Homework 1: Sequencing and Assembly

(100 points total)

Assignment guidelines

1) Submit your assignment files on canvas under module 1: Sequencing and assembly
2) Please submit your code in a file called [name].py. Your code should be easy to open in a text editor so that someone can download and use the function you write.
3) Please submit a pdf with the answers to the questions at the bottom of the assignment (and your visualization)
4) Please submit a fasta file (text file) with the output of your code called [name].fa

Complete the class assignment (50 points)

5) Create a function called "assemble reads" which takes as input a file with reads and outputs a file with a FASTA sequence of contigs labeled numerically (40 points)
    a. Code meets specifications – the function exists and makes correct input and output (10 points)
    b. No pair of reads are incorrectly merged (10 points)
    c. Code generates accurate mapping (10 points)
6) Evaluate the distribution of the reads across each sequence (30 points)
    a. Create a visualization of your choice that shows coverage of reads across each of your sequences (10 points)
    b. Based on the visualization, is the sequencing method biased? Explain (10 points)
    c.

Consider the following about your code (20 points)

1) Change the overlap parameter (k) in your code
    a. At what point does the program output change when you decrease k? (5 points)
        i. Consider the assumptions you made in your algorithm – what's the issue? (5 points)
    b. At what point does the program output change when you increase k? (5 points)
    c. What is the relationship between how high k can go and sequencing coverage? (5 points)

Consider sequencing as a whole (30 points)

1) When we look at paired end reads we gain benefit from increasing L – the distance between the paired ends.

a. Explain how mate-pair reads enable us to get a higher L than paired end reads (5 points)
b. If your average sequencing fragment size is 400bp and your mate-pair selection library size is 3000bp and your read length is 80bp estimate your average distance L for paired end **AND** mate pair sequencing. Show your work for how you got both numbers (15 points)

2) Why is it better to generate reads from the same sequence (pac-bio) than generating the same amount of reads from the replicate of that sequence? (5 points)
3) Choose a type of human genomic variation besides single nucleotide polymorphism(SNP). Explain how long reads help detect that type of variation. (5 points)