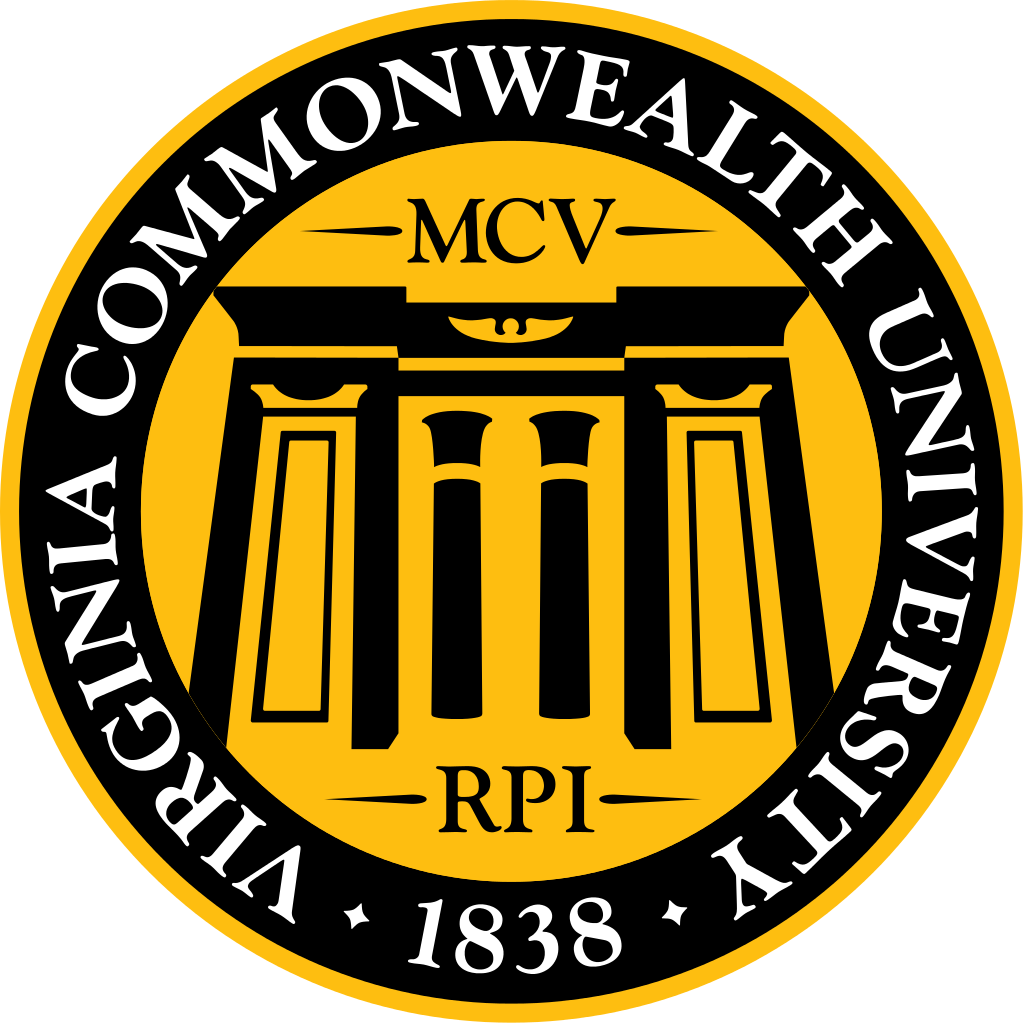
Hasan and Skyler

BNFO 420: Applications in Bioinformatics

Assignment 3: BLAST Pipeline Analysis

Due: 2/21/2017 before class (9:30am)

**The Problem:** Every year since 1997 the WHO has suggested annual reformulations of vaccine strains. You have received an email containing the protein sequence for the first vaccine strain that was recommended by the WHO. You have been asked by the WHO help provide evidence that supports their decision in selecting this particular strain of Influenza A H3N2 (Sydney 1997 like virus) for the first two flu seasons (1998 and 1999). You identify that the best way to do this would be to compare how well the Sydney 1997 strain matches against all sequences available over time (specifically 1997, 1998-1999, 2000-2001). After retrieving the file, you note in the header that this is a HA protein (Segment 4), but the idea of manually parsing page after page of BLAST output to identify the subtypes of the entire lot makes you a little queasy. Being an enterprising and resourceful scientist, you devise a plan to write a script that will cluster your sequences, making the job quite a bit less annoying by allowing the data to speak for itself. Since you know HA protein within subtypes are very similar, you decide that an acceptable threshold for similarity would be 97% identical. More plainly, your expectation is that the HA protein of the selected strain should match a significant portion of the amino acid sequences from all strains in the years leading up to it’s selection. Further, in the years afterward you expect this particular sequence to show up less.





# **The Program**

1. Starting from the blast.py program I’ve provided. You’ll need to devise a way to first compare the query against each database.
2. Then pick through and count results where the coverage was greater than 97% and see how representative this is of the total. HINT: This comes down to finding the percent of sequences that aligned at or above 97% coverage then comparing it to the total sequences in the database (subject).

# **The Questions**

Now that you have results from your program…

1. Looking at your results over time (year of the isolate), point out and justify what you believe to be antigenic shift and antigenic drift. Use examples from your data to help clarify your point.
2. Based on your results, do you think the WHO was justified in selecting this vaccine strain for each of the 1997/1998? Were they justified in changing it in the 1999 season? Should they change it for the 2000/2001 flu season?

**Deliverables:**

You need to submit to me (via BB):

1. Any scripts you created/modified (blast.py etc.), appropriately named with *eid*\_script.py
2. The answers to the questions in whatever text document you’d like.