Homework 2

Simple Count and Report

Write a program called `squared_cubed.py` and prints out three columns of data, ideally, separated by tabs. A header line should be written which is labels of the columns

```
N    Squared    Cubed
```

Column 1: numbers 0 -> 30 Column 2: Square (x^2) of column 1 Column 3: Cubes (x^3) of column 2

Output should look like

```
N   Squared Cubed
0   0   0
1   1   1
2   4   8
3   9   27
4   16  64
5   25  125
```

Count up

We will compute some statistics for a tab delimited file called GFF which lists the location of genes and exons location in a genome annotation. Remember GFF is a structured format, tab delimited, which describes locations of features in a genome.

Recall eukaryotic Genes are made up of features: exons, introns, Untranslated regions (UTR). Some exons are coded as 'CDS' for CoDing Sequences - eg the ones that code for proteins.

See Wikipedia gene page and view of Gene structure in particular

Here is a GFF file for the *Penicillium chrysosporium* genome, which is the fungus which gave us one of the first antibiotics. The FungiDB database hosts genome sequences and data files for a collection of fungi.

The GFF file is available here FungiDB-54_PchrysosporiumRP-78.gff and FastA format genome assembly is FungiDB-54_PchrysosporiumRP-78_Genome.fasta. These are two files related to location of genes and sequence data.

Write a script called `genome_stats.py` to: 1. Download these file (this can be in UNIX before you run your python script or you can incorporate this into the python). I already wrote part of this for you in the template code you can start with that executes a `curl` command from within your script. But if this doesn't make sense to you, you can remove that. 2. Summarize how many exons, CDS, protein_coding_gene, and 3. Compute the total length of the genes (length is the END - START) 4. Use the FASTA file to compute the total length of

genome (by adding up the length of each sequence in the file). Recall I lectured on a basic code to read in a FASTA file - you can also see that code template here 5. Print out the percentage of the genome which is coding

Codon compute

Use the following files to examine codon usage across these two bacteria. Remember that codons are triplets (eg ACA, GAT, ...). There are 64 total possible triplets. To count these, know that they are non-overlapping sets of three adjacent bases in the sequences, start with the very first base as the reading frame.

These files are coding sequences of the predicted genes in each of two species.

1. ftp://ftp.ensemblgenomes.org/pub/bacteria/release-45/fasta/bacteria_0_collection/salmonella_enterica_subsp_enterica_serovar_typhimurium_str_lt2/cds/Salmonella_enterica_subsp_enterica_serovar_typhimurium_str_lt2.ASM694v2.cds.all.fa.gz
2. ftp://ftp.ensemblgenomes.org/pub/bacteria/release-45/fasta/bacteria_0_collection/mycobacterium_tuberculosis_h37rv/cds/Mycobacterium_tuberculosis_h37rv.ASM19595v2.cds.all.fa.gz

Write a script called `codon_compute.py`. You can download the data outside of the python script or you can include these steps in your script. I already wrote part of this for you in the template code you can start with that executes a `curl` command from within your script.

The code you write will need to process these files in order to print out the following information:

1. The total number of genes in each species.
2. Total length of these gene sequences for each file
3. The G+C percentage for the whole dataset (eg the frequency of G + the frequency of C)
4. Total number codons in each genome.
5. Print out table with three columns: Codon, Frequency in Sp1, Frequency in Sp2