Manual of VirionLang (v1.1)

1. Operating system

Linux (VirionLang has been tested on Ubuntu)
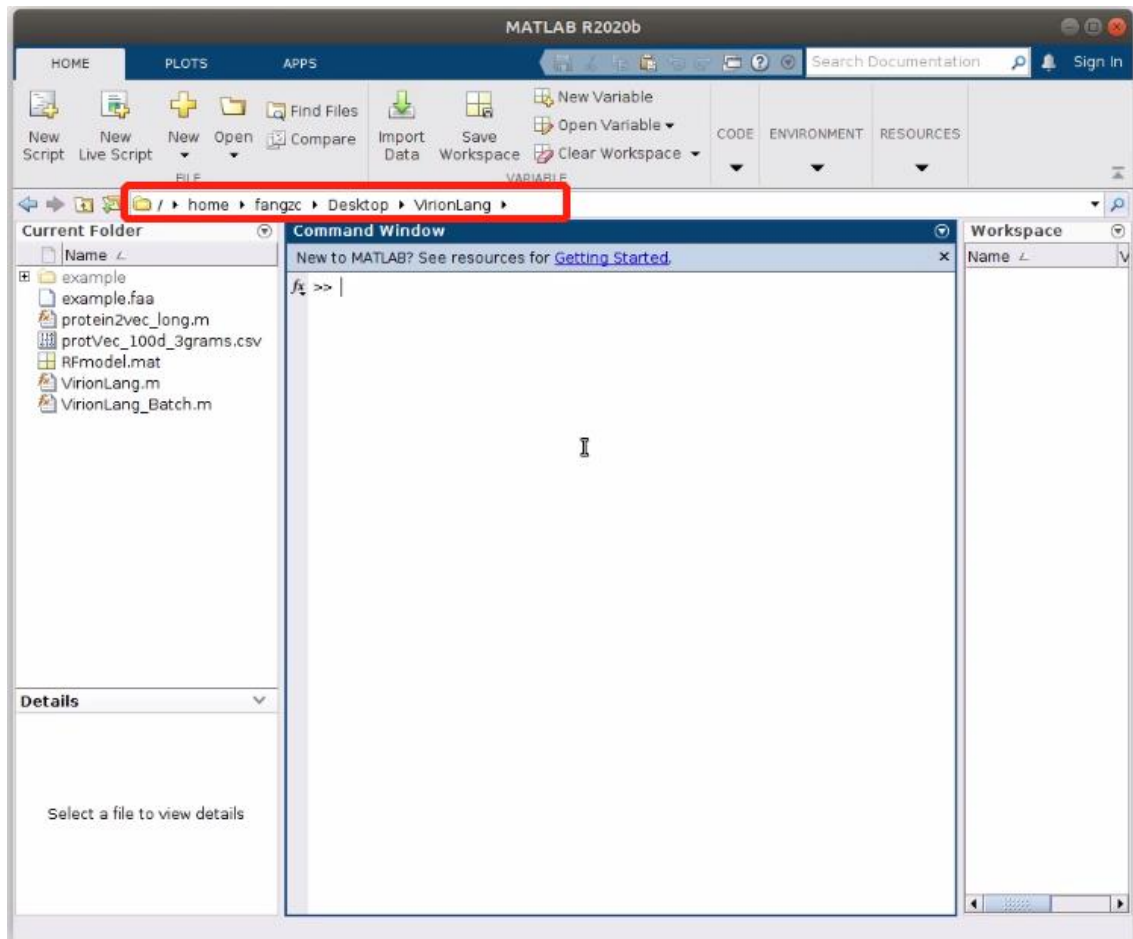
2. Requirements

MATLAB (https://www.mathworks.com/, VirionLang was developed in MATLAB 2020b.)

3. Usage

Step 1. To run VirionLang, please download VirionLang package from GitHub website https://github.com/zhenchengfang/VirionLang.

Step 2. Open the MATLAB, and change the working path of MATLAB to the program folder.

Step 3a. VirionLang takes a "fasta" file as input and outputs a table file. Please execute the following command directly in the MATLAB command window.

VirionLang('<input_file_folder>/input_file.faa', '<output_file_folder>/output_file.csv')

For example, if you want to identify the sequences in example.faa, please execute:

**VirionLang('./example.faa', './result.csv')**
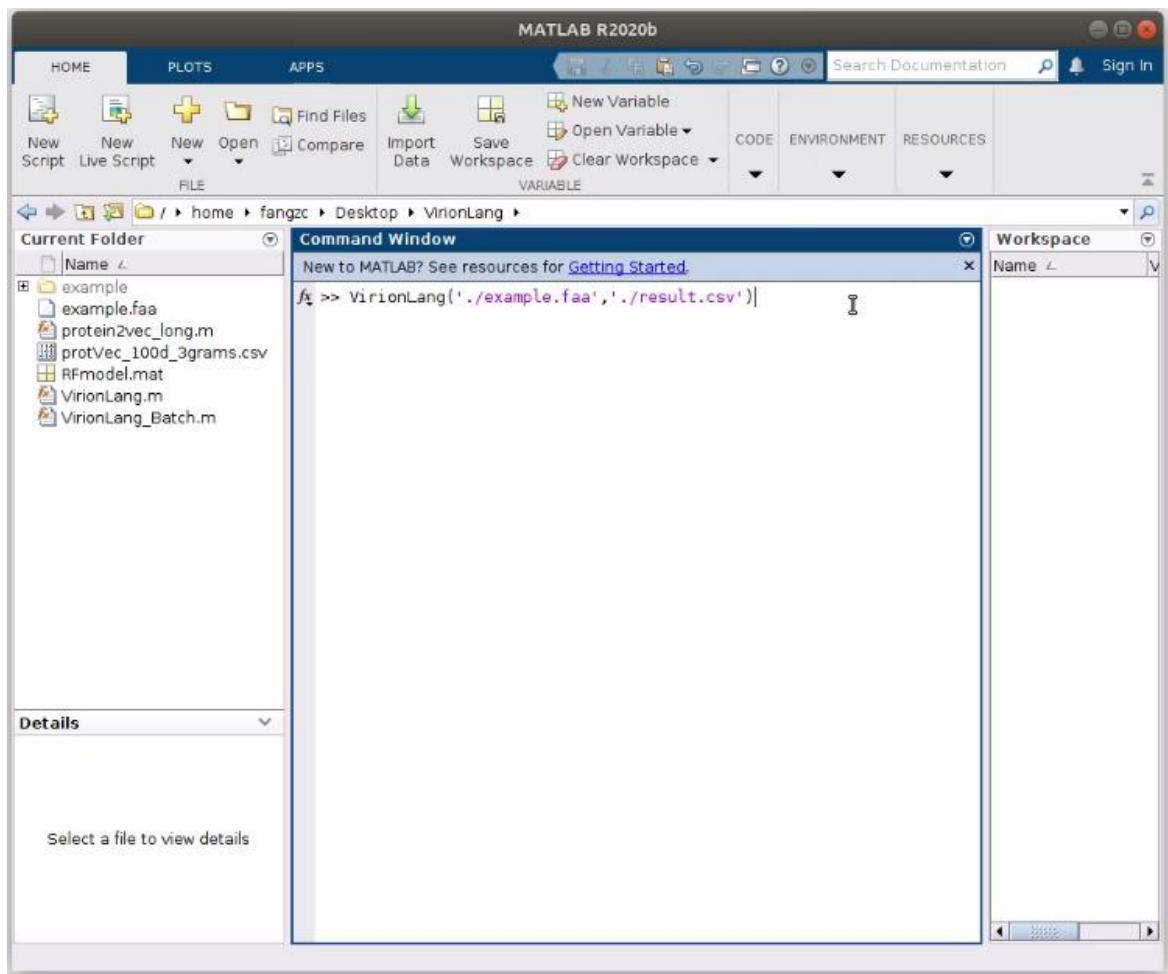
Step 3b. If you have multi files to process, you may depose all the files in one folder, and execute:

VirionLang_Batch('<input_folder_path>')

Please ensure that the input folder contain **only** the "fasta" files that need to be processed. The output files will depose under the input folder taking ".VirionLang.csv" as the suffix. For example, if you want to identify the sequences in the "example" folder, please execute:

VirionLang_Batch('./example/')
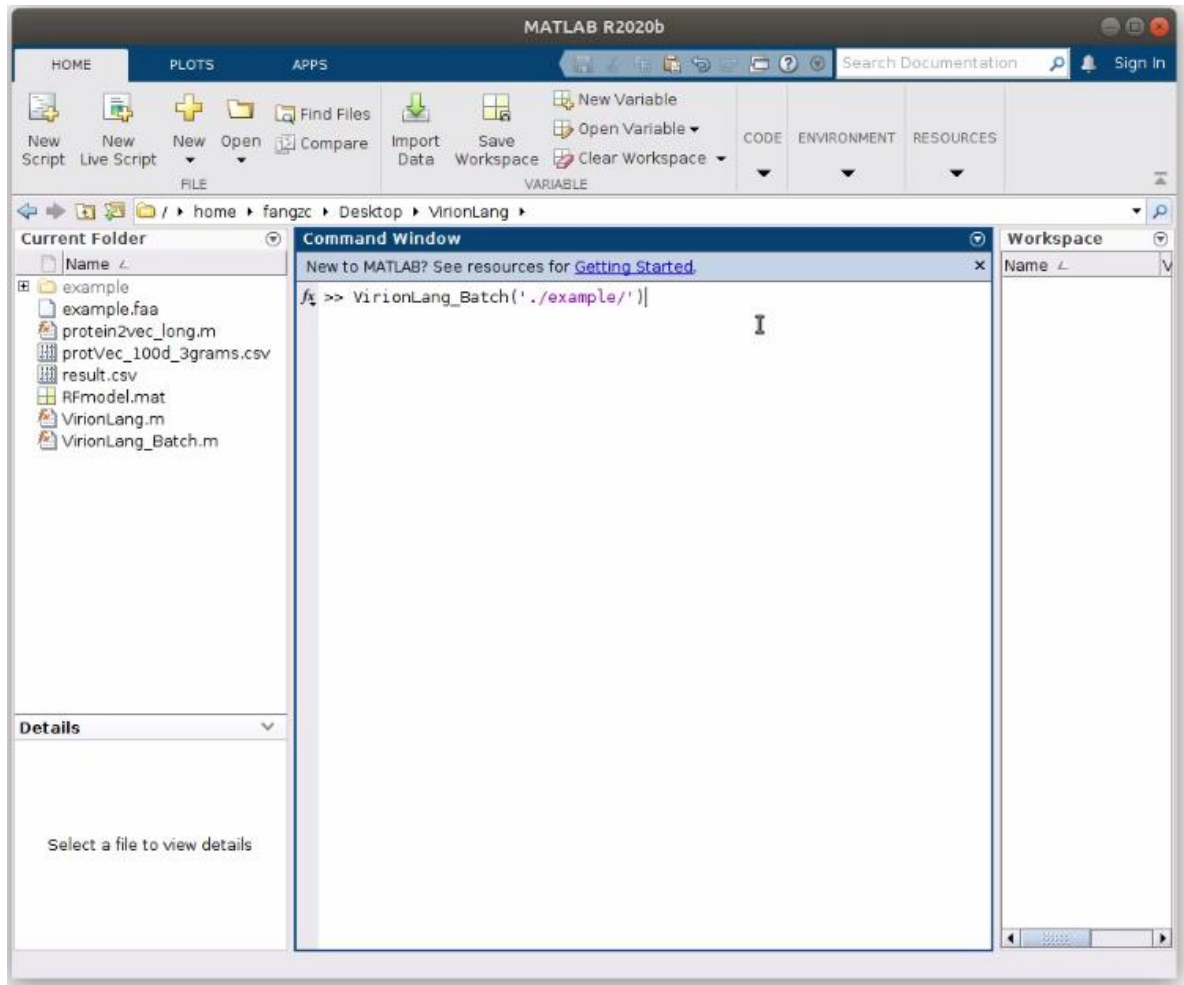
4. Output

The output of VirionLang consists of 2 columns, representing "sequence header" (the same with the corresponding header in the "fasta" file), and the probability score that the sequence belongs to a PVP. Here is a screenshot of the output file:

	A	B
1	Header	Score
2	YP_009618579.1 major capsid protein	0.8505
3	YP_009618594.1 baseplate assembly protein	0.756867
4	YP_009618598.1 tail fiber protein	0.8096
5	YP_009618710.1 major capsid protein	0.8505
6	YP_009618719.1 tail tapemeasure protein	0.644543

The $Sn=TP/(TP+FN)$ and $Sp=TN/(TN+FP)$ of VirionLang under different thresholds are shown in the following figure.
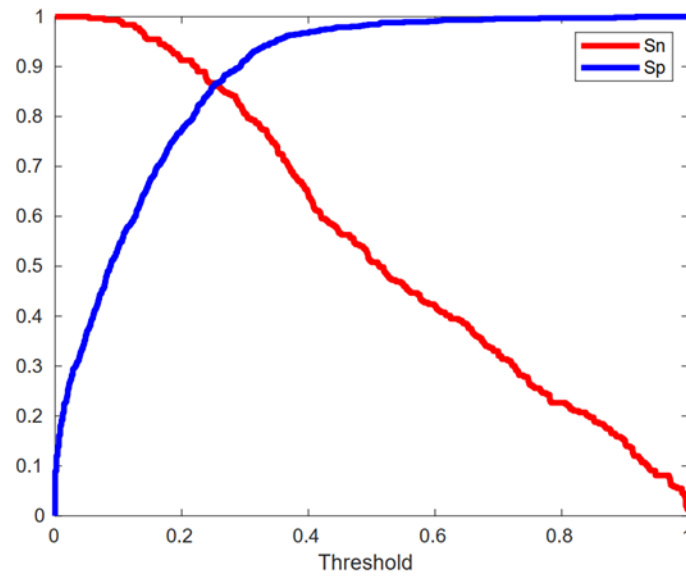
Any question, please do not hesitate to contact me: fangzc@smu.edu.cn.