Homework 2

This assignment calls for two scripts. They are both started in the template.

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 this (but going up to at least 30 for for the N column)

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

Genome Stats

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. **Print out** the number of exons, CDS, protein_coding_gene features found in the genome annotation (GFF file) 3. Compute and **print out** the total length of all the protein_coding_gene features (length is the END - START). 4. Compute and **print out** total length of all the CDS features (length is the END - START). 5. 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, **Print out** the total length. 6. *Print out* the percentage of the genome which is coding (using the numbers calculated from the protein_coding_gene)

Hints: - starter code is provided but you can solve this in a different way or just add to this script and commit it. - a dictionary will be useful for capturing the counts of the numbers or lengths of the different features as you loop through the GFF file - the `aspairs()` function returns a dictionary where the keys are sequence IDs and the values are the DNA sequence for each of the contigs.