Homework 2

Simple Sevenless

Write a program `sevenless.py` to print out all the numbers from 0 to 99, one on each line, except do not print any number perfectly divisible by 7

Open and Shut

Write a script `open_shut.py` to write a new file called 'closed.txt'

A database of all movies file is located at https://datasets.imdbws.com/title.basics.tsv.gz or you can use the already downloaded file at `/bigdata/gen220/shared/simple/title.basics.tsv.gz`

In this file please have it print out: 1. the number of movies which have 'door' (or) 'Door' in the name. 2. the number of movies which have the word 'door' or 'Door' (eg it has to be a whole word not a part of another word). 3. the number of movies with the word "Open" and the number with the word "Closed"

Count up

Compute let's calculate some statistics from this GFF file which lists the location of genes and exons locations. Remember GFF is a structured format, tab delimited, which describes locations of features in a genome.

Here is a GFF file for the E. coli K-12 genome. ftp://ftp.ensemblgenomes.org/pub/bacteria/release-45/gff3/bacteria_0_collection/escherichia_coli_str_k_12_substr_mg1655/Escherichia_coli_str_k_12_substr_mg1655.ASM584v2.37.gff3.gz

Here is a Fasta file for the genome of E. coli K-12. ftp://ftp.ensemblgenomes.org/pub/bacteria/release-45/fasta/bacteria_0_collection/escherichia_coli_str_k_12_substr_mg1655/dna/Escherichia_coli_str_k_12_substr_mg1655.ASM584v2.dna.chromosome.Chromosome.fa.gz

Write a script called `count_up.py` to: 1. Download this file. 2. Count up and print out the number genes (gene feature) 3. Compute the total length of the genes 4. Use the Fasta file to compute the total length of genome 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` which will download and process these files in order to print out

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 file
5. Print out table with three columns: Codon, Frequency in Sp1, Frequency in Sp2