Lab Exercise 5

1)

The file **viralzone.csv** contains taxonomic and genomic data of ~450 virus strains downloaded from the ViralZone database. The genomic data of each record is present in JSON format within the CSV (yes, one data format wrapping another). The genome is a list, where each element in the list (usually just one) represents a segment of the virus's genome. Inside the JSON you can also find the protein products that are coded by the genome, along with their amino-acid sequences. Unlike cellular organisms, viral genomes usually contain at most a few dozens of genes.

The virus strains **Human papillomavirus 1** and **European elk papillomavirus** are both part of the **Papillomaviridae** family (but of different genera). Find the pair of proteins coded by their two genomes, one of each, that most resemble each other. Of course, defining resemblance is somewhat arbitrary, meaning you will have to make tough choices.

2)

A)

Download all the **reviewed** human proteins from UniProt in XML format.

B)

Use Biopython to parse it.

What are the 5 most common subcellular locations of human proteins according to UniProt?

C)

Find the reviewed human protein with the highest proportion of helix secondary-structure.