Programming Assignment 1 - 10 kilobase alignment

BIOINFO M260

Due: Tuesday January 24th, 2017, 11:59 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://cm124.herokuapp.com/upload` as a **.zip** file.

Start Here

`https://cm124.herokuapp.com/ans_file_doc` should handle most of your questions on reading and writing output.

Sequence Alignment

The trivial sequence alignment algorithm is to "slide" the read along the reference genome. This is written simply as a for loop:

```
... get reads as input ...

for read in reads:
    for i in range(len(genome) - len(read)):
    mismatches = [0 if genome[i + j] == read[j] else 1 for j in range(len(read))]
    n_mismatches = sum(mismatches)
    if n_mismatches < 3:
    ...
    add this to the output
    ...
    break

.... construct consensus sequence ...
```

The statement above has some useful Python tricks; instead of iterating only over indices, Python allows you to iterate over objects. If you store your reads in an array called `reads`, you can get each one using the word `read`, rather than using many different indices.

The statement `mismatches = [0 if genome[i + j] == read[j] else 1 for j in range(len(read))` is called a list comprehension, which borrows from functional programming. Essentially, this code is the same as:

```
output = []
for j in range(len(read)):
if genome[i + j] == read[j]:
output.append(0)
else:
output.append(1)
```

However, the syntax is much cleaner, and if you can read it, you'll be a much better programmer.

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.

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://cm124.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 is under week 2 in CCLE. You also have two sets of data with solutions (also under week 2 in ccle), that should be helpful for building and debugging your algorithms. You can also submit your solutions for those datasets to `https://cm124.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.