Homework 4

Two parts: practice Python on genome and gene-table data (Part 1), then compare genes between two bacterial strains with BLAST (Part 2).

Accept the assignment through GitHub Classroom (link in Canvas), then clone your repository. Starter scripts are provided; fill in the code.

Part 1: Practicing Python

0. `setup.sh`: download and uncompress these two genomes from NCBI.
   - *Thermus aquaticus*: https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/001/399/775/GCF_001399775.1_ASM139977v1/GCF_001399775.1_ASM139977v1_genomic.fna.gz
   - *Bacillus subtilis*: https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/009/045/GCF_000009045.1_ASM904v1/GCF_000009045.1_ASM904v1_genomic.fna.gz

Also download the CDS files for two strains of the bacterium *Serratia marcescens*:

- Reference: ELP1.10 CDS
- Query: Serratia_CH31 CDS

1. `gc_stats.py`: print the percent GC (G and C bases divided by total sequence length) for each of the two genome files.

   Bonus: make the script generic so it prints the filename and GC content for any set of files passed on the command line. On the HPCC, `/bigdata/gen220/shared/bacteria/dna` has many genomes to try this on.

2. `cds_stats.py`: read each of the two Serratia CDS files (the sequences are coding sequence only, so there are no introns or UTRs) and write a table with one row per CDS to STDOUT or a file (your choice). Columns:

   - **ID**: the `locus_tag` from the FASTA header (e.g. `[locus_tag=...]`), or the first word of the header if there is none
   - **Length**: CDS length in bp
   - **GC%**: percent G+C of the CDS
   - **Start codon**: the first three bases (ATG, GTG, TTG, ...)
   - **Stop codon**: the last three bases (TAA, TAG, TGA)
   - **Complete ORF?**: `yes` if the length is a multiple of 3, it starts with a valid start codon and ends with a stop codon; otherwise `no`

   Then, for each file, print a summary:

   - total number of CDS, and total coding bp
   - mean, median, shortest and longest CDS length (report the ID of the longest and shortest)
   - mean GC% of the CDS
   - count of each start codon and each stop codon (a `collections.Counter` works well here)
   - number of CDS that are not a complete ORF

   Which start codon is most common? How does the mean CDS GC% compare to the genome GC% you got in step 1 (use the Serratia genome files if you want a direct comparison)? Add a short comment at the top of your script with your answers.

   For a larger test set, use the files in `/bigdata/gen220/shared/bacteria/cds` on the HPCC, or just use the two Serratia files above.

   Bonus: sort the table so the longest CDS comes first and the shortest is last. Or print a histogram of CDS lengths in 100 bp bins using text characters.

Part 2: Comparing strains with BLAST

Compare the genes from ELP1.10 to Serratia_CH31 with BLAST.

3. Write a script to BLAST the ELP1.10 CDS file (reference) against the Serratia_CH31 CDS file (query): make a BLAST database, run `blastn` with tabular output (`-outfmt 6`), and write `serratia_search.blastn.tsv`. Run `git add` and commit this file.

4. Complete the two scripts:

Script	Function	Returns
`blast_match.py`	`calculate_num_matched_sequences(filename, min_identity)`	number of query sequences with a match at or above `min_identity`
`blast_match_PID.py`	`find_lowID_best_hits(blastfile, threshold)`	names and percent identity of query sequences whose BEST hit is below `threshold` (default 80)

I should be able to run each (e.g. `python3 blast_match_PID.py`) as long as `serratia_search.blastn.tsv` exists.