DeePVP manual

1 Overview

DeePVP is designed to identify and classify the phage virion protein (PVP). DeePVP contains a main module and an extended model. The main module can identify whether the given phage protein belongs to PVP, and the extended module can further classify the predicted PVP into the one of the 10 major PVP class: head-tail joining, collar, tail sheath, tail fiber, portal, minor tail, major tail, baseplate, minor capsid, and major capsid. The main module and the extended module of DeePVP can be run using an integrated pipeline or be run separately. For example, if researchers have already identified PVPs using other related methods, they can directly run DeePVP's extended module for PVP classification.

DeePVP can be run on the **virtual machine**, or via the **docker**. DeePVP take a "fasta" format file containing phage protein sequences as input, and output a tabular file.

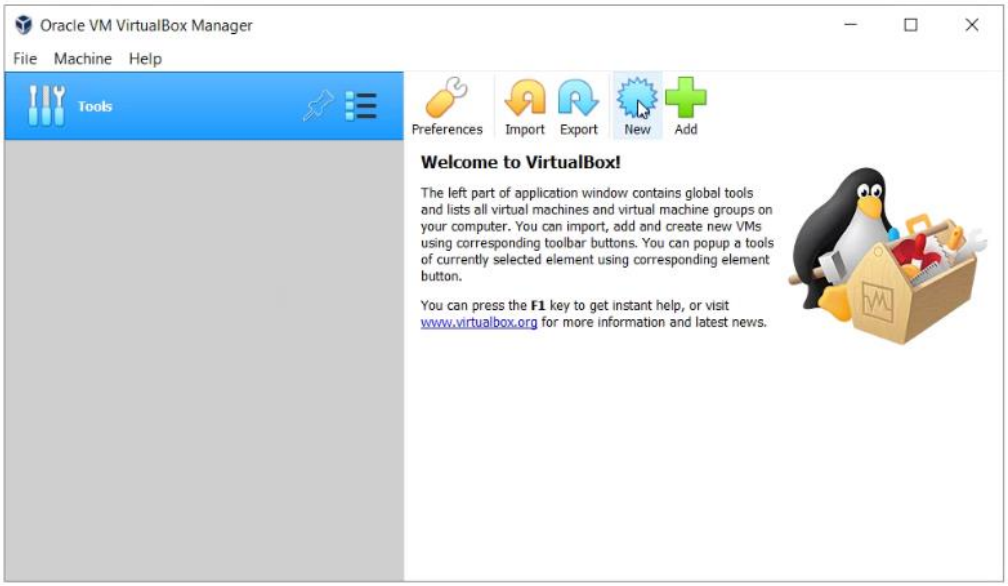
2 Virtual machine version

2.1 Installing the virtual machine of DeePVP

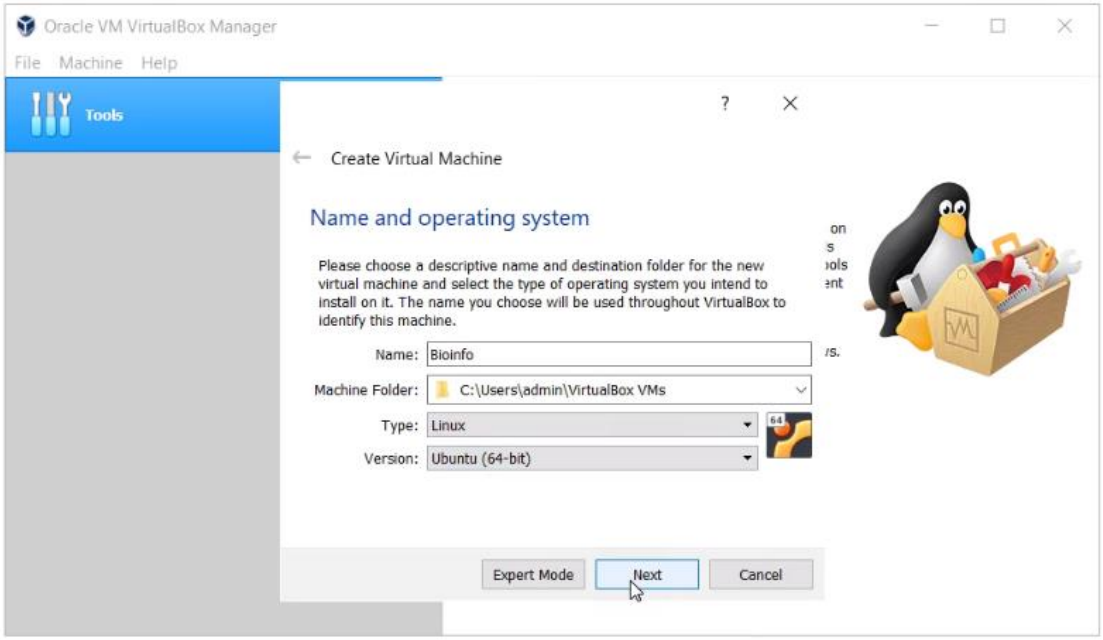
Step 1: Download the "DeePVP_VM.vdi.7z" file from https://www.dropbox.com/s/s7mipumr8f8tvi0/DeePVP_VM.vdi.7z?dl=0 or http://cqb.pku.edu.cn/ZhuLab/PPR_Meta/data/DeePVP_VM.vdi.7z. The "7z" file can easily be decompressed using current compressing software, such as "WinRAR", "WinZip", and "7-Zip". After decompressing the 7z file, a file named "VM_Bioinfo.vdi" will occur.

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.

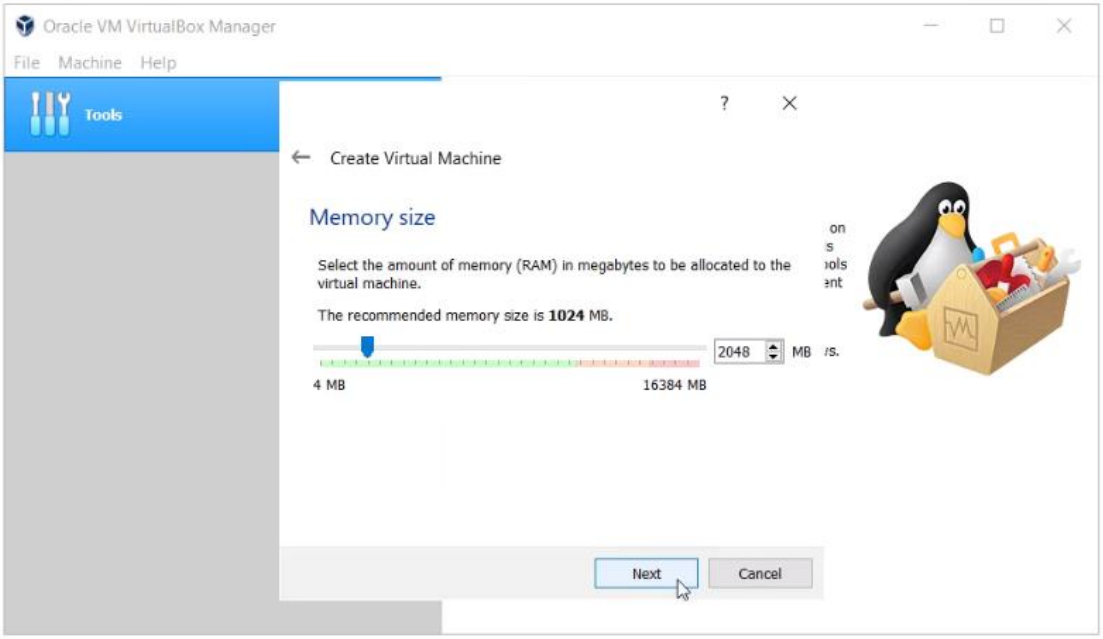
Step 3: Open VirtualBox, click the "New" button to create a virtual machine.



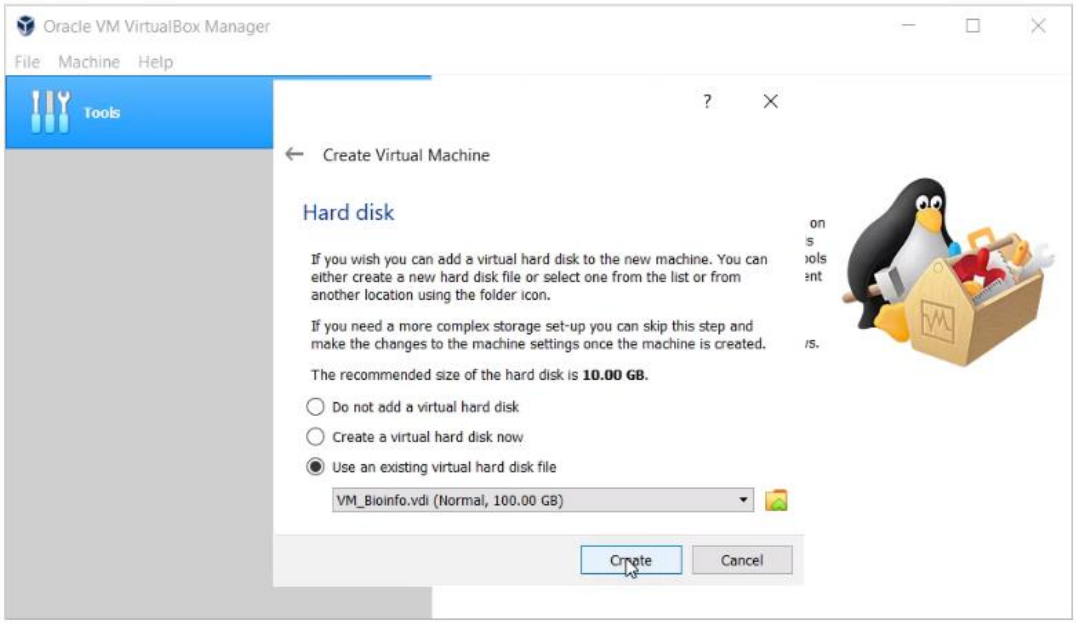
Step 4: Specify a name, select “Linux” as the operating system and select “Ubuntu” as the version of the operating system. Then, click “Next”.



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

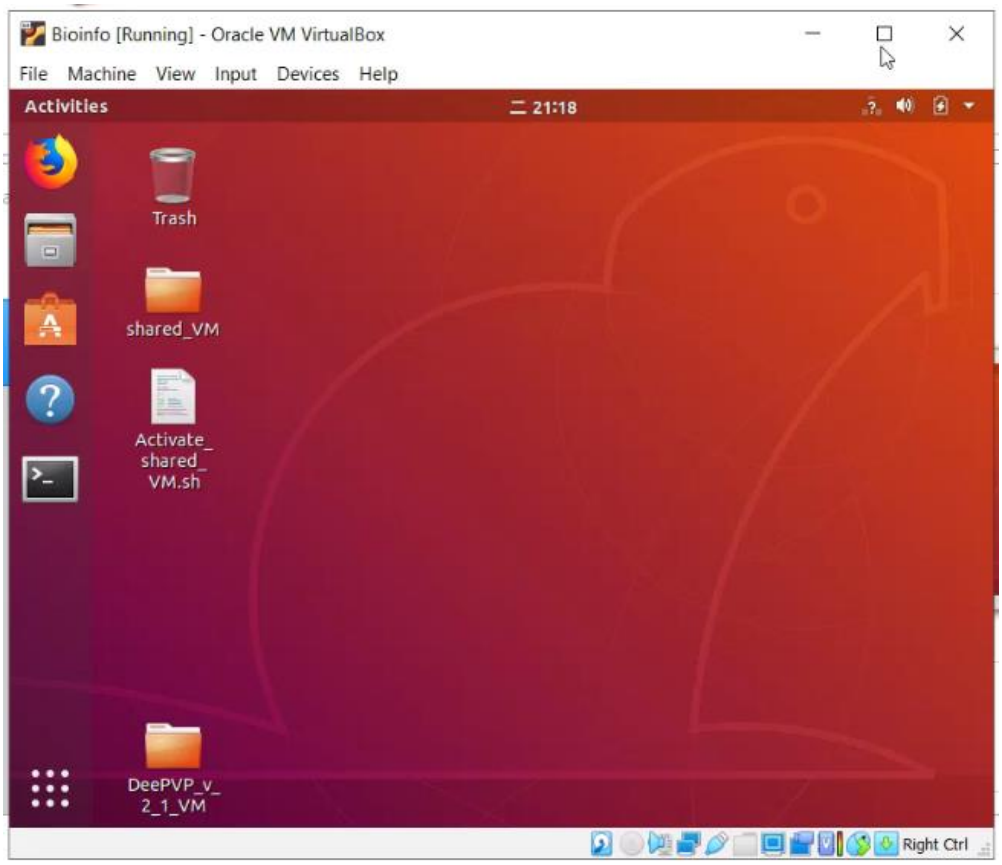


Step 6: Select “Use an existing virtual hard disk file”, and specify the “VM_Bioinfo.vdi” file.

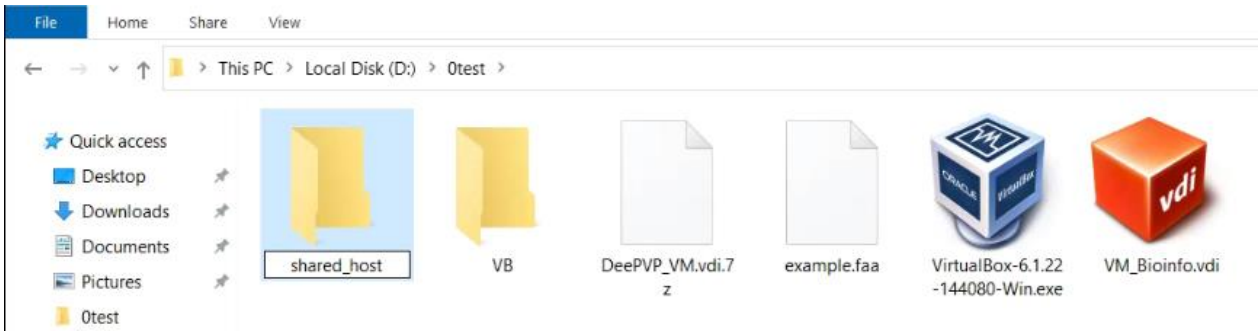


Step 7: Click “start” to open the machine.

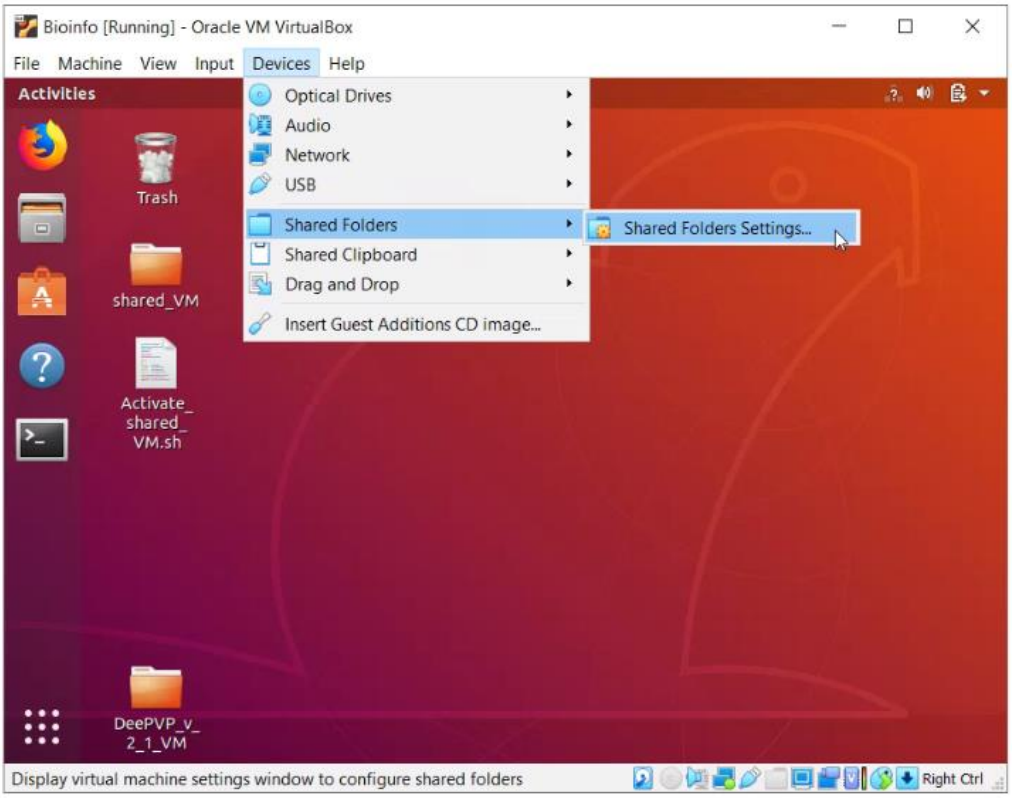
Step 8: The DeePVP folder is on the desktop.



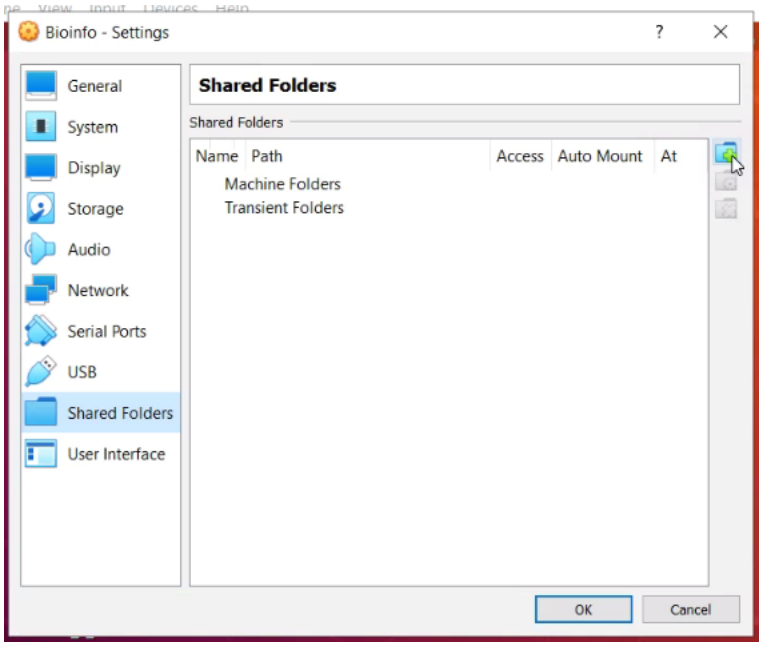
Step 9: Create a folder named "shared_host" in your physical host (not in the virtual machine, and do not change the folder name). This folder is created for the file exchanging between the physical host and the virtual machine.

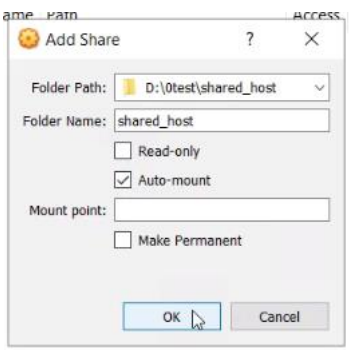


Step 10: In the window of VirtualBox, click “Devices”, “Shared Folder”, “Shared Folders Settings”

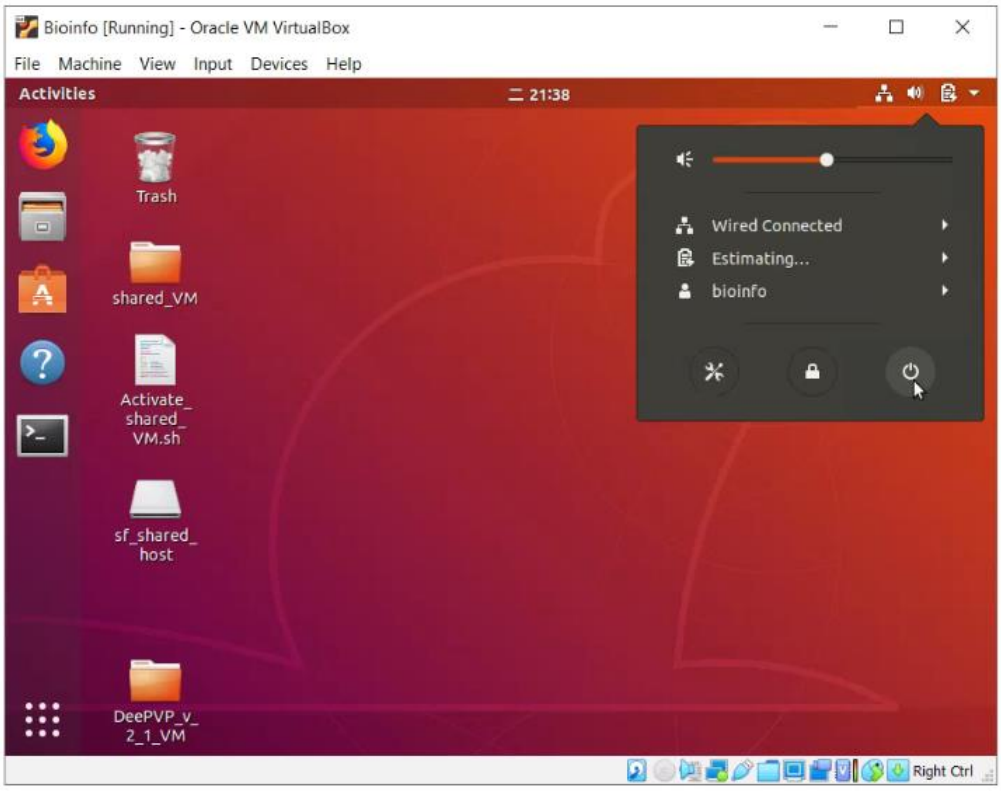


Step 11: Click ‘+’ to add the “shared_host” folder of the physical host, and then select “Auto-mount”.

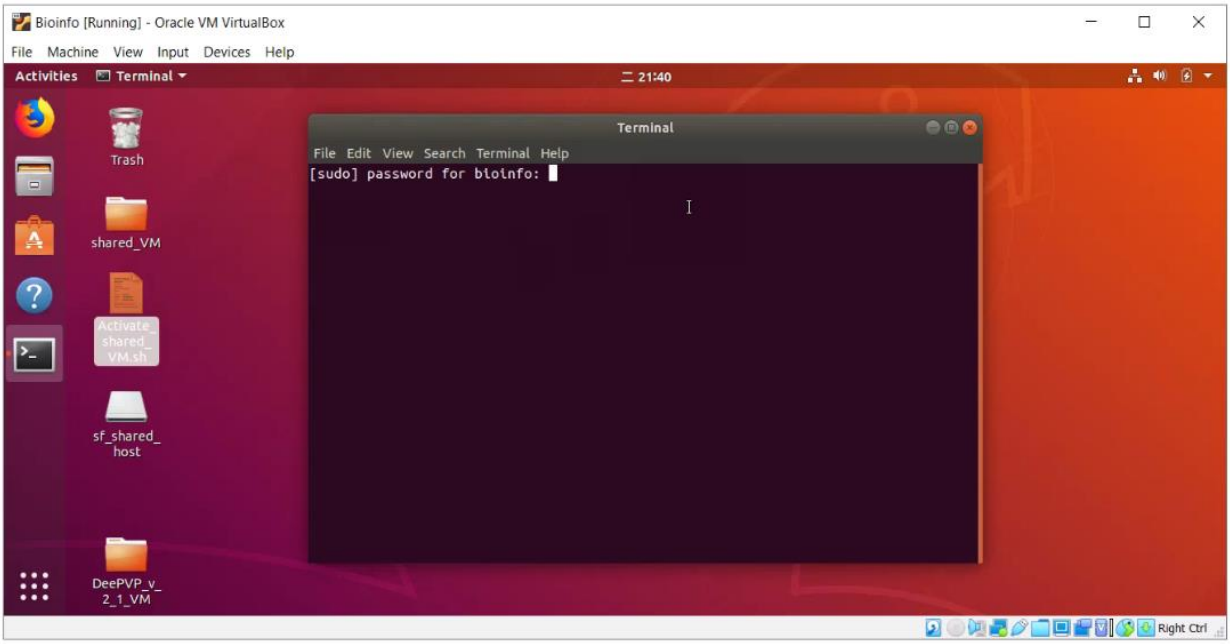




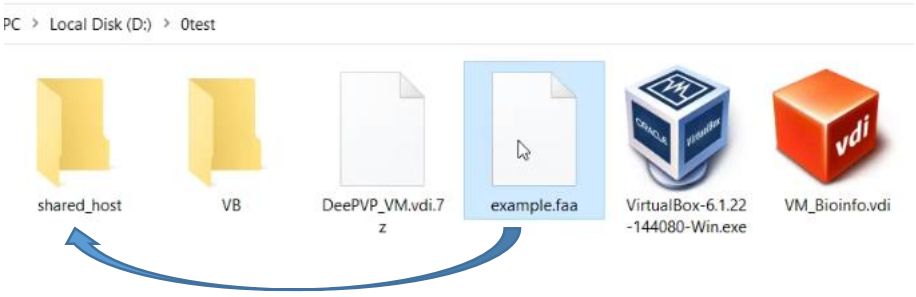
Step 12: Restart the virtual machine.



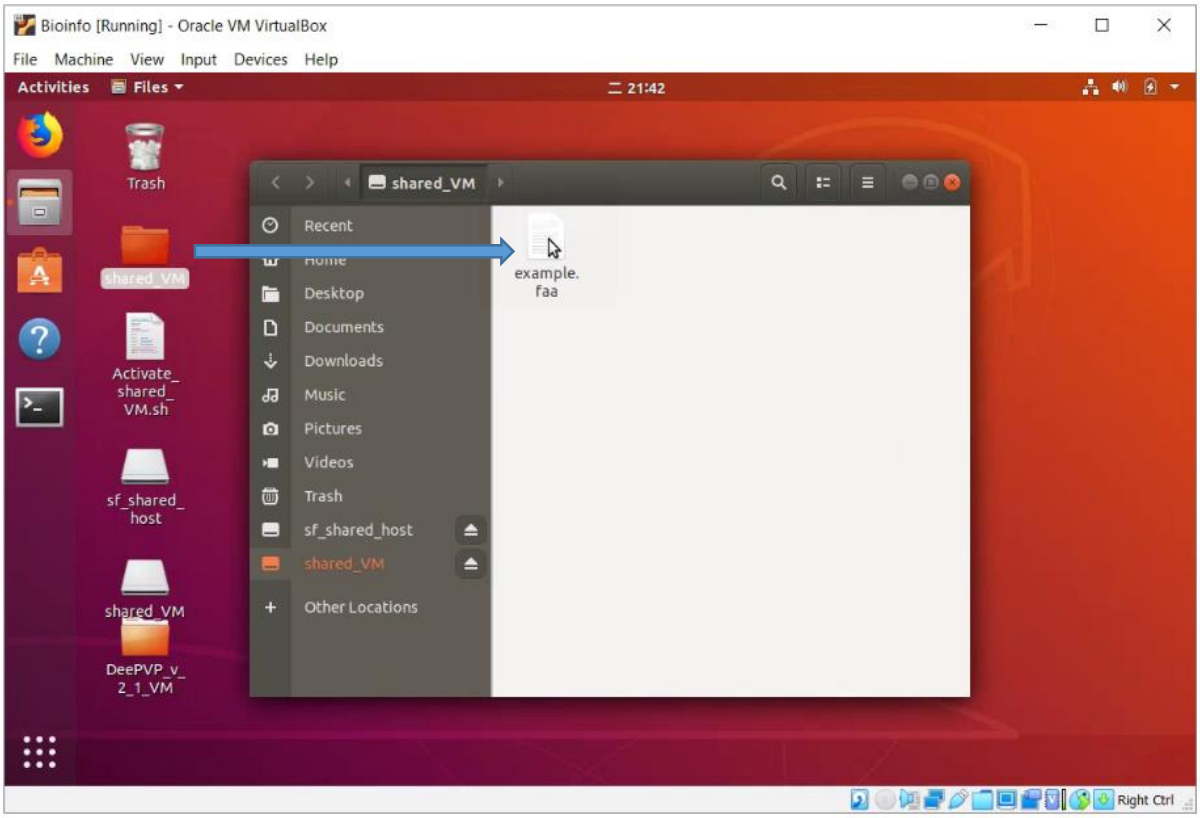
Step 13: Double click the “Activate_shared_VM.sh” file on the desktop of the virtual machine. When prompted for a password, enter "1" and hit the "Enter" key. (Note, no prompt will occur when you enter the password).



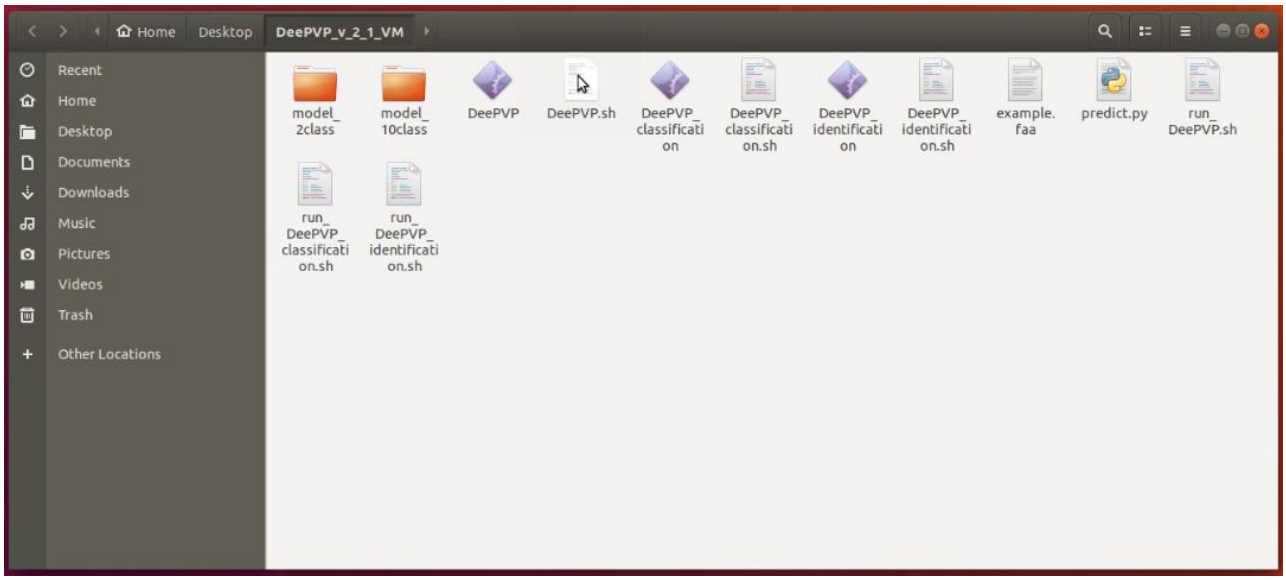
Step 14: In the physical host, copy your input file (fasta format) to the “shared_host” folder.



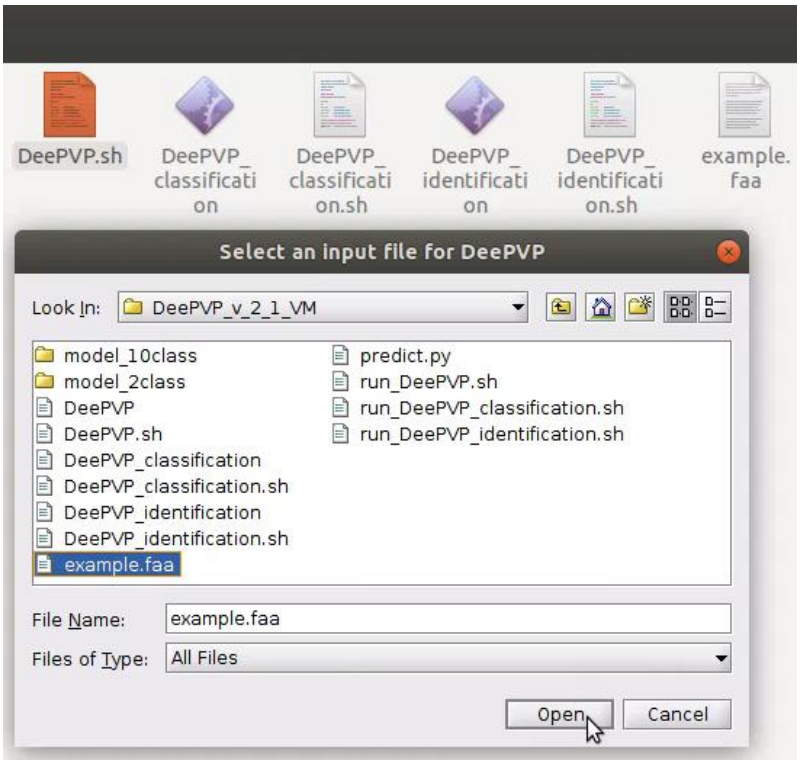
Step 15: Then the file will also occur in the “shared_VM” folder in the desktop of the virtual machine. You can copy the file to other folders in the virtual machine.



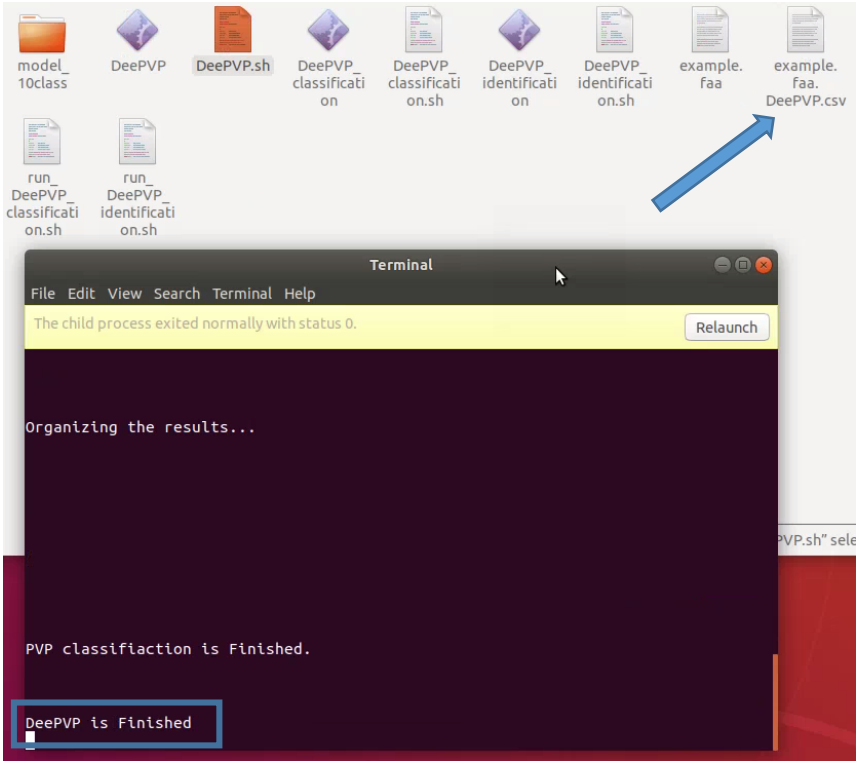
Step 16: To run the DeePVP, go into the DeePVP folder in the desktop, double click the "DeePVP.sh" file.



Step 17: Please wait for a few seconds. Then, select your input file and click the "Open" button, and the program will start running.



Step 18: When the terminal display “DeePVP is Finished”, you can close the terminal. The output file will occur under the same folder with the input file. The suffix of the output file is “.DeePVP.csv”. You can copy the output file to the “shared_VM” folder and the file will occur under the “shared_host” folder of the physical host.



To run the main module of DeePVP independently, you can double click the “DeePVP_identification.sh” file. To run the extended module of DeePVP independently, you can double click the “DeePVP_classification.sh” file. The suffix of the output file of the main module is “.DeePVP_main.csv”, and the suffix of the output file of the extended module is “.DeePVP_extended.csv”.

3 Docker version

3.1 Installing the docker

Step 1: Users can download and install Docker on Windows, Mac and Linux platforms from https://docs.docker.com/get-docker/. You can see the following options and click a corresponding platform for you. After switching to detailed installing instructions, you can install a docker. Especially, if you are a non-root user on Linux platform, you should add yourself to a docker group so that you can have access to running docker (see https://docs.docker.com/engine/install/linux-postinstall/). To test whether Docker has installed correctly, you could open your terminal and run:

docker run hello-world

If no error messages, the docker has been successfully installed.

Step 2: After installing the docker successfully, you can start to use the docker version of DeePVP in a terminal. Open a new terminal on your computer. Then change the path to the location that you want.

Step 3: Pull down the image from Docker Hub by running the command:

docker pull shufangwu/ppr-meta:1.0

Step 4: Obtain all the local docker images that you have by running the command:

docker images

```
[sfwu@localhost ~]$ docker images
REPOSITORY          TAG                 IMAGE ID            CREATED             SIZE
shufangwu/ppr-meta  1.0                 dea8c47cd82b        8 days ago          4.91GB
```

Step 5: Start the docker container based on downloaded images by running the command:

docker run -it shufangwu/ppr-meta:1.0 bash

Then you can enter the docker container and begin to use DeePVP.

```
[sfwu@localhost ~]$ docker run -it shufangwu/ppr-meta:1.0 bash
root@844a99829241:/#
```

3.2 Reenter the container and transfer files between your host and the docker container that was used before.

Step 1: If you want to exit from a current using container and stop this container, you can just type ‘exit’ in the command line and then click ‘Enter’ on the keyboard. While you just want to exit and keep the container running, you can press Ctrl+P+Q. Then you can come back to your host.

Step 2: Obtain the information about all the available containers:

docker ps -a

Step 3: You need to pay attention to 'CONTAINER ID' and 'STATUS' columns. For example, the container (CONTAINER ID: 844a99829241)'s STATUS is 'Exited', which means this container is stopped when using the 'exit' command to exit. However, the container (CONTAINER ID: ebb8039036ec)'s STATUS is 'Up', which means this container is running when using 'Ctrl+P+Q' to exit. The running containers could transfer files between themselves and the host while the stopped containers could not. You can also rerun a stopped container by:

docker start 844a99829241

('844a99829241' is the CONTAINER ID of the stopped container)

Step 4: Transfer input_file.faa file from your host to the container by running the command:

docker cp <Folder_path_on_host>/input_file.fna CONTAINER_ID:<Destination_path_on_container>

For example, you can use the command in your host:

docker cp ./input_file.fna ebb8039036ec:/home/DeePVP

('ebb8039036ec' must be a running CONTAINER ID that you want to use)

Transfer output.csv file from the container to your host by running the command:

docker cp CONTAINER_ID:<Foder_path_on_container>/ output.csv <Folder_path_on_host>

For example, you can use the command in your host:

docker cp ebb8039036ec:/home/DeePVP/output.csv ./

('ebb8039036ec' must be a running CONTAINER ID that you want to use).

Step 5: After transferring an input file to the container, you can reenter this container and then use DeePVP to make a prediction. A running container can be reentered. If the container's statute is 'Exited', you should first start it by using the command:

docker start CONTAINER_ID

Then you can enter this running container by command:

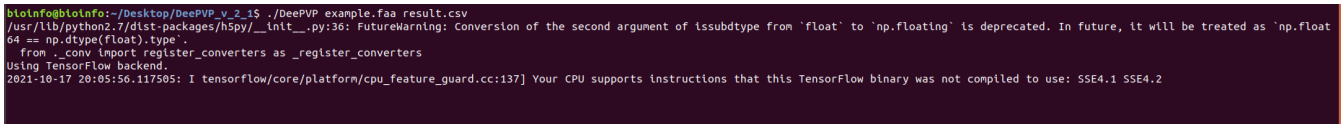
docker attach CONTAINER_ID

(or **docker exec -it CONTAINER_ID bash**)

You will enter the container and start to make a prediction

Step 6: Under the DeePVP folder, user can directly type the following command to run the integrated pipeline:

./DeePVP example.faa result.csv



The “example.faa” should be replaced by the input file’s name containing the phage protein sequences in the “fasta” format, and the “result.csv” should be replaced by the output file’s name specified by the user.

To run the main module of DeePVP independently, user can type the following command:

./DeePVP_m example.faa result_m.csv

To run the exanted module of DeePVP independently, user can type the following command:

./DeePVP_e example.faa result_e.csv

Note: When running the program, some warning massage about the device may occur, and users can ignore the massage.

4 Output

4.1 The output of the integrated pipeline

When running the integrated pipeline of DeePVP, the output file contains 14 columns:

The 1st column contains the headers of the sequences in the input file.

The 2nd column contains the PVP scores (between 0 and 1) calculated by the main module of DeePVP.

The 3rd column contains the prediction results of the main module. By default, the proteins with the PVP scores higher than 0.5 are regarded as PVP, while the others are regarded as non-PVP.

The 4th-13th columns contain the 10 likelihood scores representing the probability of the proteins belonging to the major capsid, minor capsid, baseplate, major tail, minor tail, portal, tail fiber, tail sheath, collar, and head-tail joining. The sum of these 10 scores is equal to the PVP score.

The 14th column contains the PVP class predicted by the extended module of DeePVP. By default, the class with the highest score will be served as the prediction. It is worth noting that these predictions will not make sense if the corresponding proteins are not predicted as PVP by the main module.

4.2 The output of the main module

When running the main module of DeePVP independently, the output file contains 3 columns:

The 1st column contains the headers of the sequences in the input file.

The 2nd column contains the PVP scores (between 0 and 1) calculated by the main module of DeePVP.

The 3rd column contains the prediction results of the main module. By default, the proteins with the PVP scores higher than 0.5 are regarded as PVP, while the others are regarded as non-PVP.

4.3 The output of the extended module

When running the extended module of DeePVP independently, the output file contains 13 columns:

The 1st column contains the headers of the sequences in the input file.

The 2nd-12th columns contain the 10 likelihood scores representing the probability of the proteins belonging to the major capsid, minor capsid, baseplate, major tail, minor tail, portal, tail fiber, tail sheath, collar, and head-tail joining. The sum of these 10 scores is equal to 1.

The 13rd column contains the PVP class predicted by the extended module of DeePVP. By default, the class with the highest score will be served as the prediction.