The Microbializer pipeline aims to process bacterial genomes and analyze them. The first part finds orthologous sets across all given species. The flow for finding orthologous sets will be handled by a wrapper that will run different modules one by one:

1. blast\_all\_vs\_all.py

**Input**: (1) 2 input paths for 2 (different) bacterial genomes, g1 and g2 (2) an output file path (with a suffix as follows: i\_vs\_j\_blast.tsv. especially relevant for the wrapper).

**Output**: a tsv file where the first column contains g1 genes and the second column includes the corresponding best match.

Precisely, for each gene x in g1, blast x among all the genes of g2. Let y be the gene in g2 that is the most similar to x among all g2 genes. Append a row to the output file with: ‘{x}\t{y}’.

1. filter\_putative\_pairs.py

**Input**: (1) a path for a i\_vs\_j\_blast.tsv file (2) an output path (with a suffix as follows: i\_vs\_j\_filtered.tsv. especially relevant for the wrapper).

**Output**: the same format of the input file containing only pairs that passed the filtration. For each row in the input file (pair of genes), apply the following filters:

1. at least X% similarity

2. at least X% of the length

3.

write each pair to the output file if it passed all the above filters.

1. find\_reciprocal\_hits.py

**Input**: (1) a path for a i\_vs\_j\_filtered.tsv file and a path for a j\_vs\_i\_filtered.tsv file (2) an output path (with a suffix as follows: i\_vs\_j\_reciprocal\_hits.tsv

**Output**: the same format of the input file containing only reciprocal best-hit pairs, i.e., if x’s best hit was y in the first file, x\ty will appear in the output file only if y’s best hit in the other file will be x.

1. construct\_putative\_orthologs\_table.py

**Input**: (1) a path for a i\_vs\_j\_reciprocal\_hits.tsv file (2) a path for a putative orthologs table (each row in this table is a putative orthologous set).

**Output**: updates the table with the info from the reciprocal hit file.

1. construct\_final\_table.py

**Input**: (1) a path to directory containing tabular BLAST results (2) a path to the putative orthologs table (3) a path for a directory that will contain the final orthologs table (each row in this table is a final orthologous set), and two subdirectories input\_mcl and output\_mcl that will hold the input and output files of mcl respectively.

**Output**: final orthologs table (only orthologs that passed the clustering criterion).

1. extract\_sequences.py

**Input**: (1) final orthologs table (2) a directory with sequences files (3) a directory for output files (each OG will have its own fasta file).

**Output**: fasta file for each ortholog groups, with the sequences of each one.