Working in Unix

Question: Take a look at the large file /info/DD2404/appbio16/data/gpcr.tab using head. This file contains data concerning G-protein coupled receptors . from a number of species.head. How many columns are there (if you count by eye)?

Answer: head gpcr.tab
There are 7 columns: Accession, Entry name, Status, Protein names, Gene names, Organism, and Length

Q: How many lines is there in the file?

```
wc -l gpcr.tab
```

gives 29 305 lines

Q : Use grep and wc to find out how many human GPCRs there are listed. Do you search for "human" or "Homo sapiens"?

```
grep Human gpcr.tab | wc
```

gives 2255

```
grep "Homo sapiens" gpcr.tab | wc
```

Gives 2244

We get 11 more hits when we search for Human instead of Homo sapiens. If we include the parentheses, then we get 2244 hits same as 'Homo sapiens'. “grep -P '(HUMAN)' gpcr.tab | wc”

Q: How long is the shortest sequence listed in the same file? Use cut and sort!

Answer: The file is tab delimited. Length on column 7. Using the command:

```
cut -f7 gpcr.tab | sort -h |head
```

yields the result that the shortest value is 10. Length is the 7th column in the data, so we can use cut -f7 to select this field. We don't need to specify a delimiter, because TAB delimited is the default. Then we use sort with the -n argument to do a numeric sort on text strings. Then finally we'll use head to just show the first 10 lines.

Q: How many species are named in gpcr.tab?

Answer: The idea for this will be cut to select the organism field, followed by sort to make duplicates adjacent, then uniq to deduplicate, finishing with wc. We subtract 1 to disregard the field name. 11136 species.

cut -f 6 gpcr.tab | ~~uniq -u~~ | wc   sort -u

"cut -f 6 gpcr.tab", cuts away everything except the 6th column, where the species are listed.w "| uniq -u", only prints out the uniq lines from the cut "| wc", uses the data from uniq to count the amount of lines.

Issues: Is the organisms name sorted? If not, sort the data before piping to uniq, since it only looks at adjacent lines. Are the names mixed? i.e lower and upper case etc. In that case, add -i to ignore case. Does some names have extra spaces, not have paranthesis when other does not? add -w to uniq to only compare a set amount of characters.

Q: Use a for-loop to apply multi-sequence alignment program *muscle* to the data files in /info/appbio15/data/testatin/*.fa. If you work on your own computer:

Answer: Muscle is a program that aligns multiple sequences by doing pairwise alignment. You select multiple sequences (files) and the program finds the optimum alignment between all the sequences (Note: a gap will always be gaps). Each file in the directory that we got are a fasta file with multiple sequences that needs to be aligned.

Program:

```
awk 'NR%2' *.fasta | awk 'FS="." {print $1}' | sort | uniq -d > listOfGenes.txt

for i in $(cat listOfGenes.txt)

do

        grep -h --no-group-separator -A1 $i *.fasta > ${i}.fasta

        ~/Users/prashant/muscle -in ${i}.fasta -out ${i}_aligned.fasta

done
```