Downloading genomes from the online database, NCBI
Week 10 Activity - due Nov 1, 2021 12pm

Aim

For this activity, we are going to download a genome from the NCBI database, and calculate the ATGC content using our lab03.py scripts. We are going to be working on Trypsin, because reference genomes can be large and can take up a lot of harddrive space.

Protocol

1. Log on to genome@trypsin.sfsu.edu (Reminder: password is DNA4rosalind!)
2. Navigate to your student directory
3. Download the reference genome to your student directory.
    a. To download the reference genome, search for "Homo sapiens" and choose "Genome" dropdown menu.
    b. The genome can be downloaded by clicking the "genome" link next to "Download sequences in FASTA format for".
    c. Right click the "genome" link and copy the link address.
    d. On trypsin, paste the URL after the word wget
        i. `wget [URL]`
    e. The reference genome comes compressed, unzip the reference genome using the following command
        `gunzip [name of reference file]`
4. Upload your lab03.py script to Trypsin.
    a. Open another command line window
        i. If you're on a Mac, navigate to your lab03.py directory using the usual commands (cd, ls, pwd)
        ii. If you're on a PC, navigate to your lab03.py directory.
            1. ls = dir
    b. Use the scp command to upload your lab03.py to Trypsin
        i. Find the directory on Trypsin you want to upload your script to. This should be your student directory address. To find the absolute path of your student directory, type in pwd when in your student directory to find its address.
        ii. Next, use the scp command where your lab03.py is:

`scp` *`[lab03.py name]`* `genome@trypsin.sfsu.edu:`*`[absolute path to your student directory]`*

5. Check to see if lab03.py and the reference genome is now in your student directory on Trypsin (use the command ls while in your student directory)
6. If both are there, run lab03.py on the reference genome to output the nucleotide composition of the human genome.

7. Check to see if the GC content your script computed is the same as the GC content listed in "https://www.ncbi.nlm.nih.gov/genome/?term=homo+sapiens" under "Summary" and subheading "Statistics".
8. Submit your ATCG composition output to iLearn Assignment "Week 10 - NCBI Activity". If your output contains the DNA codon composition, you can submit that too.

Using multiple input files
Week 10 Activity 2 - due TBA

Aim
The aim of this exercise is to extract the gene sequences from a reference genome. We will use the gene locations stored in a file called a "general feature format" (GFF) file.

Protocol
1. Download the GFF file of the human genome to trypsin using the wget command using the links below. The GFF file shows where genes are in a genome. It also holds information like where exons are. For the purposes of this activity, only use "gene" annotations.

FASTA File:
https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/001/405/GCF_000001405.39_GRCh38.p13/GCF_000001405.39_GRCh38.p13_genomic.fna.gz

GFF File:
https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/001/405/GCF_000001405.39_GRCh38.p13/GCF_000001405.39_GRCh38.p13_genomic.gff.gz

2. Unzip the FASTA and GFF files

3. Write a python3 script that reads in (1) the GFF file, and stores the gene locations in a data container like a dictionary, (2) Reads in the human genome and stores the DNA sequence as a string or list. (3) Extracts all of the genes from the genome.

Notes
(1) locations are 1-based in a GFF file, 0-based in python3. Subtract 1 from the gene start and stop locations when indexing.