ME.440.825 Quantitative Neurogenomics

Problem Set 1

Part 1: basic bash scripting

First, download this folder of genome sequence files. Then, answer each question using Bash scripting.

You'll need to have your folder of sequence files in your working directory. Print your working directory and its contents.

```
echo Working directory:
pwd
echo $'\n'Contents of working directory:
ls
```

```
## Working directory:
## /Users/acspiegel/Documents/Hopkins/Neuroscience/Courses/Quant_mol_neuro_2022/modules/module_1/pset
##
## Contents of working directory:
## Kcna1_sequence.txt
## findSeqMotif.sh
## ion_channel_sequences
## pset1_introToBashAndR.rmd
```

We'll start with Kcna1.fa, a FASTA file that contains the genome sequence of a mouse voltage-gated potassium channel. The first line contains a carat (">"), followed by a unique sequence identifier. The actual sequence starts on the next line. See for yourself by printing the first three lines of the file. Note that FASTA files can contain multiple sequences, but this one has just one.

```
head -3 ion_channel_sequences/Kcna1.fa
```

```
## >ref|NM_010595.3|:1-8970 Mus musculus potassium voltage-gated channel, shaker-related subfamily, meml
## GGGGGCTCCTCAGAGGCTCCGCAGCGGTGGAAGGACTGGAGCTGCTGGCTGCCTCCTCCGGTGCAGCCTG
## TATCCAGGTGCAGCGGCACTGGGGACGCGGTGCATATCCCTTGCTCAGACTGCCACTGTGACCCTTGCGC
```

Print just the sequence into a new file called Kcna1_sequence.txt

```
tail -n +2 ion_channel_sequences/Kcna1.fa > Kcna1_sequence.txt
```

How many lines are in Kcna1_sequence.txt? How many characters are in the file? If you subtract the number of lines from the number of characters, you should get the number of nucleotides in the Kcna1 gene. Why? Answer in a comment.

```
echo Number of characters:
wc -m Kcna1_sequence.txt

echo $'\n'Number of lines:
wc -l Kcna1_sequence.txt

# The wc command counts all the characters in the file, including newline characters.
```

```
## Number of characters:
##     9099 Kcna1_sequence.txt
##
## Number of lines:
##      129 Kcna1_sequence.txt
```

Count the number of times each nucleotide appears in the Kcna1 gene.

```
echo A count:
grep -o "A" Kcna1_sequence.txt | wc -l
echo G count:
grep -o "G" Kcna1_sequence.txt | wc -l
echo C count:
grep -o "C" Kcna1_sequence.txt | wc -l
echo T count:
grep -o "T" Kcna1_sequence.txt | wc -l
```

```
## A count:
##     2267
## G count:
##     2166
## C count:
##     2207
## T count:
##     2330
```

Voltage-gated potassium channels have a signature selectivity filter with the amino acid sequence TVGYG. Confirm that the Kcna1 gene has a sequence that would encode these amino acids. Your first step should be to remove newline characters from Kcna1_sequence.txt.

```
cat Kcna1_sequence.txt | tr -d "\n" |
grep -o "AC[TCAG]GT[TCAG]GG[TCAG]TA[TC]GG[TCAG]" | wc -l
```

```
##        1
```

Using *your favorite text editor*, write a bash script that takes a FASTA file (with a single sequence) and a target sequence motif as input and outputs the number of times that the motif appears in the FASTA file. You will use it here to determine which of the FASTA files in the ion_channel_sequences folder are likely to encode a voltage-gated potassium channel.

After writing your script, check its file permissions and make it executable

```
ls -l findSeqMotif.sh
#chmod +x findSeqMotif.sh
```

```
## -rwxr-xr-x@ 1 acspiegel  staff  240 Aug 28 22:46 findSeqMotif.sh
```

Use a loop to check whether each of the FASTA files in the ion_channel_sequences folder contains the TVGYG motif. Print the names of just the files that do contain the motif.

```
tvgyg="AC[TCAG]GT[TCAG]GG[TCAG]TA[TC]GG[TCAG]"

echo voltage-gated potassium channels:

for filename in ion_channel_sequences/*.fa
do
  motifCount=$(./findSeqMotif.sh $filename $tvgyg)
  if [ $motifCount -gt 0 ]
  then
    echo $filename
  fi
done
```

```
## voltage-gated potassium channels:
## ion_channel_sequences/Kcna1.fa
## ion_channel_sequences/Kcna2.fa
```

```
## ion_channel_sequences/Kcnb1.fa
```

R section

Include some R plotting

End of Problem Set