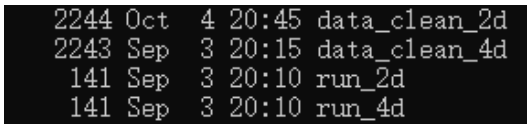
Data preparation:

Utilizing Ubuntu windows (bash), to clean and prepare the data.

1. Folder *Original_data*
Save all the original datasets which offer by clients.

2. Folder *Clean_data*
Data:
Each patient has 8 datasets, which are activated or resting TCR record in 4 time points, a day before vaccinating and a week after vaccinating in 2016 and 2017.
After cleaning the data, each cdr3 correspond to its V-gene name, counts, and frequency. (cleaned data in sub-folder '2-digits' and '4 digits', as the client's request, the data cleaned into two versions, 2 and 4 digits which represent the digits number of V-Gene name.)

If test needed:



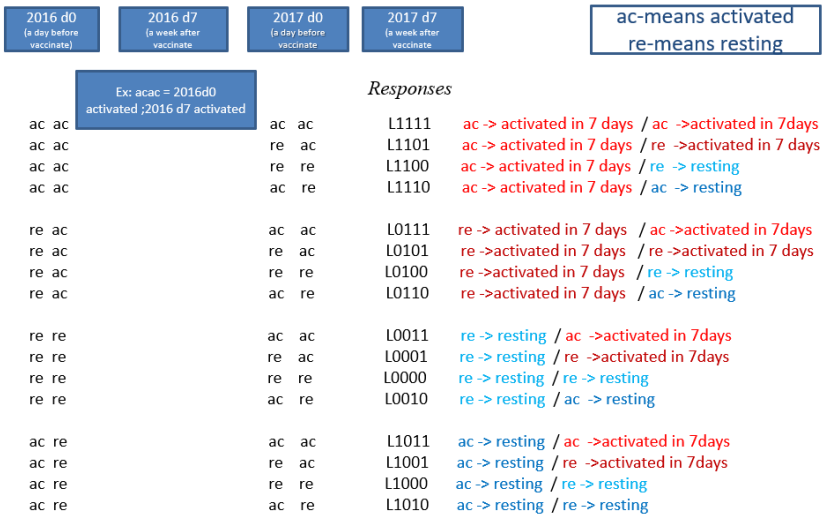
There are 4 bash script in folder Clean_data.

Just type **(./run_2d new_folder**), notice the first parameter is new folder name, don't use the already existing one.
The script will create a new folder, all cleaned data would save in that folder.

Same as 4 digits.

PS: if meet the permission limited.

3. Folder *data_preparation*



FigureA: Label activated TCR as'1', and resting as'0'.

Data:

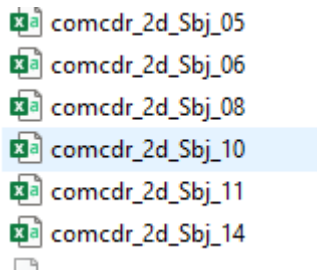
Then Intersect data sets (common cdr3 as the main key), we get 16 kinds of responses (FigureA). In joined dataset (common cdr3), except cdr3 and V-gene names, we have 8 attributes columns which are 4 time points' TCR count and frequency, and a target column which is the response.

If test needed:

In folder common cdr3, for example in the sub-folder 'VGene-2digits', which include cleaned data and several bash script.

```
VGene-2digits$ ./cmcdr 06 2016 2017 2d
VGene-2digits$ ./cmcdr 08 2016 2017 2d
VGene-2digits$ ./cmcdr 10 2016 2017 2d
```

While we just need to run script 'cmcdr', the **first parameter** is the patient number, **second** is the year 2016, **third** is the year 2017, **fourth** is 2d which represents handle 2 digits V-Gene datasets.



The common cdr3 dataset would appear in the same folder, Which would be used in machine learning and analysis.

```
VGene-4digits$ ./cmcdr 06 2016 2017 4d
VGene-4digits$ ./cmcdr 08 2016 2017 4d
VGene-4digits$ ./cmcdr 10 2016 2017 4d
```

Generate 4 digits datasets is the same.