Homework Project 2 - Megabase-scale alignment: 1-Million Length Genome

BIOINFO M122/M222

Due: Friday May 18th, 2021, 11:59 pm

This programming assignment is designed to expand your understanding of sequencing and the difficulty of mapping insertions and deletions.

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 that the data relates to. In the reference file, the reference 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. These reads are formatted as two 50 bp-long ends, which are separated by a 90-110 bp-long separator.

Starter Code

Starter code for the project has been pushed to the Github repository under **HP2** folder at https://github.com/rosie068/CM122_starter_code. Use `git pull` in the CM122_starter_code repository you cloned to get the updated code, or you can just redownload the files directly from the link. As with HP1, you should read the content of the HP2 code, and see if you can understand what it is doing. You should also look to see where your input/output is going to go.

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.

Again, do NOT download from the Heroku site, instead download your practice and for-grade data from:
10K Length Example Genome (Not for credit):
https://studentdownloads.s3-us-west-1.amazonaws.com/practice_W_3.zip

1-Million Length Example Genome (Not for credit):
http://studentdownloads.s3-us-west-1.amazonaws.com/practice_E_1.zip

HP2A 1-Million Length Genome (For Credit):
http://studentdownloads.s3-us-west-1.amazonaws.com/hw2undergrad_E_2.zip

HP2B 100-Million Length Genome (For Credit):
http://studentdownloads.s3-us-west-1.amazonaws.com/hw2grad_M_1.zip

We are providing you with the skeleton for one script:
1. `basic_hasher.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 scripts 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 10K length genome practice data provided for project 2.

1. Download the 10K practice data into the HP2 folder and unzip it. The commands below assume that you have a folder named practice_E_1 in the HP2 folder. If you download and save things in a different place you'll have to adjust the file paths below.

2. Use `basic_hasher.py` to align reads to the genome.

```
python basic_hasher.py -g ref_practice_E_1_chr_1.txt -r reads_practice_E_1_chr_1.txt -o test_output.txt -t practice_E_1_chr_1
```

This will generate a file of changes in test_output.txt and a zipped version of that file formatted correctly for submission.

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.

Smith-Waterman Reconstruction

For more enrichment on variant calling using the Smith-Waterman Algorithm, see UCLA Professor Chris Lee's lecture here: https://www.youtube.com/watch?v=EWJnDMKBEv0

Chapter 5 in the textbook also goes over these ideas.

Grading

** NOTE: The "HP2A 1-Million Length Genome" is for undergrads. The "HP2B 100-Million Length Genome" is for graduate students. **

SNP Score	No Credit	Full Credit
Undergrad	55	75
Grad	70	90

Indel Score	No Credit	Full Credit
Undergrad	3	13
Grad	15	25

Your total score will be the average of (Your Score - No Credit Score)/(Full Credit Score - No Credit Score) for both SNPs and Indels.