Programming for Bioinformatics | BIOL 7200

Week 6 Exercise

September 26, 2023

Instructions for submission

- Run a Linux or Mac terminal on your computer
- You may want to create a directory to work in (e.g., "~/biol7200/class6/ex6")
- Download "aligned.fna" into your working directory
- Name your script file(s): "gtusername_questionnumber.py"
- Write your script(s) to be compatible with Python version 3.11

NOTE starting this week, unless explicitly instructed to write in Bash, only Python solutions will be accepted. No Bash scripts!

Grading Rubric

This assignment will be graded out of 100.

You only need to write one script for this assignment. **No Bash code allowed this week!!!** Your solution must be all Python code in a script.

Question 1 (100 points)

Write a Python script that reads the FASTA file, "aligned.fna" provided on canvas and prints the sequences to the terminal along with symbols indicating which positions matched. The output of your script should look as follows:

Given the following sequences in a FASTA file,

```
>seq1
ATGCAAGTCGAGCGGATGAAGGGAGCTTGCTCCTGGATTCAGCGGCGGAC
>seq2
ATGCAAGTCGAGCGGCAGCACAGAGGAACCTTGGGTGGCGAGCGGCGGAC
```

Your script should produce the output

```
ATGCAAGTCGAGCGGATGAAGGGAGCTTGCTCCTGGATTCAGCGGCGGAC
|||||||||||||||  | |  |||    ||   |     ||||||||||
ATGCAAGTCGAGCGGCAGCACAGAGGAACCTTGGGTGGCGAGCGGCGGAC
```

Where a pipe symbol is printed between the sequences at positions where the bases are identical, and a space is printed at positions where the bases differ.

The sequences in the file "aligned.fna" have already been aligned, so you need only process the sequences and print the output in the described format.

Your script must take command-line input. Do not hard-code the path to the sequence file. The usage of your script should be

`<script name>.py <FASTA file>`

Extra credit (20 points)

If you submit a solution for this question. Name your script "gtusername_EC.py"

It is an effective strategy when learning a second programming language to write scripts in both languages to see the similarities and differences between the languages. For this question, your task is to rewrite the final Bash assignment in Python, but with a few differences. The differences are:

1. You should run BLAST outside of the Python script using the following command:

   ```
   tblastn -query data/HK_domain.faa -subject data/Vibrio_cholerae_N16961.fna -
   outfmt '6 std qlen' > Vc_blastout.txt
   ```

2. You should not use `awk` to process the BLAST output. Instead, your script should read in the unprocessed BLAST output and the "Vibrio_cholerae_N16961.bed" file and all processing should be performed within a Python script using only Python code.

To remind you, BLAST hits should be processed (using Python) to only keep hits with greater than 30% identity and >= 90% length.

Your script should write the unique list of identified homolog genes to an output file (specified in the commandline) and should print the number of homologs identified.

The usage of your script should be

`<script name>.py <blast output> <BED file> <output file>`

Using the specified BLAST command, you should get 34 homologs for *Vibrio cholerae*. You may use the example Bash scripts included in the slides of the demo session as the inspiration for this script if the script you wrote did not get the correct number of hits.