Manual of PPR-Meta Release Version 1.0

Table of Contents

Any question, please do not hesitate to contact me: fangzc@pku.edu.cn

PPR-Meta can run either in the physical host or virtual machine. The physical host version can speed up with GPU. For non-computer professionals, we recommend using the virtual machine version. On PPR-Meta homepage, there is a brief video guide to show how to run PPR-Meta in virtual machine.

Part Ⅰ. Physical host version

1. Operating system

Linux (PPR-Meta has been tested on Ubuntu 16.04)

2. Requirements

●Python 2.7.12 (https://www.python.org/)
●Python packages:
  ○numpy 1.13.1(http://www.numpy.org/)
  ○h5py 2.6.0(http://www.h5py.org/)
●TensorFlow 1.4.1 (https://www.tensorflow.org/)
●Keras 2.0.8 (https://keras.io/)
●MATLAB Component Runtime (MCR) R2018a
(https://www.mathworks.com/products/compiler/matlab-runtime.html)

or

MATLAB R2018a (https://www.mathworks.com/products/matlab.html)

Note:
1. For compatibility, we recommend installing the tools with the similar version as described above.
2. If GPU is available in your machine, we recommend installing a GPU version of the TensorFlow to speed up the program.
3. PPR-Meta can be run with either an executable file or a MATLAB script. If you run PPR-Meta through the executable file, you need to install the MCR while MATLAB is not necessary. If you run PPR-Meta through the MATLAB script, MATLAB is required.

3. Preparation

Please install numpy, h5py, TensorFlow, Keras, MCR (or MATLAB) according to their manuals.

The numpy, h5py, TensorFlow, and Keras are python packages, which can be installed with "pip". If "pip" is not already installed in your machine, use the command "sudo apt-get install python-pip python-dev" to install "pip". Here are example commands of installing the above python packages using "pip".

pip install numpy

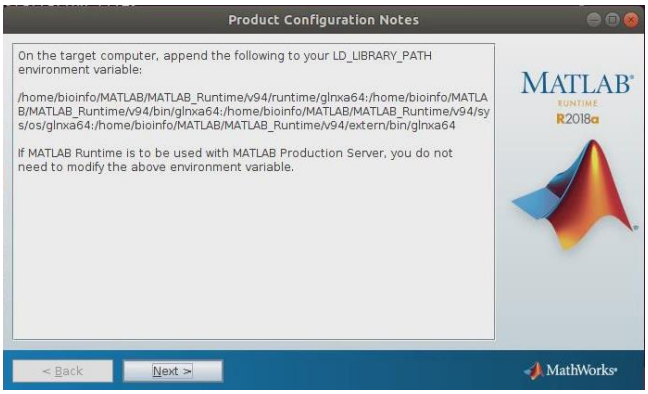
pip install h5py
pip install tensorflow==1.4.1 #CPU version
pip install tensorflow-gpu==1.4.1 #GPU version
pip install keras==2.0.8

If you are going to install a GPU version of the TensorFlow, specified NVIDIA software should be installed. See https://www.tensorflow.org/install/gpu to know whether your machine can install TensorFlow with GPU support.

When running PPR-Meta through the executable file, MCR should be installed. See https://www.mathworks.com/help/compiler/install-the-matlab-runtime.html to install MCR. On the target computer, please append the following to your LD_LIBRARY_PATH environment variable according to the tips of MCR:

<MCR_installation_folder>/v94/runtime/glnxa64
<MCR_installation_folder>/v94/bin/glnxa64
<MCR_installation_folder>/v94/sys/os/glnxa64
<MCR_installation_folder>/v94/extern/bin/glnxa64

A screenshot of the tips when installing MCR is shown below:



When running PPR-Meta through the MATLAB script, please see https://www.mathworks.com/support/ to install the MATLAB.

4. Usage

To run PPR-Meta, please download PPR-Meta package using git, and then change directory to PPR-Meta:

git clone https://github.com/zhenchengfang/PPR-Meta.git
cd PPR-Meta

Also, you can download PPR-Meta as a zipped file, and then unpack the zipped file and change directory to PPR-Meta:

wget http://cqb.pku.edu.cn/ZhuLab/PPR_Meta/PPR_Meta_v_1_0.zip
unzip PPR_Meta_v_1_0.zip
cd PPR_Meta_v_1_0

4.1. Run by executable file (in command line)

In this form, please simply executes the command:

./PPR_Meta <input_file_folder>/input_file.fna <output_file_folder>/output_file.csv

The input file must be in fasta format containing the sequences to be identified. The users can use the file "example.fna" which contains 300 sequences in the program folder to test the PPR-Meta by simply executing the command:

./PPR_Meta example.fna result.csv

4.2. Run by MATLAB script (in MATLAB GUI)

In this form, please execute the following command directly in the MATLAB command window.

PPR_Meta('<input_file_folder>/input_file.fna', '<output_file_folder>/output_file.csv')

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

PPR_Meta('example.fna', 'result.csv')

Remember to set the working path of MATLAB to the program folder before running the program.

4.3. Run with specified threshold (-t option)

For each input sequence, PPR-Meta will output three scores (between 0 to 1), representing the probability that the sequence belongs to a phage, chromosome or plasmid. By default, the prediction of PPR-Meta is the category with the highest score. Users can also specify a threshold. In this way, a sequence with a highest score lower than the threshold will be labelled as "uncertain". In general, with a higher threshold, the percentage of uncertain predictions will be higher while the remaining predictions will be more reliable. For example, if you want to get reliable phage and plasmid sequences in the file "example.fna", you can take 0.7 as the threshold. Please execute:

./PPR_Meta example.fna result.csv -t 0.7 (Run by executable file)

or

PPR_Meta('example.fna', 'result.csv', '-t', '0.7') (Run by MATLAB script)

4.4. Run PPR-Meta over a large file (-b option)

If the RAM of your machine is small, or your file is very large, you can you -b option to let the program read the file in block to reduce the memory requirements and speed up the program. For example, if you want to let the program to predict 1000 sequences at a time, please execute:

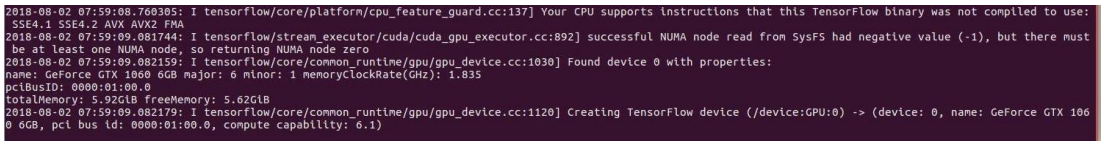
./PPR_Meta example.fna result.csv -b 1000 (Run by executable file)

or

PPR_Meta('example.fna', 'result.csv', '-b', '1000') (Run by MATLAB script)

The default value of -b is 10000.

Note: When running PPR-Meta, you can ignore the warning about the information of the CPU/GPU, as shown in the screenshot below.



5. Output

The output of PPR-Meta consists of six columns, representing “sequence header” (the same with the corresponding header in the fasta file), “sequence length”, “the probability that the sequence belongs to the phage”, “the probability that the sequence belongs to the chromosome”, “the probability that the sequence belongs to the plasmid” and “the possible source of the sequence”, respectively. Here is a screenshot of the output file:

	A	B	C	D	E	F
1	Header	Length	phage_score	chromosome_score	plasmid_score	Possible_source
2	phage1_source="Arthrobacter phage Mudcat"	7484	0.941588231	0.002345994	0.056065768	phage
3	phage2_source="Bacillus phage AR9"	8912	0.831541201	0.035068852	0.13338995	phage
4	phage3_source="Bacillus phage Aurora"	6983	0.730311051	0.181169275	0.088519651	phage
5	phage4_source="Bacillus phage Belinda"	1885	0.999458821	7.78E-06	0.000533381	phage
6	phage5_source="Bacillus phage DIGNKC"	4363	0.951073721	0.005356195	0.043570097	phage
7	phage6_source="Bacillus phage DirtyBetty"	3608	0.991743643	0.000850173	0.007406195	phage
8	phage7_source="Bacillus phage Eldridge"	5233	0.894087191	0.020961364	0.08495145	phage
9	phage8_source="Bacillus phage Nemo"	186	0.97037226	0.009645794	0.019981954	phage
10	phage9_source="Bacillus phage Nigalana"	3508	0.98406637	0.000495979	0.015437657	phage

Note:
The current version of PPR-Meta uses “comma-separated values (CSV)” as the format of the output file. Please use “.csv” as the extension of the output file. PPR-Meta will automatically add the “.csv” extension to the file name if the output file does not take “.csv” as its extension”

Part Ⅱ. Virtual machine version

Running PPR-Meta in the virtual machine is much easier for user who is not familiar with the command line, and the virtual machine can be installed in any PC. Note that the running time of the virtual machine version may be longer because the virtual machine could not speed up with GPU. We also modified the code to separate the input sequences into more batches to reduce memory requirements, which may increase the running time.

The following is the step by step guide to run PPR-Meta in virtual machine. Users can also refer to the brief video guide (http://cqb.pku.edu.cn/ZhuLab/PPR_Meta/Video_Guide.mp4). The version of the VirtualBox we used was 6.0.4.

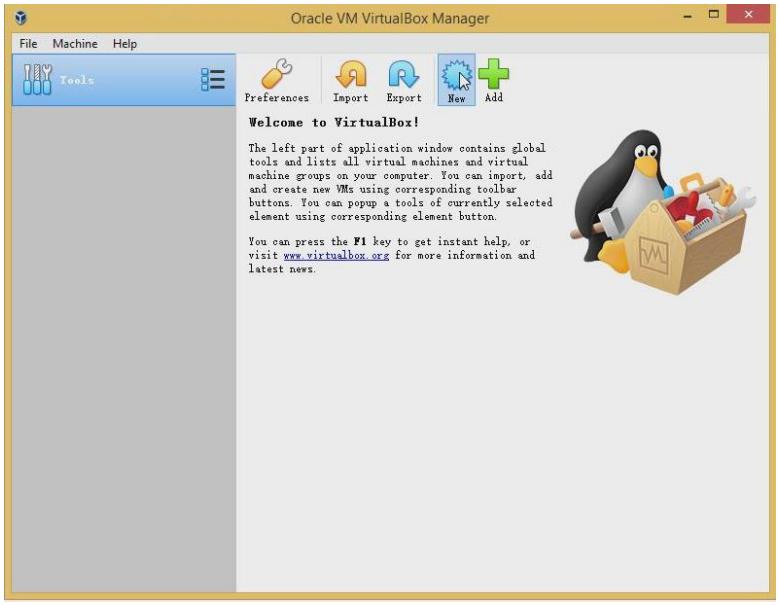
1. Install the virtual machine and run PPR-Meta in virtual machine.

Step 1: download the “VM_Bioinfo.vdi.7z” file from the PPR-Meta homepage (http://cqb.pku.edu.cn/ZhuLab/PPR_Meta/VM_Bioinfo.vdi.7z). The “7z” file can

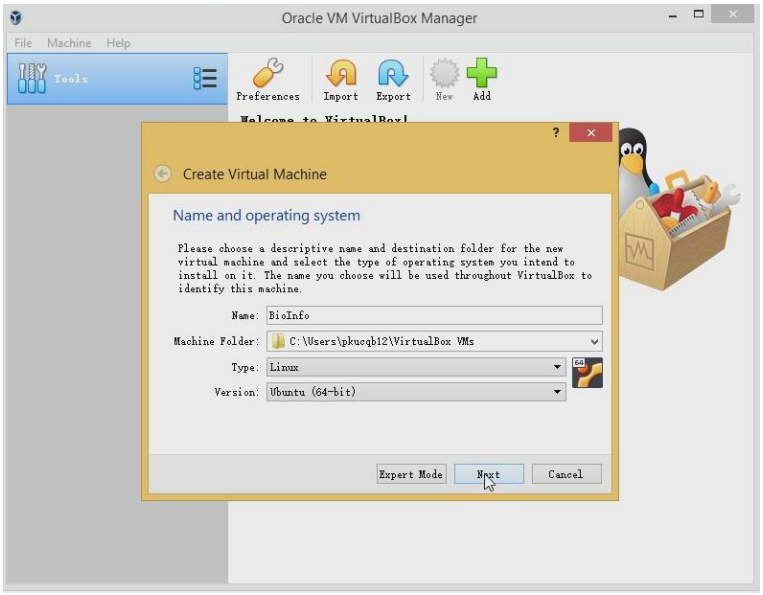
easily be decompressed using a current compressing software, such as “WinRAR”, “WinZip”, and “7-Zip”.

Step 2: download the VirtualBox software form https://www.virtualbox.org and install the VirtualBox. The VirtualBox is easy to install, you just need to select an installation folder and click the “next” button in each step.

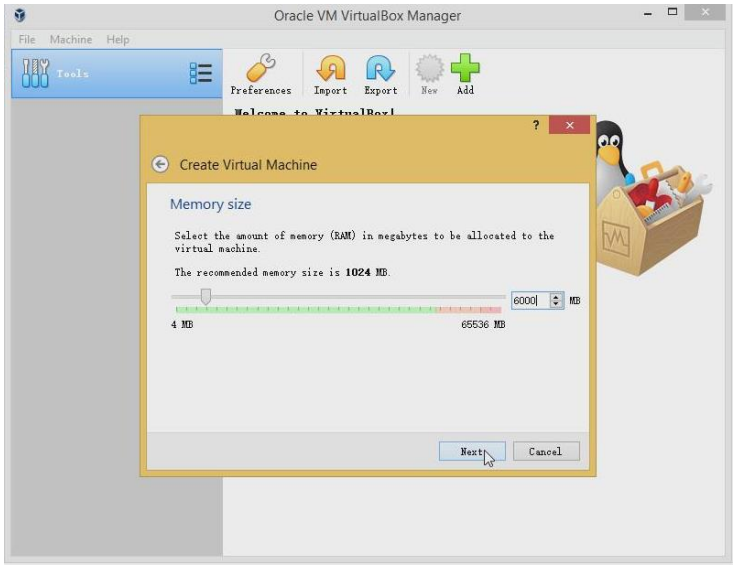
Step 3: Open VirtualBox, click the “New” button to create virtual machine.



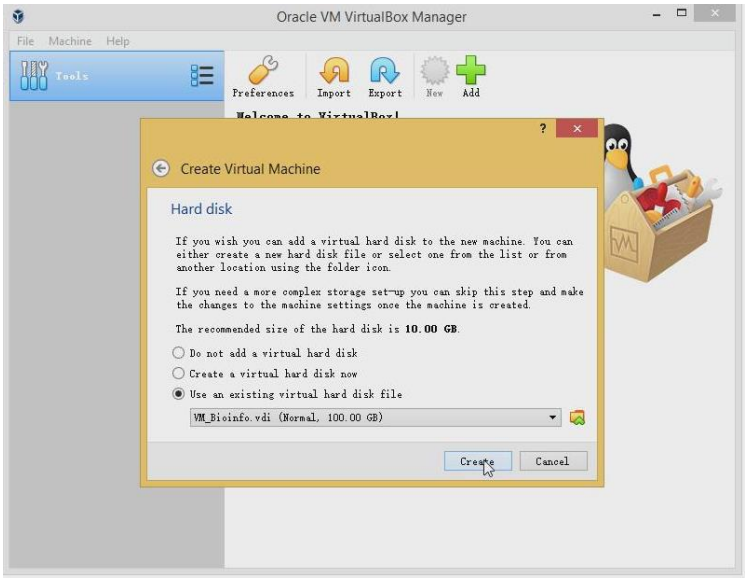
Step 4: Specify a name, select the “Linux” as the operating system and select “Ubuntu” as the version of the operating system. Then, click “Next”.



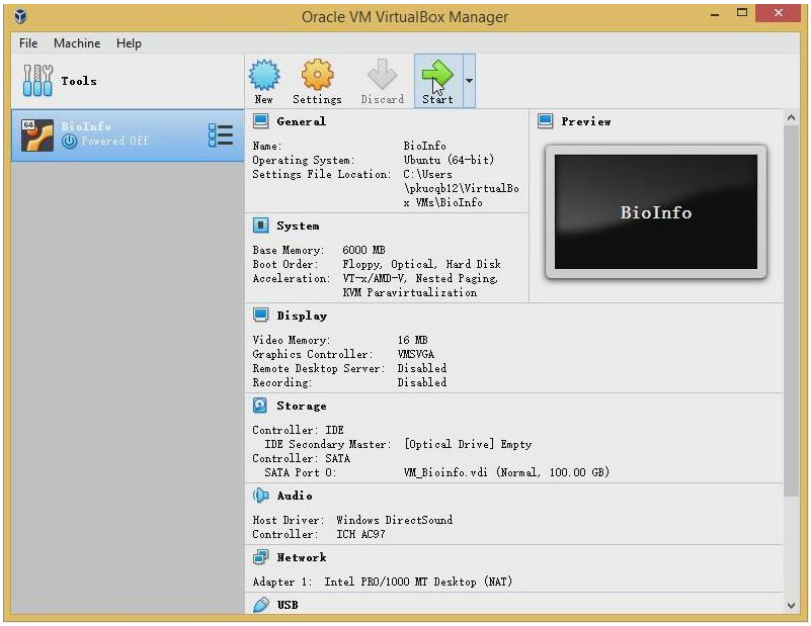
Step 5: If possible, allocate a larger amount of memory to the virtual machine. Click “next”.



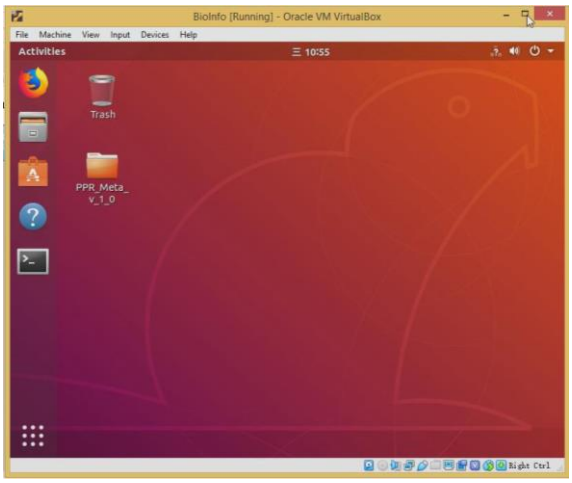
Step 6: Select “Use an existing virtual hard disk file”, and specify the “VM_Bioinfo.vdi” file downloaded from PPR-Meta homepage. Click “Create”.



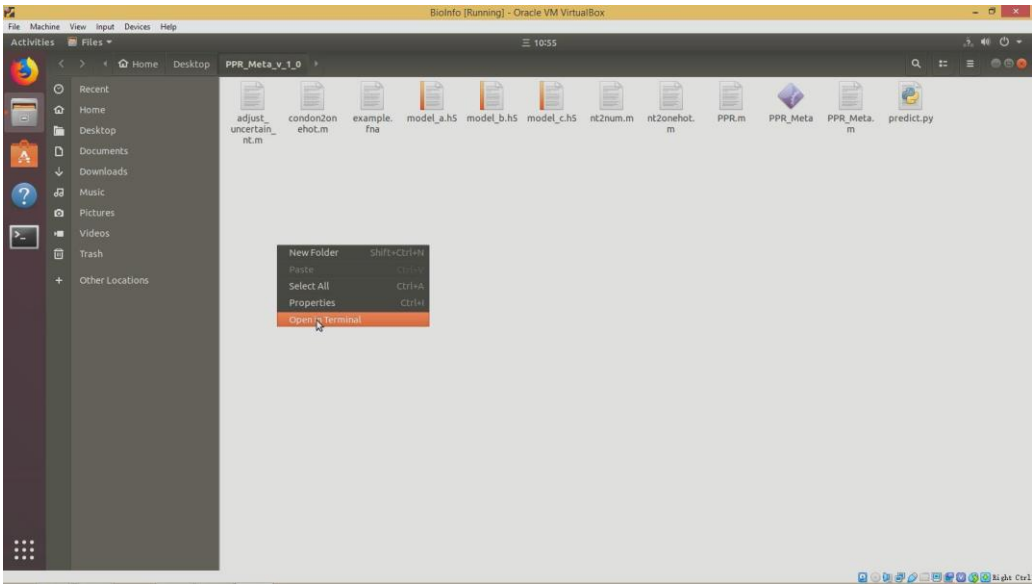
Step 7: Click “start” to open the machine.



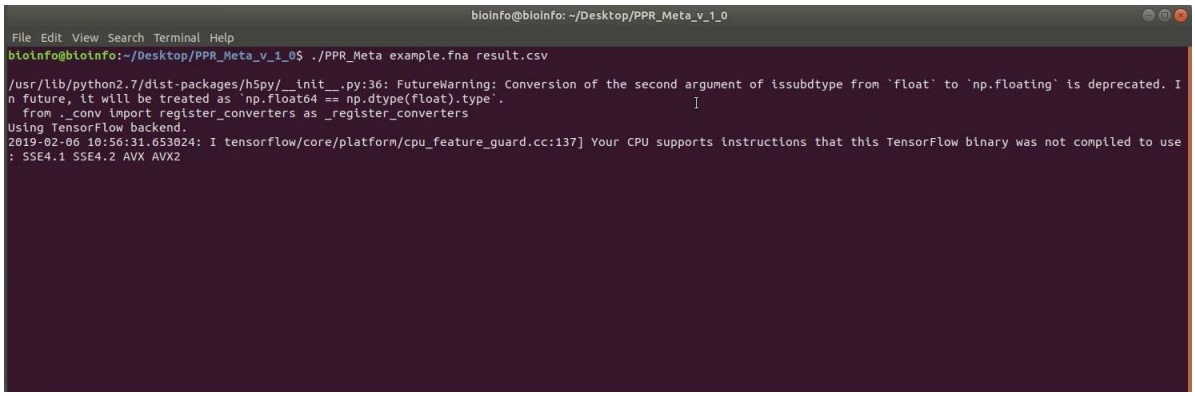
Step 8: The PPR-Meta is on the desktop.



Step 9: Go into the "PPR_Meta_v_1_0" folder, click the right click and open the terminal.

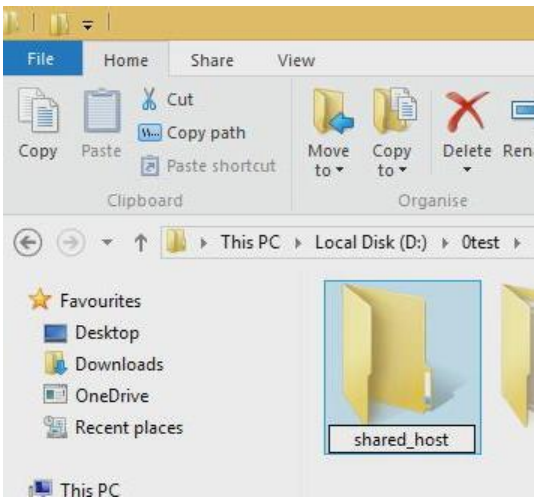


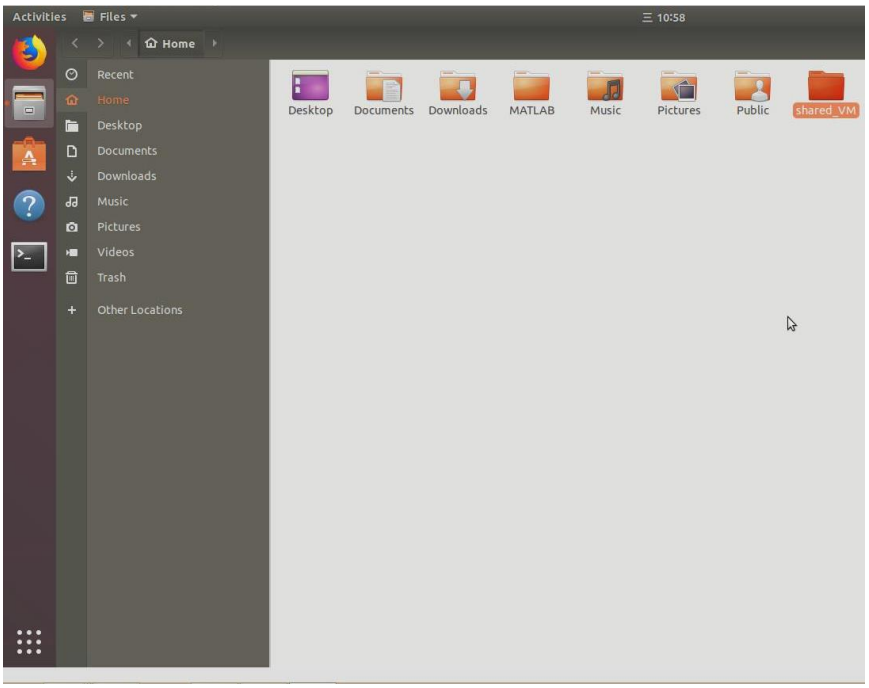
Step 10: Now you can run PPR-Meta (see also part Ⅰ， section 4.1 and 4.3). You can ignore the “FutureWarning” and the information about the CPU.



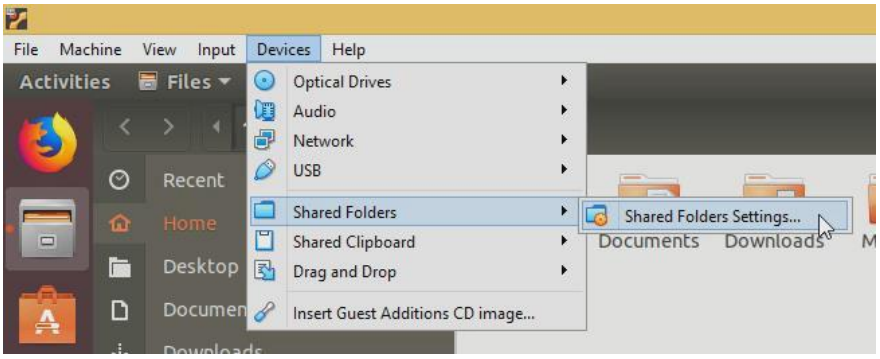
2. Exchange file between physical host and virtual machine.

Step 1: Create shared folders in both physical host (shared_host) and virtual machine (shared_VM).

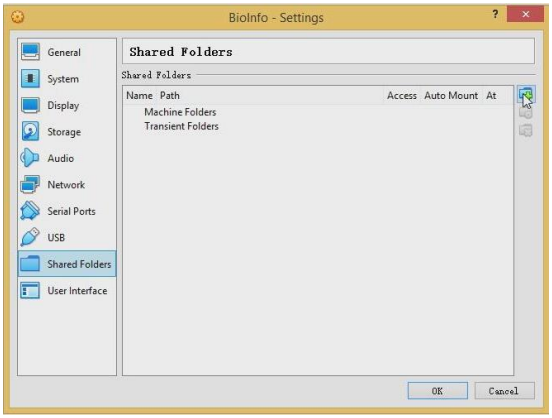


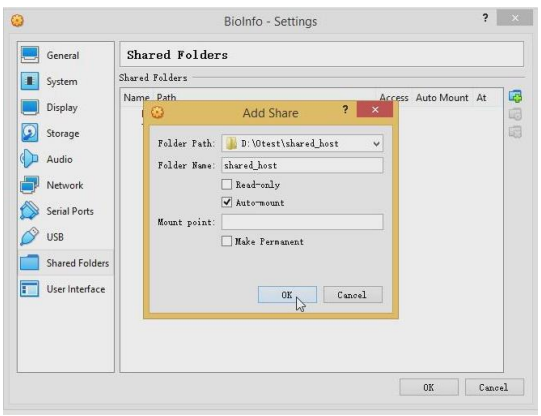


Step 2: In the window of VirtualBox, click “Devices”, “Shared Folder”, “Shared Folders Settings”.



Step 3: Add shared folder of the physical host and select “Auto-mount”.



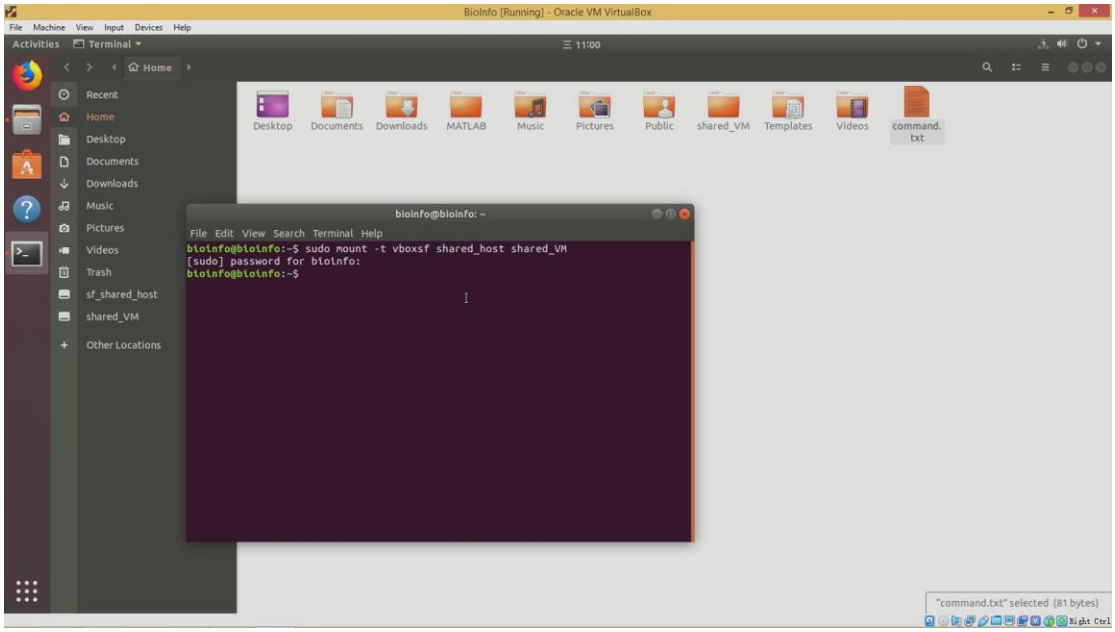


Step 4: Restart the virtual machine.

Step 5: Go to the parent folder of the shared folder in the virtual machine, click the right click and open the terminal. Copy the command in the "command.txt" file to the terminal:

sudo mount -t vboxsf shared_host shared_VM

Note, you should replace "shared_host" and "shared_VM" to the shared folders' name that you specify. **The password of the virtual machine is 1.**



Step 6: Now, you can exchange files between virtual machine and physical host. For example, you can copy the file in the virtual machine to the "shared_VM" folder, and this file will also exist in the "shared_host" folder in the physical host, vice versa. If you want to know more about the file exchange, click "Help" -> "Contents" -> "Guest Additions" -> "Shared folders" in the VirtualBox window for more details.

Note:

The version 1.1 of the program allows run mutiple tasks in parallel. However, running mutiple same tasks (with the same input file under the same '-t' and '-b' setting will throw error.