but it is not usable yet.

v2.0.0 released December 3, 2014
Version 2.0.0 is faster and easier to use, providing new output options (--internal-structure and --log), new control options (--boundary, --fastidious, --no-otu-breaking), and built-in OTU refinement (no need to use the python script anymore). When using default parameters, a novel and considerably faster algorithmic approach is used, guaranteeing **swarm**'s scalability.

v1.2.21 released February 26, 2015
Version 1.2.21 is supposed to fix some problems related to the use of the SSSE3 CPU instructions which are not always available.

v1.2.20 released November 6, 2014
Version 1.2.20 presents a production-ready version of the alternative algorithm (option -a), with optional built-in OTU breaking (option -n). That alternative algorithmic approach (usable only with $d = 1$) is considerably faster than currently used clustering algorithms, and can deal with datasets of 100 million unique amplicons or more in a few hours. Of course, results are rigourously identical to the results previously produced with swarm. That release also introduces new options to control swarm output (options -i and -l).

v1.2.19 released October 3, 2014
Version 1.2.19 fixes a problem related to abundance information when the sequence identifier includes multiple underscore characters.

v1.2.18 released September 29, 2014
Version 1.2.18 reenables the possibility of reading sequences from *stdin* if no file name is specified on the command line. It also fixes a bug related to CPU features detection.

v1.2.17 released September 28, 2014
Version 1.2.17 fixes a memory allocation bug introduced in version 1.2.15.

v1.2.16 released September 27, 2014
Version 1.2.16 fixes a bug in the abundance sort introduced in version 1.2.15.

v1.2.15 released September 27, 2014
Version 1.2.15 sorts the input sequences in order of decreasing abundance unless they are detected to be sorted already. When using the alternative algorithm for $d = 1$ it also sorts all subseeds in order of decreasing abundance.

v1.2.14 released September 27, 2014
Version 1.2.14 fixes a bug in the output with the --swarm_breaker option (-b) when using the alternative algorithm (-a).

v1.2.12 released August 18, 2014
Version 1.2.12 introduces an option --alternative-algorithm to use an extremely fast, experimental clustering algorithm for the special case $d = 1$. Multithreading scalability of the default algorithm has been noticeably improved.

v1.2.10 released August 8, 2014
Version 1.2.10 allows amplicon abundances to be specified using the usearch style in the sequence header (e.g. ">id;size=1") when the -z option is chosen.

v1.2.8 released August 5, 2014
Version 1.2.8 fixes an error with the gap extension penalty. Previous versions used a gap penalty twice as large as intended. That bug correction induces small changes in

clustering results.

v1.2.6 released May 23, 2014

Version 1.2.6 introduces an option --mothur to output clustering results in a format compatible with the microbial ecology community analysis software suite Mothur (<http://www.mothur.org/>).

v1.2.5 released April 11, 2014

Version 1.2.5 removes the need for a POPCNT hardware instruction to be present. **swarm** now automatically checks whether POPCNT is available and uses a slightly slower software implementation if not. Only basic SSE2 instructions are now required to run **swarm**.

v1.2.4 released January 30, 2014

Version 1.2.4 introduces an option --break-swarms to output all pairs of amplicons with *d* differences to standard error. That option is used by the companion script ‘swarm_breaker.py‘ to refine **swarm** results. The syntax of the inline assembly code is changed for compatibility with more compilers.

v1.2 released May 16, 2013

Version 1.2 greatly improves speed by using alignment-free comparisons of amplicons based on *k*-mer word content. For each amplicon, the presence-absence of all possible 5-mers is computed and recorded in a 1024-bits vector. Vector comparisons are extremely fast and drastically reduce the number of costly pairwise alignments performed by **swarm**. While remaining exact, **swarm** 1.2 can be more than 100-times faster than **swarm** 1.1, when using a single thread with a large set of sequences. The minor version 1.1.1, published just before, adds compatibility with Apple computers, and corrects an issue in the pairwise global alignment step that could lead to sub-optimal alignments.

v1.1 released February 26, 2013

Version 1.1 introduces two new important options: the possibility to output clustering results using the uclust output format, and the possibility to output detailed statistics on each OTU. **swarm** 1.1 is also faster: new filterings based on pairwise amplicon sequence lengths and composition comparisons reduce the number of pairwise alignments needed and speed up the clustering.

v1.0 released November 10, 2012

First public release.