Pickle 1.0

User Manual

August 2019

**1. Introduction**

**Pickle** is the software designed to search for two- and three-dimensional patterns in tabulated OTU data files. There are four pipelines available for the specific pattern identification:

1. 2D\_patterns\_main

Performs search for two dimensional co-exclusion, co-presence, and one-way relation patterns

1. 3D\_All\_together\_or\_alone

Perform search for three dimensional all together or alone pattern

1. 3D\_Switch\_Pattern

Perform search for three dimensional X23 present if X1 present pattern

1. 3D\_Type2\_Coexclusion

Perform search for three dimensional type2 co-exclusion pattern

Questions, suggestions, and concerns related to the performance and functionality of the software should be addressed via e-mail: gegolovk@utmb.edu

**2. Compile Pickle**

In order to compile C++ source code please use the following command:

g++ -O3 2D\_Patterns\_Main.cpp –w –o 2D\_Patterns

g++ -O3 3D\_All\_Together\_Or\_Alone.cpp –w –o 3D\_All\_Together\_Or\_Alone

g++ -O3 3D\_Switch\_Pattern.cpp –w –o 3D\_Switch\_Pattern

g++ -O3 3D\_Type2\_Coexclusion.cpp –w –o 3D\_Type2\_Coexclusion

*Note:* it is possible to use any other C++ compiler

**3. Run Pickle**

There are three groups of parameters (nine values total) which can affect results of the search for both 2D and 3D patterns and can be defined using the command line.

Presence threshold optimization parameters:

-n min\_presence\_threshold (default value 0.0);

-m max\_presence\_threshold (default value 0.005 or 0.5%);

-s threshold\_presence\_step (default value 0.0001 or 0.01%).

These parameters are used to identify if the organism considered to be present or absent in the sample.

The population threshold optimization parameters are defining the minimal fraction of experimental observations required to be present in each of quarters (in multidimensional cases sectors) defining each particular pattern:

-p min\_population\_threshold (default value 0.1 or 10%);

-t max\_population\_threshold (default value 0.2 or 20%);

-r population\_threshold\_step (default value 0.01 or 1%).

The patterns score threshold parameters are used to filter which combination of the of the population threshold/pattern score will be presented in the output file.

-z min\_score (default value 0.9 or 90%);

-x max\_score (default value 1 or 100%);

-c score\_step (default value 0.01 or 1%).

It is important to mention that while the presence and population threshold parameters are contributing to the performance of the application (time complexity), the patterns score threshold parameters affect only the total number of the identified patterns and as result only affect output file size.

**Command line example**

To run any of the 2D or 3D Pickle pipelines in a Linux environment, the user must provide the command line with 2 parameters separated by a single space:

Example of command with default parameters: 2D\_patterns\_main -i ./Anterior\_nares\_genus\_example.txt -o ./Anterior\_nares\_genus\_manual\_

Example of full command: 2D\_patterns\_main -i ./Anterior\_nares\_genus\_example.txt -o ./Anterior\_nares\_genus\_manual\_ -n 0.0 -m 0.005 -s 0.0001 -p 0.1 -t 0.2 -r 0.01 -z 0.9 -x 1 -c 0.001

**5. Input File Format**

Input file for each of the available pattern search pipelines is a tabulated OTU file. The first line of the input file is a header with a list of samples separated by tab. Starting from the second line, file should contain OTU name and corresponding normalized values separated by tabs. Each line should end with taxonomical lineage associated to the OTU.