Lab Exercise 3

1)

Use **herpesvirus_genome.json**.

A)

Find the frequency of each amino-acid in the herpesvirus's proteome.

B)

Find the frequency of each amino-acid in genes falling into each of the three categories: **envelope**, **membrane** and **capsid** (i.e. where the name of the protein product contains the relevant word). In each category, find the two amino acids with the most extreme frequencies (both highest and lowest) compared to the background frequencies in the entire proteome of the virus.

2)

The file **viperdb.csv** contains summary of structural data about viruses downloaded from the VIPERdb database. Each row is an analysis of a PDB (Protein Data Bank) record based on crystallography. Each row is associated to a specific viral genus (and a family, which is a higher taxonomy group). The size of the virus's capsid, which is roughly spherical in those viruses, is described by inner, outer and average radiuses (measured in Angstrom units).

A)

For each genus, find the record with the highest outer radius.

B)

At the highest level, viruses are categorized by their genome type, which can be either DNA or RNA, and can be either single-stranded (ss) or double-stranded (ds). For example, dsRNA refers to viruses with a double-stranded RNA genome.

For each of the four groups (ssRNA, dsRNA, ssDNA and dsDNA), find the average surface area of the capsids (described by the "Outside SASA" field). Take into account only the biggest record of each genus that you found earlier (ignore all other records).