Classify Sequences with CLARK

CLARK (CLAssifier based on Reduced K-mers) is a tool for supervised sequence classification based on discriminative k-mers.

UGENE provides the GUI for CLARK and CLARK-l variants of the CLARK framework for solving the problem of the assignment of metagenomic reads to known genomes.

Parameters in GUI

Parameter	Description	Defaultvalue
Input data	To classify single-end (SE) reads or contigs, received by reads de novo assembly, set this parameter to "SE reads or contigs". To classify paired-end (PE) reads, set the value to "PE reads".	SE reads or contigs
Classification tool	Use CLARK-l on workstations with limited memory (i.e., "l" for light), this software tool provides precise classification on small metagenomes. It works with a sparse or "light" database (up to 4 GB of RAM) while still performing ultra accurate and fast results.	CLARK-l
Database	A path to the folder with the CLARK database files (-D). It is assumed that "targets.txt" file is located in this folder (the file is passed to the "classify_metagenome.sh" script from the CLARK package via parameter -T).	
Minimum k-mer frequency	Minimum of k-mer frequency/occurrence for the discriminative k-mers (-t). For example, for 1 (or, 2), the program will discard any discriminative k-mer that appear only once (or, less than twice).	0
Mode	Set the mode of the execution (-m): • "Full" to get detailed results, confidence scores and other statistics. • "Default" to get results summary and perform best trade-off between classification speed, accuracy and RAM usage. • "Express" to get results summary with the highest speed possible.	Default
Gap	"Gap" or number of non-overlapping k-mers to pass when creating the database (-). Increase the value if it is required to reduce the RAM usage. Note that this will degrade the sensitivity.	4
Load database into memory	Request the loading of database file by memory mapped-file (--ldm). This option accelerates the loading time but it will require an additional amount of RAM significant. This option also allows to load the database in multithreaded-task (see also the "Number of threads" parameter).	False
Number of threads	Use multiple threads for the classification and, with the "Load database into memory" option enabled, for the loading of the database into RAM (-n).	8
Output file	Specify the output file name.	auto

Parameters in Workflow File

Type: clark-classify

Parameter	Parameter in the GUI	Type
sequencing-reads	**Input data**	*string*
tool-variant	**Classification tool**	*number*
database	**Database**	*string*
k-min-freq	**Minimum k-mer frequency**	*number*
mode	**Mode**	*bool*
gap	**Gap**	*number*
preload	**Load database into memory**	*bool*
threads	**Number of threads**	*number*
output-url	**Output file**	*string*

Input/Output Ports

The element has 1 *input port*:

Name in GUI: Input sequences:

URL(s) to FASTQ or FASTA file(s) should be provided. In casethe of SE reads or contigs use the "Input URL 1" slot only.

In case of PE reads input "left" reads to "Input URL 1", "right" reads to "Input URL 2". See also the "Input data" parameter of the element.

Name in Workflow File: in

Slots:

SlotInGUI	Slot in Workflow File	Type
Input URL 1	**url**	*string*

The element has 1 *output port*:

Name in GUI: CLARK Classification:

A map of sequence names with the associated taxonomy IDs, classified by CLARK.

Name in Workflow File: out

Slots:

SlotInGUI	Slot in Workflow File	Type
Taxonomy classification data	tax-data	*tax-classification*