Homework Project 1 - 10 kilobase alignment

CS/BIOINFO M122/M222

Due: Monday, May 4th, 2021, 5:00 pm

This short programming assignment is designed to help you get an understanding for the basics of sequence alignment. You can use any language for this project, but Python is strongly recommended, and you will receive starter code in Python. You will submit your response to https://cm122.herokuapp.com/upload as a `.zip` file.

Overview
In the first two programming assignments for this class, you will solve the computational problem of re-sequencing, which is the process of inferring a donor genome based on reads and a reference.

You are given a reference genome in FASTA format, and paired-end reads.

The first line of each file indicates which project the data relates to. In the reference file, the genome is written in order, 80 bases (A's, C's, G's, and T's) per line.

The paired end reads are generated from the unknown donor sequence, and 10 percent of the reads are generated randomly to mimic contamination with another genetic source. There is also a 1% sequencing error rate that is independent of location in the read.These reads are formatted as two 50 bp-long ends, which are separated by a 90-110 bp-long separator.

Your task is to map the reads to the reference genome and then determine where the reads indicate there is a difference (a variant). The different kinds of variants that may be found in the donor genome are explained at https://cm122.herokuapp.com/variant_doc

Starter Code

Starter code for all the class projects is available at https://github.com/rosie068/CM122_starter_code . It is strongly recommended that you use git for these programming assignments, and to set up your own account on github.com. The tutorials at http://try.github.io/ might come in handy. If you are completely unfamiliar with git and github, you can obtain a copy of the code by running
```
git clone https://github.com/rosie068/CM122_starter_code.git
```
This will create a folder named CM122_starter_code in your current directory.

Tutorial

The starter code provided handles reading the reads and reference genome into Python lists, as well as converting a list of SNPs into the proper output format. You will be responsible for aligning the reads to the reference, and calling SNPs.

You should download the practice from https://studentdownloads.s3-us-west-1.amazonaws.com/practice_W_1.zip and "for-credit" data from https://studentdownloads.s3-us-west-1.amazonaws.com/hw1_W_2.zip . **DO NOT DOWNLOAD FROM THE LINKS ON HEROKU.** If you want to follow the tutorial below, download and extract these files into the HP1 folder. Via command line that would look like this:

```
cd CM122_starter_code
cd HP1
wget http://studentdownloads.s3-us-west-1.amazonaws.com/practice_W_1.zip
wget http://studentdownloads.s3-us-west-1.amazonaws.com/hw1_W_2.zip
unzip practice_W_1.zip
unzip hw1_W_2.zip
```

Obviously, you can just also open the browser and download and unzip the files using your file manager.

We are providing you with the skeleton for one script:
1. `basic_aligner.py` takes in a reference genome, a set of reads, an output file and an output header and outputs the SNPs called based on its produced alignment

Running the above script with the `-h` option should be self explanatory, but here is an example of running them to create a file that can be submitted on the website for the practice data.

```
python basic_aligner.py -g ref_practice_W_1_chr_1.txt -r reads_practice_W_1_chr_1.txt -o test_output.txt -t practice_W_1_chr_1
```

This will generate the file test_output.txt.zip that you can submit on the website. It also generates a .txt file that you can look at. **Note that the -t parameter HAS to be practice_W_1_chr_1 when submitting the practice data and hw1_W_2_chr_1 when submitting the real assignment. This will let the online submission system know on which leaderboard to place you.**

Read the content of HP1, and see if you can understand what it is doing. You can submit your results as many times as you want to achieve a passing score.

I/O Details

https://cm122.herokuapp.com/ans_file_doc should handle most of your questions on reading and writing output.

Questions to consider
The genome from which the reads are generated has not only SNPs, but insertions, deletions, and repeated sequences that are not present in the reference. In addition, there is a 1% sequencing error rate independent of location in the read.

What will these non-SNP mutations look like when you try to align them to the genome? Try writing it out on a piece of paper.

More generally, what is the "signature" of a SNP mismatch in the consensus sequence? What is the "signature" of an insertion or deletion?

Grading

Remember to submit your solutions to https://cm122.herokuapp.com/upload as a `.zip` file. You can submit as many times as you want without penalty.

You will be graded on your performance on the test set, which can be found at https://studentdownloads.s3-us-west-1.amazonaws.com/hw1_W_2.zip and under week 4 on CCLE. Again, **DO NOT DOWNLOAD FROM THE LINK ON HEROKU**. You can also submit your solutions for the practice data to https://cm122.herokuapp.com/upload to see how your solution is performing.

Undergrads will get full credit with a score of 45 on SNPs, and no credit for a score of 25 or below. Grad students will get full credit with a score of 60 on SNPs, and no credit for a score of 40 or below.