Homework 4

this may need to get updated - do not use yet

Using the Biopython library construct scripts to process sequence file.

To use biopython on the cluster you will need to do

```
module load miniconda3
source activate GEN220
```

Short proteins report short_proteins.py

Write a script (`short_proteins.py`) that will read in a fasta file of proteins from a bacteria isolated from an earthworm. Download the proteins for Verminephrobacter eiseniae and write a script to write out new file with proteins which are less than 150 amino acids long.

See the SeqIO and Tutorial give some guidance as well as lecture notes.

Split files by count split_file.py

Write a script (`split_file.py`) that will read in the same file of proteins from Verminephrobacter eiseniae. This time your script will split the file into smaller chunks - each chunk will have 500 sequences in it. (bonus - make the 500 something the user can specify on the commandline instead). For each chunk of 500 write out a file. e.g. the first file could be: `chunk1.fa` and will have first 500 sequences, the file `chunk2.fa` will have the sequences 501-1000, etc.

You'll need to have a variable that keeps track of which 'chunk' you are on and one that keeps track of the sequences you want to write out as SeqIO likes to be given an array of sequences to write out all at once.

Regular expressions count_swissprot.py

Write a script `count_swissprot.py` that will ownload the Swissprot Database and open the file (no need to use biopython) using the simple open we have used before.

1. have the script count the number of proteins which come from the Arabidopsis genus using a regular expression to match lines which that. The species for a protein is in the "OS" field. Eg below you can see a "Solanum tuberosum" protein.

```
>sp|P93784|14335_SOLTU 14-3-3-like protein 16R OS=Solanum tuberosum OX=4113 PE=2 SV=1
```

2. Print out the number of sequences which come from a virus.
3. Print out the number of proteins which have a description which contains P450

4. How many have 'MAP kinase' - bonus how would you only count those which were exactly 'MAP kinase' but not 'MAP kinase kinase' or 'MAP kinase kinase kinase'.