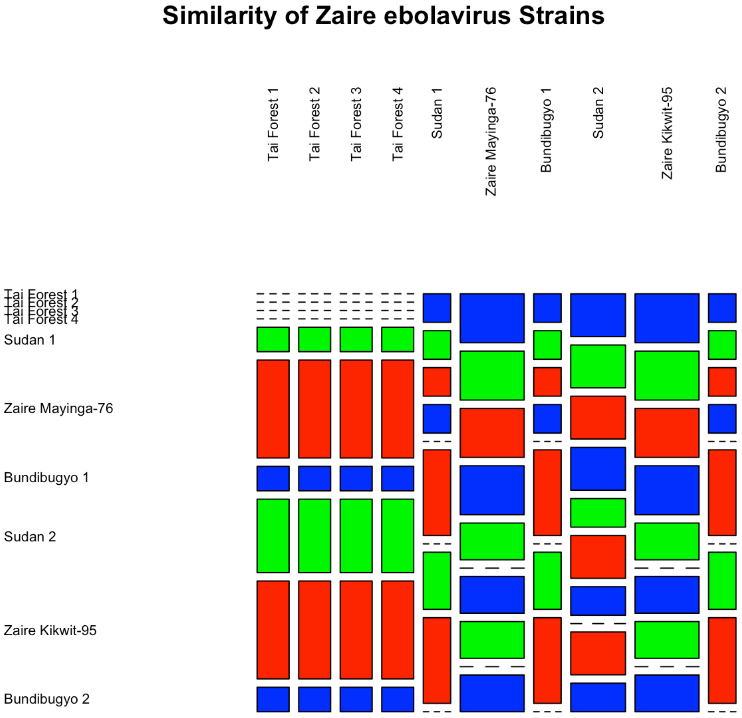
Hello!
My name is Revanth and I am a Computer Science and Cognitive Science Major. My interests include using my programming and data analytics skillset to solve problems in bioinformatics and life science. For this assignment, I chose to look at problem that an epidemiologist might face in real life.

Data

According the World Health Organization (WHO), Ebola has a 50% mortality rate. I looked at the proteins associated with **VP35** gene in various strains of the Ebola virus. My dataset consists of 10 protein sequences of length 216 amino acids, which I collected from UniProt Protein Database. The **VP35** gene has proven to contribute to virulence of the Ebola virus, because it causes the production of *interferons*, which prevent the host's immune system from combatting the virus. As such, I looked at the various strains of the virus and how they differ in produced protein sequence.

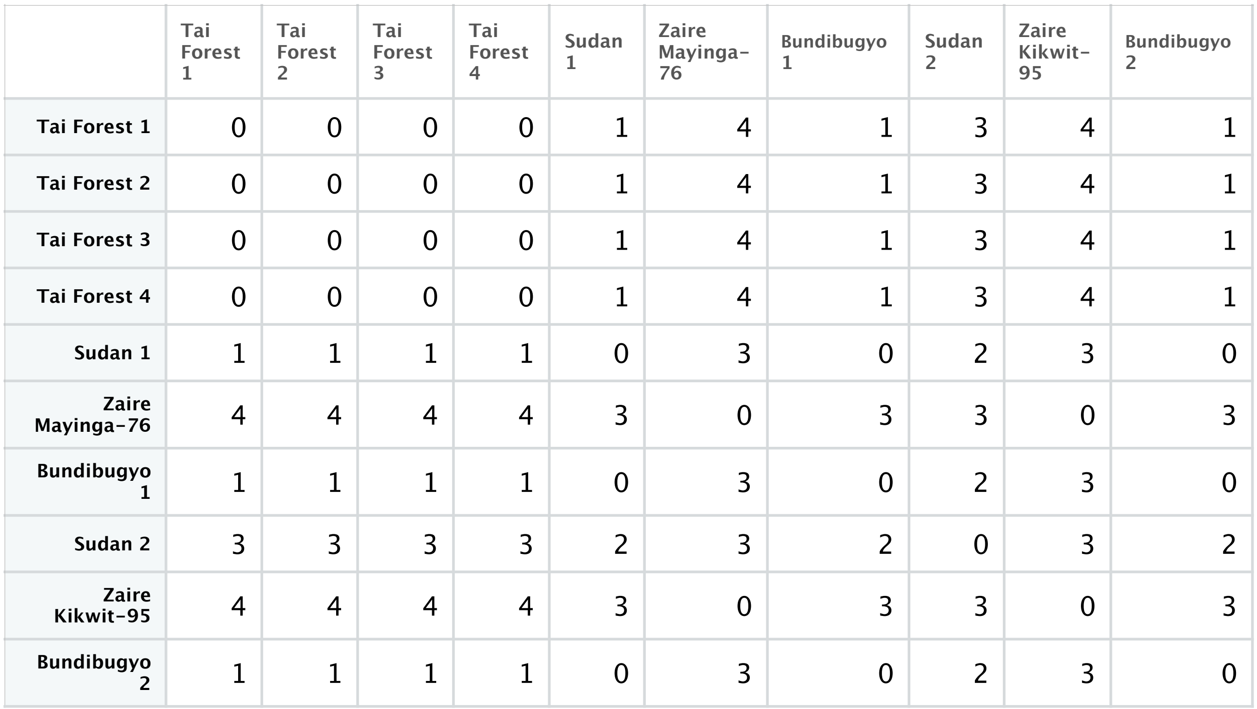
Due to random chance or human error, there is a chance that protein sequences have been misread or the wrong sequence has been uploaded. When dealing with proteins it is possible that a sequence has been misread and parsed incorrectly. However, the UniProt database is fairly reputable as it is has widespread use and the data has been annotated and cross-validated by many researchers. As such, I believe that the data source is trustworthy, but there is always a risk miscoding information when using data that is extremely granular. Additionally, my data set is small and I should be wary of the Law of Small numbers when interpreting the results.

I conducted a Z-test to see if the protein sequences were closely related or not differentiable from a random sequence of amino acids. There was overwhelming evidence that the proteins sequences were distinctly similar to each other than random amino acid sequences.



Method

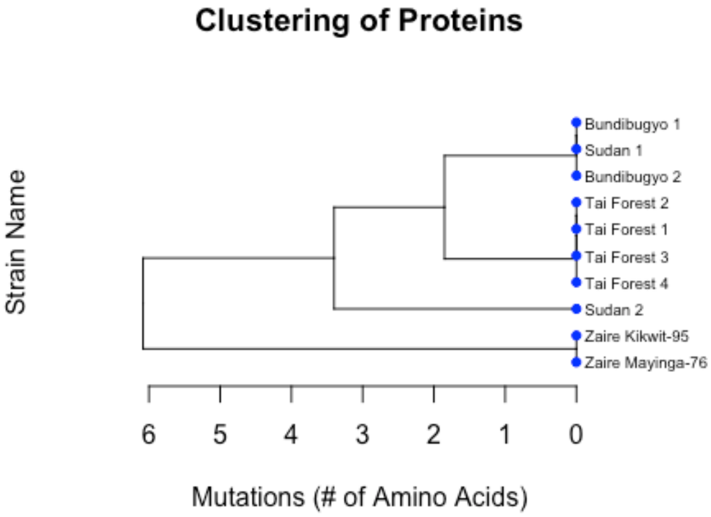
To further compare sequences, I first look at the **Hamming Distance** of all pairs of protein sequences (Hamming Distance = how many amino acids are in the same position when comparing two proteins). I saved all of these distances in a distance matrix (see table below), which I then converted into a mosaic plot (left).



	Tai Forest 1	Tai Forest 2	Tai Forest 3	Tai Forest 4	Sudan 1	Zaire Mayinga-76	Bundibugyo 1	Sudan 2	Zaire Kikwit-95	Bundibugyo 2
Tai Forest 1	0	0	0	0	1	4	1	3	4	1
Tai Forest 2	0	0	0	0	1	4	1	3	4	1
Tai Forest 3	0	0	0	0	1	4	1	3	4	1
Tai Forest 4	0	0	0	0	1	4	1	3	4	1
Sudan 1	1	1	1	1	0	3	0	2	3	0
Zaire Mayinga-76	4	4	4	4	3	0	3	3	0	3
Bundibugyo 1	1	1	1	1	0	3	0	2	3	0
Sudan 2	3	3	3	3	2	3	2	0	3	2
Zaire Kikwit-95	4	4	4	4	3	0	3	3	0	3
Bundibugyo 2	1	1	1	1	0	3	0	2	3	0

Using the distance matrix (and an R library), I create a dendrogram to show how similar each strain is with other. In a dendrogram, elements which are closer in the branches are more similar. An example of a dendrogram is shown below. Although this plot was not discussed in class, I found it to be a more interesting efficient way of explain some information.

Results



Of the 10 protein sequences in the dataset, we can see that certain strains of the virus produce more similar proteins than others. The dendrogram and the mosaic plot highlight this. Particularly, all of the Tai Forest strains produce identical proteins as did the Zaire strains. Interestingly, one of the Sudan strains proved to be more similar to the Bundibugyo strains than the other Sudan Strain. One should be cautious when reading these graphs, because the branches are shown *relative* to each other but not shown in context of the whole sequence length. In this dataset, even the most distant protein only showed a difference of 6 mutations out of a total length of 216 amino acids. As such, one should take the results with a grain of salt.

The dendrogram is useful because it could be used to map how the virus has mutated over time. An epidemiologist could infer the fatality rate or the virulence of a new strain based on how closely its protein sequence matches existing strains. Similarly, they could also infer how likely a particularly vaccine may work on a new strain. This would be a very crude estimate, however, it shows an efficient way of clustering information when conducting exploratory data analysis.

data source (UniProt protein database):
https://www.uniprot.org/uniref/UniRef90_Q05127?columns=id%2centry+name%2creviewed%2cprotein+names%2corganism%2corganism-id%2cclusters%2clength%2crole&offset=0

sources:

1. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3061251/
2. https://www.who.int/news-room/fact-sheets/detail/ebola-virus-disease