**User guide**

This is a user guide for scripts written in undergraduate project on bioinformatics-based metabolic gene cluster annotation of Cordyceps militaris and related entomopathogenic fungi, available at https://github.com/jennyjenny8/Research\_project\_codes. The example of input and output files from the project are also included in the resources. This program was developed to identify specific and shared secondary metabolites gene cluster predicted from Antibiotics and Secondary Metabolite Analysis SHell (antiSMASH) fungal version 4.1.0 between main and related species. The results of shared secondary metabolites gene cluster will be illustrate as gene cluster diagram and phylogenetic tree.

**Operation system**

This program was built using Windows 10 64-bit operating system on Intel Core i7-8550U CPU with 8 GB RAM.

**Program requirement**

Python version 3.6.2

NCBI BLAST + executable version 2.7.1

Multiple Alignment using Fast Fourier Transform (MAFFT) command-line version 7.395

**Python modules requirement**

* Biopython version 1.7
* Pandas version 0.22.0
* Xlrd version 1.1.0
* Matplotlib version 3.0.3
* Reportlab version 3.4.0

**Installation**

After install all program required, pip install all module required. Then download and unzipped file from https://github.com/jennyjenny8/Research\_project\_codes.

**Inputs**

The input file require for this program are FASTA file of all protein sequence in the genome and result from antiSMASH. There are two type of AntiSMASH results used, the first type is excel spreadsheet (.xls) file which record the gene accessions of gene in each gene cluster. The second type of AntiSMASH results is GenBank file for each of the predicted gene cluster. All input files must located in specific folder order as shown below inside the input folder. The input files should be separate into main and related folder for main and related species, respectively. Inside the folder of each species should include protein sequences, excel file and a folder of GenBank file. Order of the input file is shown below.

* Input
  + main
    - Protein sequences (filename: GenBankassembly\_protein.faa)
    - Excel file (filename: GenBankaccession.geneclusters.xls)
    - cluster
      * gene cluster GenBank files (GenBankaccession.cluster.gbk)
  + related
    - Species name
      * Protein sequences (filename: GenBankassembly\_protein.faa)
      * Excel file (filename: GenBankaccession.geneclusters.xls)
      * cluster
        + gene cluster GenBank files (GenBankaccession.cluster.gbk)

**Execution**

After install all program requirement, set Python and NCBI BLAST as path in the environment variables. Then change the location of mafft.bat in phylogenetic\_tree.py. Double-click on run.bat or right-click run.bat and select Run as administrator. A new window will pop up, wait until ‘PAUSE Press any key to continue . . .’ show up which means the program finish.

**Output**

The results will be save in a folder inside output folder with date and time as folder name. The output files included text file containing all pair of shared secondary metabolites gene cluster between main and related species (shared\_gene\_cluster\_file.txt), image file of a horizontal stacked bar graph created from shared secondary metabolites gene cluster result (shared\_gene\_cluster\_barplot.png), PDF file of each group of shared secondary metabolites gene cluster (main species gene cluster number.pdf) and tree file (COMBINED\_nj.nhx). The tree file can be open using graphical viewer of phylogenetic trees program.