Homework 3

made by German Ilya

1. Data

We get 9 different genomes of coronavirus SARS-CoV-2 for people from different countries. And one extra gemone of SARS-CoV-1.

As in homework №2 we find each genome on site of National Center for Biology Information and adding them to only file that MEGA will work with. (I made one file for each country and then merged it by `cat country_name.fasta >> merged.fasta`, the same thinks I made with SARS-CoV-1.)

You can find files here

And merged file here

2. Alignment

Then we have a file, that we can use in mega. The process of alignment of sequences the same as in prevous hw, so it's not necessary to show it here.

The result if pretty good: the sequences are really similar. Especially the sequences of SARS-CoV-2.

Some screens you can find here

3. Trees

We constructed trees by different methods (as we did it in hw2).

You can find trees here.

All of them are cohere that the man from USA was infected first. But there is a divergence, who was the last. So, we follows NJ tree and finds mutation between coronavirus from usa-man and morocco-man.

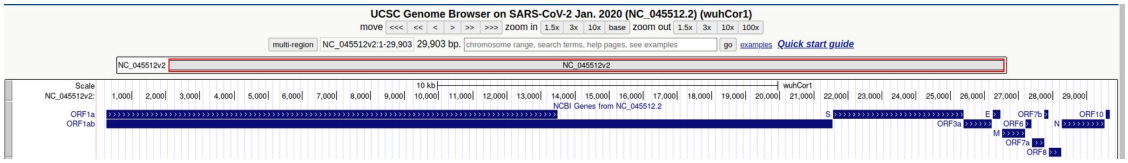
4. Mutations

We have chosen USA and Morocco for comparisong. To understand in which genes the mutation was, we find differences in sequences and remember their numbers.

Screens_here

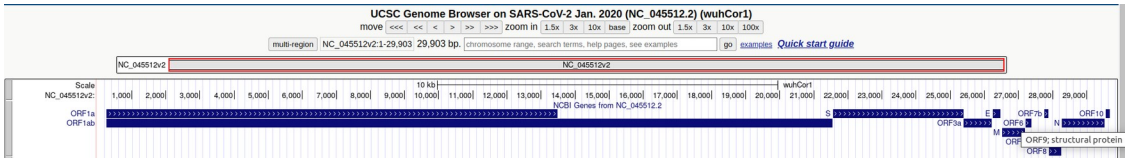
(Numbers: 9565, 10060, 10229, 15745, 28477)

Finally, we can open UCSC Genome Browser where we can find in which genes our mutations were.



According to this information, we can define genes:

9565, 10060, 10229 and 15745 - in ORF1ab gene 28477 - in N gene (but it also descripted as ORF9a if we put mouse on it).



Interesting mini-research.

The interesting thing is that the last mutation is actually removing some nucleobases. Maybe, this part of genome has no useful functions, so it can be removed by evolution process. (This is not an isolated case, so you can find the same case in extra-screens)

Another interesting thing is that the majority of mutation was in the place 25000+ in sequences.

I show another situation, because I thought that as closer mutations are, as much possible that they are in one gene. So, I tried to find mutation in different parts of gemone. In fact, the situation is opposite.

To explain that, the gene ORF1ab carries the main info about coronavirus. It matches big part of sequence, but it is more stable and more resistance to mutation.

Screens, prove my explanetion.