Classify Sequences with DIAMOND Element

In general, DIAMOND is a sequence aligner for protein and translated DNA searches similar to the NCBI BLAST software tools. However, it provides a speedup of BLAST ranging up to x20,000. Using this workflow element one can use DIAMOND for taxonomic classification of short DNA reads and longer sequences such as contigs. The lowest common ancestor (LCA) algorithm is used for the classification.

Element type: diamond-classify

Parameters

Parameter	Description	Defaultvalue	Parameter in Workflow File	Type
Database	Input a binary DIAMOND database file.		**database**	*string*
Genetic code	Genetic code used for translation of query sequences (--query-gencode).	The standard genetic code	**genetic-code**	*number*
Sensitive mode	The sensitive modes (--sensitive, --more-sensitive) are generally recommended for aligning longer sequences. The default mode is mainly designed for short read alignment, i.e. finding significant matches of >50 bits on 30-40aa fragments.	Default	**sensitive-mode**	*string*
Top alignments percentage	DIAMOND uses the lowest common ancestor (LCA) algorithm for taxonomy classification of the input sequences. This parameter specifies what alignments should be taken into account during the calculations (--top). For example, the default value "10" means to take top 10% of the best hits (i.e. sort all query/subject-alignments by a score, take top 10% of the alignments with the best score, calculate the lowest common ancestor for them).	10%	**top-alignments-percentage**	*number*
Frameshift	Penalty for frameshift in DNA-vs-protein alignments. Values around 15 are reasonable for this parameter. Enabling this feature will have the aligner tolerate missing bases in DNA sequences and is most recommended for long, error-prone sequences like MinION reads.	Skipped	**frame-shift**	*number*
Expected value	Maximum expected value to report an alignment (--evalue/-e).	0.0010	**e-value**	*number*
Matrix	Scoring matrix (--matrix).	BLOSUM62	**matrix**	*string*
Gap open penalty	Gap open penalty (--gapopen).	Default	**gap-open**	*number*
Gap extension penalty	Gap extension penalty (--gapextend).	Default	**gap-extend**	*number*
Block size	Block size in billions of sequence letters to be processed at a time (--block-size). This is the main parameter for controlling the program's memory usage. Bigger numbers will increase the use of memory and temporary disk space, but also improve performance. The program can be expected to use roughly six times this number of memory (in GB).	0.5	**block-size**	*number*
Index chunks	The number of chunks for processing the seed index (--index-chunks). This option can be additionally used to tune the performance. It is recommended to set this to 1 on a high memory server, which will increase performance and memory usage, but not the usage of temporary disk space.	4	**index-chunks**	*number*
Number of threads	Number of CPU threads (--treads).	8	**threads**	*number*
Output file	Specify the output file name. The output file is a tab-delimited file with the following fields: * Query ID * NCBI taxonomy ID (0 if unclassified) * E-value of the best alignment with a known taxonomy ID found for the query (0 if unclassified)	auto	**output-url**	*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.

The input files may contain single-end reads, contigs, or "left" reads in case of the paired-end sequencing (see "Input data" parameter of the element).

Name in Workflow File: in

Slots:

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

The element has 1 *output port*:

Name in GUI: DIAMOND Classification:

A list of sequence names with the associated taxonomy IDs, classified by DIAMOND.

Name in Workflow File: out

Slots:

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