Homework 2 – git and pseudocode – for biomedical data analysis projects
DUE: Thursday, September 26, 2024, at 11:59am
TOTAL 20 points
PROBLEMS: 2

##
##

PROBLEM 1: (10 points) – Pseudocode for fundamental concepts in DNA sequence analysis
----

Background

This part of the assignment yill introduce you to fundamental concepts in DNA sequence analysis.

Task Overview

Analyze a short DNA sequence using basic bioinformatics techniques and create simple algorithms in pseudocode.

Detailed Steps
----

1. (1 point) Sequence Generation

   - Create a DNA sequence of exactly 20 base pairs using the random generator: http://www.faculty.ucr.edu/~mmaduro/random.htm

2. (2 points) Basic Sequence Analysis

   - Write pseudocode for an algorithms that:
     a) Count the frequency of each nucleotide (A, T, C, G) in your sequence\

3. (1.5 point) Custom Pattern Design

   - Research common DNA patterns (e.g., TATA box, CpG islands, start codons)
   - Design/choose a custom pattern that is:
     a) 4–6 base pairs long
     b) Inspired by a common DNA pattern you've researched
   - Include the pattern in the solution file and explain in 2–3 sentences why you chose this pattern and its potential biological significance

   - Resources:
     â€¢ A link to sequence motifs (see sequence logo – you can look at the logo and that can be the pattrn you choose): https://jaspar.elixir.no/collection/core/
     â€¢ Click on the motif ID in the table at the link above and then on validation and it takes you to a paper describing the validation of that motif
     â€¢ Link to a wiki page describing DNA motifs: https://biocorecrg.github.io/CRG_Bioinformatics_for_Biologists_2021/dna_motifs.html
     â€¢ Link to a bioinformatics center main wiki page: https://biocore.crg.eu/wiki/Main_Page
     â€¢ Link to a class project that researched the topic of finding motifs: http://engr.case.edu/li_jing/papers/00798gpattern.pdf
     â€¢ Link to a paper on promoter binding site prediciton algorithms: https://link.springer.com/protocol/10.1007/978-1-60761-854-6_5

4. (2 points) Pattern Search

   - Add your custom pattern in three random places in your sequence
   - Write pseudocode for an algorithm that finds the start index for all occurrences of your custom pattern in your sequence, the result will be a list of indices where the pattern occurs in your sequence

5. (2 points) Complementary Sequence

   - Write pseudocode for an algorithm that generates the complementary strand of your DNA sequence
   - Remember: A pairs with T, C pairs with G

6. (1.5 poins) Conclusion

   - In 2–3 sentences, explain why identifying specific patterns in DNA is important for biological research
   - In 2–3 sentences, describe how the frequency of your custom pattern might affect its biological role

Submission Requirements

- Submit file called b575hw2pb1_pseudocode.txt that contains the solution to the steps 1–6.
  For each pseudocode algorithm include the walkthrough and result.

##
##

PROBLEM 2: (10 points) – Collaborative Bioinformatics Project: Git, GitHub, and Sequence Analysis
----

This part of the assignment will introduce you to version control using git and GitHub in the context of a collaborative bioinformatics project.

Task Overview

You'll practice essential git commands, contribute to a shared repository, and perform basic sequence analysis tasks.

Detailed Steps
----

1. (1.5 point) Forking and Cloning

   - Fork the Repository: Go to the GitHub repository provided by your instructor: https://github.com/mitreacristina/b575_hw02 and fork it to your GitHub account.
     â€¢ There is a button fork on the main page of the repository.
     â€¢ Details about how to fork the repository: https://docs.github.com/en/pull-requests/collaborating-with-pull-requests/working-with-forks/fork-a-repo
   - Clone Your Fork: Clone your forked repository to your local machine

2. (0.5 points) Creating a Branch

   - Create a new branch for developing a sequence analysis script called sequence_analysis

3. (1.5 point) Creating a Bash Script for Sequence Analysis – Making changes

   - Add Sequence Data: Create a sample DNA sequence with 213 nucleotides using the sequence generator at: http://www.faculty.ucr.edu/~mmaduro/random.htm and add it to the sequence.fasta file (break it on three lines and use the id: ref|sequence5_ID|Homo Sapiens)
   - Write the Script: Create a new file named analyze_sequence.sh with the following content:

     ```
     #!/bin/bash

     # Count the number of sequences
     seq_count=$(grep -c "^>" sequence.fasta)

     # Count the total number of bases
     base_count=$(grep -v "^>" sequence.fasta | tr -d '\n' | wc -c)

     # Calculate GC content
     gc_count=$(grep -v "^>" sequence.fasta | tr -d '\n' | tr -cd 'GCgc' | wc -c)
     gc_percent=$(echo "scale=2; $gc_count / $base_count * 100" | bc)

     echo "Number of sequences: $seq_count"
     echo "Total bases: $base_count"
     echo "GC content: $gc_percent%"
     ```

4. (2 points) Tracking changes

   - Stage and commit the changes (the modification of data file and addition of the sequence analysis script)
   - Push Changes: Push your branch to your forked repository

5. (3 points) Creating Pull Requests

   - Go to Your Forked Repository: Navigate to your forked repository on GitHub
   - Create Pull Requests: Create pull requests for both your sequence-analysis branch to the original repository, describing your changes and assign it to me: cristinamitrea.
     â€¢ Note you are creating a pull request against the main branch of the original repository – the one that is under my GitHub username: cristinamitrea
     â€¢ Here is a link form the GitHub documentation with the steps to follow: https://docs.github.com/en/pull-requests/collaborating-with-pull-requests/proposing-changes-to-your-work-with-pull-requests/creating-a-pull-request-from-a-fork

6. (1.5 points) Conclusion

   - Create a file named conclusion.txt in your local repository and answer the following questions in 2–3 sentences:
     â€¢ What was the most challenging part of this assignment?
     â€¢ What is one element earn about git and GitHub through this experience?
   - Submit Conclusion: Add, commit, and push the new file conclusion.txt

Submission Requirements

- Submit a file called b575hw2pb2_GitHub.txt that contains your GitHub username.

##
##