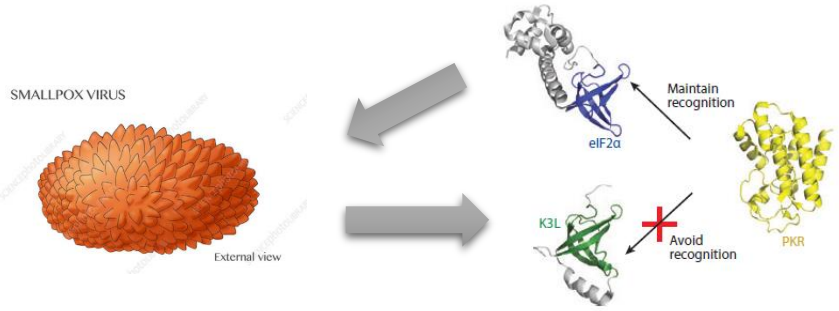
Disorder regions Prediction in Host-Virus domain – motif interactions

Overview:

- ELM- Eukaryotic Linear Motif - are a class of protein interaction interface that are central to cell physiology. they most common in **disorder region** that allows them to be accessible for protein–protein interaction. (Kumar et al., 2020).
- These motifs are pervasively mimicked by virus through convergent evolution to hijack and deregulate host cellular functions.
- As a result of that the host domain proteins are in kind of conflict:
  in one hand the want to be in contact with their own protein that contain the motif (ELM) that important to function regulation, and on their other hand they want to avoid contact from virus that contain the same motif (ELM)
- For example, the protein kinase R (PKR) protein blocks viral replication via phosphorylation of the translation initiation factor eIF2α. Poxviruses inhibit PKR using an eIF2αmimic, K3L.
  Thus, PKR is a defense protein that needs to maintain binding to its substrate eIF2α while avoiding interactions with the K3L mimic. colored regions of eIF2α (blue) and K3L (green) are structurally homologous. (Daugherty and Malik, 2012)



- **Research question** : Are the regions of motifs in viral proteins really characterized by being in regions of disorder structure? whether there is a difference between human and viral proteins in this respect ?

Data:

The analysis has been performed on two data: elm instances and elm classes.

- **Elm instances** - Contains all of experimental annotated ELM instances in tsv format file of all type and spices. divided by columns of - Accession, ELMType, ELMIdentifier, ProteinName, Primary_Acc, Accessions, Start, End, References, Methods, InstanceLogic, PDB, Organism.
- **ELM classes** - Contains all of experimental annotated ELM classes in tsv format file of all Accession ELM. divided by columns of – Accession, ELMIdentifier, FunctionalSiteName, Description, Regex, Probability, Instances, Instances_in_PDB.
- The project will focus mainly on analyzing the information according to : ELMIdentifier, Primary_Acc, Start-end motif, type of Organism and regular expression.

Methods:

Handling data: mining data and analyzed between two tsv dataframe, query online data and external software. organized the data in a convenient pythonic data structure by using string, lists, tuples, dictionary, read and write into files and handle with errors. In addition, using python libraries such as pandas, requests, biopython, re, os, scipy, mathplotlib and seaborn.

Analysis:
Query online databases – download sequence from Uniprot database : accessible database of protein sequence and functional information.
Sequence analysis - with list comprehension Biopython and regular expressions (regex).
External software – Iupred for predicting disorder area in the motif: uses a quadratic expression in the amino acid composition, which considers that the contribution of an amino acid to order/disorder depends not only its own chemical type, but also on its sequential environment, including its potential interaction partners. residues with values **above 0.4** can be regarded as disordered.
Comparison - to human data ELM proteins that act as a control group against the virus ELM proteins.
Statistical analysis - Pearson correlation coefficient and Mann–Whitney U test.
Visualization data - by using figures of Scatter plot with regression line, boxplot and histogram.

Results:

Human Vs Virus disorder value paired by amino acid: it can be seen in the figure of the paired scatter plot that each point represents an average of the disorder motif value values for each amino acid (in one letter). The x-axis represents the human proteins and the y-axis represents the viral proteins. It showed that there is a general trend that the average values of the disorder motif region in most amino acids tend to be in disorder region and in higher value in human proteins compared to viral (under the x=y line, they more tend to disorder area). This aspect is also expressed in the regression line in blue which is more inclined towards the x-axis and indicates higher values in human proteins. In addition, we can see the **Persson correlation** is very low and the p-value is not significant.

Human vs Virus average Iupred disorder value: it can be seen in the boxplot that the distribution of disorder motif values between human proteins and viral proteins, both of them are tend to be in disorder value (above 0.4). Here, as parallel to the scatter plot it can be seen that the median of human proteins tends to be higher than the viral proteins (more tend to disorder). In addition, the number of amino acids that contained in each group is indicated in the center of the box plot. In the **histogram plot** we can see the distribution of the two groups and it can be seen in general that human proteins tend to be in higher concentration in regions of high values according to the trend we have seen before.
In addition, it can be seen that a **Man whitneyu test** (2-taild) was performed and again the p-value is not significant between the two groups.

Conclusion: It can be seen that in general the area of the motif in human and virus proteins tend to be in regions that are disorder according to the lecture. The Human motif tend to be in more disorder area compare to the virus but not in a significant way.

Possible causes for non-significance:

- There is no significant difference between viruses and humans in disorder value in motifs area.
- Our data was not large and comprehensive enough for a clear difference between the groups.
- Iupred may not be the right tool to perform the analysis for compering disorder averages in motifs between human and viral proteins.