Homework 1

CSCI 5481, Computational Techniques for Genomics
University of Minnesota
Instructor: Dan Knights

Instructions

- Please turn this assignment in on the course Canvas page.
- If there are multiple files to turn in, all text and code should be placed into a single folder with a name like *lastname_homework1*. The folder should then be compressed and submitted as a single archive (.zip or .tgz)
- You are encouraged to discuss this project with others, but you need to write your own code.
- Please write the names of anyone with whom you discussed this assignment at the top of your code.

Background

The purpose of this exercise is to get you prepared to write convenience wrapper scripts that run 3rd party software tools for performing alignment, assembly, phylogeny inference, secondary structure prediction, etc. You will learn how to write a script in Python that uses command-line parameters, reads a file, and can execute a command and retrieve its standard output and error output. **You must be using a linux operating system to complete this assignment.**

Tasks

1. Download the Homework1 folder: https://www.dropbox.com/s/lw8a5l4g6b9yfjv/Homework1.zip?dl=1
2. Modify Homework1/template.py for Python 2 or Homework1/template-python3.py for Python 3 to do the following:
    a. Call the BURST command from within python using the following syntax:
    ./burst -r ref.fna -q query.fna --taxonomy taxonomy.txt -o output.txt
    b. Print out the return value, standard output, standard error.
3. Verify that your program runs with:
    python homework1.py -q query.fna -r ref.fna -t taxonomy.txt -c /path/to/burst -o output.txt -V
4. The output file shows information for every query sequence that hit a reference sequence at 97% similarity or above. Answer these questions, using this guide to the meaning of the first 12 columns in the output.txt file: https://github.com/seqan/lambda/wiki/BLAST-Output-Formats. The 13th column shows the taxonomic assignment of each query sequence.
    a. What fraction of the original input query sequences had a match in the database at 97% or above? Note that a few matches fell below 97%. You can exclude these. Inspect query.fna, which is in FASTA format, to ensure that you are counting the correct number of input sequences.
    b. What is the most common bacterial species in the query set? Note that any taxonomy string ending with something after the last “s__” is a species name. If a string ends with “s__” and nothing else, then it was ambiguous and does not count as a species.

c. What is the average percent similarity of the matches? You can exclude those below 97%.

Deliverables

1. Please turn in, via moodle:
    a. Your well-commented code. Note in your code the names of any people with whom you discussed the assignment. This includes code for answering question 4 above.
    b. The output printed by your program (this is what is printed to standard out, not the BURST output file).
    c. The output.txt file generated by BURST
    d. Answers to questions above in txt/doc/pdf