Solutions for Assignment 3: *Mapping cDNA Reads*

Usage Statement:
Write the usage statement for map_sequence_starter.py in the README

python3 map_sequence_starter.py <cDNA_file> <seq_reads_file>
python3 map_sequence_starter.py /home/assignments/assignment3/scott_mouse_cDNAs.fa /home/assignments/assignment3/mckinley_raw_reads.txt

Question 1
Run `map_sequence_starter.py` *on the Scott cDNAs and the McKinley sequencing reads to align the sequencing reads. Paste the output in your* README.txt. *What genes are highly expressed in this sample? Which are lowly or not expressed? What is the function of each of these genes, and explain why do you think it is highly or lowly expressed in the sample.*

Question 1.1
What is the output from running `map_sequence_starter.py` *on the Scott cDNAs and the McKinley sequencing reads?*

```
Name    Reads   Reads per BP
Abcg2   7       0.0035460992907801418
Atp1A1  370     0.12044270833333333
Cd34    7       0.006092254134029591
ChAT    0       0.0
Gap43   36      0.05263157894736842
Gfap    38      0.029389017788089715
Mbp     304     0.4037184594953519
Myod1   0       0.0
Olig2   6       0.006172839506172839
Tubb3   154     0.11382113821138211
```

Question 1.2
What genes are highly expressed in this sample? Why do you think that is?

Consider the three most highly expressed genes in this set. Mbp is myelin basic protein, a well known marker of glial cells. Tubb3 is tubulin beta class 3, a well-known neuronal marker. Atp1A1 is an Na+/K+ ATPase, also known to be highly expressed in the CNS. The fact that all three of these are highly expressed is consistent with the fact that this sample was isolated from brain tissue.

Question 1.3
Which are lowly expressed? Why do you think that is?

The two genes that are not expressed at all, Myod1 and ChAT, are markers for muscle cells and motor neurons respectively. (Motor neurons are located exclusively in the peripheral nervous system). Thus, the fact that they are not expressed is also consistent with the origin of the sample.

Question 2

The Scott lab has further analyzed the expression of an additional 300 Mus musculus genes in this experiment. Of the top 20 most expressed of these genes, 10 are annotated as having brain-specific expression. Of the 300 genes they analyzed, 30 are annotated as having brain-specific expression. Is there an enrichment of genes annotated as having "brain specific" expression in their list of the top 20 most expressed genes? What statistical test did you use, and why? Was the test one-tailed or two-tailed? What was the exact p-value? (Do not state p < X). ***Show your work to get partial credit.***

Is there an enrichment…?
Yes

What test did you use?
The question asks to test if two categorical variables, expression (top expression and not top expression) and brain-specificity (brain specific and not brain specific) are independent. Furthermore, since the question asks if there is an enrichment of top expressed genes that are brain specific, the statistical test will be one-sided. The null and alternative hypotheses are:

$$H_0: Pr(gene\ is\ top\ expressed\ \&\ brain\ specific) = Pr(gene\ is\ top\ expressed) * Pr(gene\ is\ brain\ specific)$$
$$H_1: Pr(gene\ is\ top\ expressed\ \&\ brain\ specific) > Pr(gene\ is\ top\ expressed) * Pr(gene\ is\ brain\ specific)$$

The appropriate tests for this are Fisher's exact test (one-tailed) or the $\chi 2$ test for association.

The 2x2 contingency test for this is:

	Top expressed	Not top expressed	Total
Brain specific	10	20	30
Not brain specific	10	260	270
Total	20	280	300

Was the test one-tailed or two-tailed?
One-tailed

What was the p-value?
To calculate the p-value, you can use the Scipy package:

```
>>> import scipy.stats
>>> contingency_table = [[10,20], [10,260]]
>>> p_value = scipy.stats.fisher_exact(contingency_table,
alternative="greater")[1]
>>> p_value
2.0614035225265323e-06
```

Since the p-value < 0.05, enrichment of brain specific genes in the highly expressed gene set.

Question 3

In `map_sequence_starter.py` *why did we calculate hits as raw number divided by length of the gene, rather than just using raw number of counts?*

We divided by the length of the gene to normalize the RNA counts so that we can get a rough comparison of the relative molar amounts of each RNA molecule. If you do not normalize, long genes will be disproportionately represented in your sample. In general, one would also normalize by the number of reads obtained to get RPKMs (reads per kilobase DNA per million reads). This would allow you to compare genes across RNA-seq experiments.

Question 4

If we were trying to map RNA-seq data for the whole transcriptome rather than just 10 genes, what would you do differently? Name at least three limitations in the approach we have followed in this assignment.

Here is a non-exhaustive list of limitations:

1. Requires a perfect match to map back the read. This does not allow for indels, SNPs or sequencing errors in the read.
2. Does not utilize memory efficiently as the number of genes analyzed increases. Use dynamic programming to utilize memory more efficiently.
3. Does not check to see if two or more coding sequences share 25-mers. This is potentially a problem even for the smaller set of genes.
4. Does not allow for gaps.
5. Does not take into account splice isoforms.
6. Does not work with extremely short transcripts, e.g., microRNAs which are shorter than the given kmer.

Commenting

Comment the provided code. Specifically, make sure you look up and define all of the functions used.

The provided code is well commented. All of the functions are defined.
The student's code is well commented.