I am doing a project that I need to train a model. By collecting over 10000 viral genomes (ACGT) from NCBI, then transform each the genomes into 3rd Markov model transition matrices, then using jupyter notebook with python, and using Neural Network (NN). By labeling the taxonomy for each virus, which belong to specific family, and train them with NN, and 80% for training and 20% for testing, and then after training, I want to check the accuracy of my model, and finally, I want to input a new viral genome, and my model can distinguish which taxonomy it belong to (family).

I already download about 17000 viral genomes and the location is “/Users/kenneth/Desktop/fyp\_doing/sequences.fasta”  
and each virus the content is like this:  
>Phenuiviridae

ACACAAAGGCGCC……

----

I am doing a project that I need to compare two virus genomes (ACGT), and by using 2nd Markov model, each virus genome will become a matrix that each entry is kind of probability (0 to 1).

If I want to compare two transition matrices (each entry range from 0 to 1, and each row sum up to 1), and if I want to find out their similarity is high or low, what approach is good for me?  
  
remarked that I want to simply compare the similarity of row to row. So, no need considers current state or next state or long run or as a whole.  
  
Total Variation (TV) distance  
Cosine Similarity  
Kullback–Leibler (KL) divergence  
Jensen–Shannon (JS) divergence  
Hellinger Distance  
Frobenius Norm

I am using Visual Studio Code with python, and I am doing a project topic that need to compare transition matrix of virus genome sequence by 1st order to 5th Markov order to find out their matrix similarity. Compare the virus 1 by 1, so 1 matrix compares to another matrix.

Now I am already downloaded 1000 virus data from NCBI in fasta format.

I want to have 3 kinds of comparison:

1. Comparing the virus with same Species
2. Comparing the virus with same Genus but different Family
3. Comparing the virus with same Family but different Genus

From 1st order to 5th order and by using these two approaches:

Total Variation (TV) distance  
Jensen–Shannon (JS) divergence

For example, 2 viruses with same Species comparing their transition matrix in 1st order, and the approach will be using TV distance. And because there are many comparisons for each 2 viruses out of 1000 viruses, so will be take overall average.

So, finally will be print out total 30 results, because 3 kinds of comparison times 5 different orders times 2 approaches.

3\*5\*2=30  
-----  
The virus data of the fasta file will like this:  
>Betanucleorhabdovirus rhododendri |Betanucleorhabdovirus|Rhabdoviridae

TATCACTTA……

So, the virus data is like:  
>Species|Genus|Family

genome sequences

However, some virus may miss family or genus or species name, for example, this virus is missing family and genus name, then you should skip this kind of virus. i.e. only consider the data that family, genus, species all exist in the beginning.

>Wuhan heteroptera virus 3 ||

CCCAGGTTCAATATAGTATTT

-----

I am using Visual Studio Code with python, and I am doing a project topic that need to compare transition matrix of virus genome sequence by 1st order to 5th order Markov Model to find out their matrix similarity. Compare the virus 1 by 1, so 1 matrix compares to another matrix. Please give me whole python code in one.

Now I am already downloaded 1000 virus data from NCBI in fasta format. The file path is “/Users/kenneth/Desktop/fyp\_doing/compare.fasta”

Using Total Variation (TV) distance and Jensen–Shannon (JS) divergence to compare:

Virus with all different (Different Family, different Genus, different Species)

Virus with all same (same Family, same Genus, same Species)

and because there are many comparisons for each 2 viruses out of many viruses, so will be take overall average.

-----  
The virus data of the fasta file will like this:  
>Species|Genus|Family

genome sequences

However, some virus may miss family or genus or species name, for example, this virus is missing family and genus name, then you should skip this kind of virus. i.e. only consider the data that family, genus, species all exist in the beginning. For example:

>Wuhan heteroptera virus 3 ||

CCCAGGTTCAATATAGTATTT

So, some virus data is like:  
>Species|Empty|Empty

genome sequences

----

I am using Visual Studio Code with python, and I am doing a project topic that need to compare transition matrix of virus genome sequence by 3rd order Markov order to find out their matrix similarity. Compare random two virus. Please give me whole python code in one.

Now I am already downloaded 1000 virus data from NCBI in fasta format. The file path is “/Users/kenneth/Desktop/fyp\_doing/compare.fasta”

Using this approach: Total Variation (TV) distance

The virus data of the fasta file will like this:  
>Betanucleorhabdovirus rhododendri |Betanucleorhabdovirus|Rhabdoviridae

TATCACTTA……