NAME

swarm — find clusters of nearly-identical nucleotidic amplicons

SYNOPSIS

swarm [*options*] [*filename*]

DESCRIPTION

Environmental studies generate large volumes of amplicons (usually rRNA SSU sequences) that need to be clustered into molecular operational taxonomic units. Traditional clustering methods are based on greedy, input-order dependent algorithms, with arbitrary selection of cluster size and cluster centroids. To address that problem, we developed **swarm**, a fast and exact method that recursively groups amplicons with *n* or less differences. **swarm** produces stable clusters (or "swarms"), free from centroid selection induced input-order dependency.

Exact clustering is impractical on large data sets when using a naïve all-vs-all approach (i.e. a 2-combination without repetitions), as it implies unrealistic numbers of pairwise comparisons. **swarm** is based on a maximum number of differences, and focuses only on close relationships. An astute use of comparisons results obtained during the process allows to avoid up to 98% 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 (SSE4.1) of modern x86-64 CPUs. If SSE4.1 instructions are not available, **swarm** exits with an error message.

swarm reads the named input *filename*, a fasta file of nucleotidic 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. 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. Default is to read from standard input if no file is named, or the file name is "-".

Options

swarm recognizes the following command-line options:

-d, **--differences** *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 "advanced options" for parameters influencing the pairwise alignment). Any *integer* between 1 and 256 can be used, but aligning two very distant amplicons is difficult and results should be considered with caution. Default number of differences is 1.

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

-o, **--output-file** *filename*
: output result to *filename*. Result is a list of swarms, one swarm per line. A swarm is a list of amplicon identifiers separated by spaces. Default is to write to standard output.

-u, **--uclust-file** *filename*
: output results in UCLUST-like file format to the specified file. Default is not to output in this format.

-s, **--statistics-file** *filename*
: output statistics to the specified file. Default is not to output statistics. The file is tab separated and each line contains the following information about a swarm: number of unique amplicons in the swarm, total number of amplicons in the swarm, identifier of the initial seed, initial seed abundance, number of singletons (amplicons with an abundance of 1), maximum number of generations and the maximum radius.

-t, **--threads** *integer*
number of computation threads to use. The number of threads should be lesser or equal to the number of available CPU cores. Default number of threads is 1.

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

Advanced options

swarm recognizes advanced command-line options modifying the pairwise global alignment scoring parameters:

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

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

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

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

As **swarm** focuses on close relationships, final results are resilient to model parameters modifications. Modifying model parameters only impacts swarms with a large number of subseed levels, or analysis using a high number of differences.

EXAMPLES

swarm -t 4 -o *myfile.swarms myfile.fasta*
Divide the data set *myfile.fasta* into swarms with the finest resolution possible (1 difference) using 4 computation threads, and write the results in the file *myfile.swarms*.

zcat file.fas.gz | **swarm** | awk "{print NF}" | sort -n | uniq -c
Use swarm in a pipeline to read a compressed fasta file and to get its swarm size profile (with default parameters).

AUTHORS

Concept by Frédéric Mahé, implementation by Torbjørn Rognes.

REPORTING BUGS

Report bugs to <mahe@rhrk.uni-kl.de> and <torognes@ifi.uio.no>.

AVAILABILITY

The software is available from <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>).