Mia Tarantola
Progress Diary
CS321 Bioinformatics Final Project

Hour 1
Set out to: Review how to generate read pairs and paired debruijn graphs
Accomplished: Reviewed how to generate read pairs and paired de bruijn graph in the book chapters. Generated pseudocode for the necessary functions (paired composition, paired dbgraph, eulerian cycle/path, reconstruction…)

Hour 2
Set out to: Compile relevant code from previous Rosalind assignments
Accomplished: Pulled from Eulerian Path Rosalind, established code-sharing platform, researched more on read-pairs

Hour 3
Set out to: Generate Read Pairs from full data sequence
Accomplished: Pulled from Generate the k-mer Composition of a String Rosalind, established code-sharing platform, researched more on read-pairs

Hour 4
Set out to: Make paired DeBruijn graph
Accomplished: Successfully generated a paired DeBruijn graph using text example ("TAATGCCATGGGATGTT",k=3,d=1) worked for this scenario, but not others

Hour 5
Set out to: Fix paired DeBruijn graph to work with other inputs
Accomplished: Successfully refined paired DeBruijn graph code to also work for ("TAATGCCATGGGATGTT",k=3,d=1) *need to find other testers

Hour 6
Set out to: Write code to find eulerian path
Accomplished: successfully generated the in/out scores for each node and found the start and end nodes *had difficulty modifying code to work with a list as dictionary key (changed to tuple)

Hour 7
Set to: Create Eulerian Cycle
Accomplished: added edge from the end node to the start node, found the eulerian cycle *again difficulty dealing with tuples vs. lists

Hour 8
Set to: Create eulerian path and reconstruct genome

Accomplished: Split the eulerian cycle between start and end node to create an eulerian path, generating two reconstructed portions of the string (one using first kmer and one using second). Merged the two together (disregarding overlap (only seen once))

Hour 9
Set to: Meet with Prof. Linderman and figure out the next step for furthering our project
Accomplished: brainstormed ideas of data analysis: sweep k and sweep d and compare N50 or randomly generate portions of the genome and compare N50

Hour 10/11

Set to: divide the genome into reads of 1000 and for each 1000 reconstruct the genome with k =

3→ 30 with d = 4

Accomplished: tried to run our program on the rosalind problem extra practice data set, did not work as the eulerian function was stuck in an infinite loop. *Need to ask what to do if other nodes besides start and finish are not balanced.

Hour 12
Set to: have our program only run on reads that are not naturally occurring cycles
Accomplished: altered our program to run on reads that aren't cycles

Hour 13
Set to: run our program once and gather data
Accomplished: tried to run our data once, but splitting our entire genome into reads of 100 was unreasonable. Randomly generated 10 reads of length 1000 to analyze. Wrote to txt file for our data as k and d swept

Hour 14
Set to: Analyze data to figure out what range of k and d to utilize.
Accomplished 7 seems like the minimum length of k for correct reads (1 -> 25). Write to csv file instead of txt (easier to use data later). (office hours)

Hour 15/16
Set to: research how to make data visualizations
Accomplished: Imported csv file data into python and wrote code using matplotlib to generate our 4 graphs

Hour 17
Set to: embellish graphs for better readability
Accomplished: Researched how to bold text, increase size and add different elements to matplotlib graphs for more clarity

Hour 18
Set to: import graphs to our poster

Accomplished: graphs were extremely small and blurry, had to photoshop the text of the graph so students standing farther away could see (tried to increase dpi of graphs but wouldn't work)

Hour 19
Set to: write a concrete methodology and figure legends
Accomplished: wrote a concise methodology that explained how we modified our class programs to fit our new goal. Wrote figure legends that explained what conclusions could be made from each graph.

Hour 20
Set to: add comments to code so that it can be more easily understood. Compile all used code, txt file inputs and data for our final submission
Accomplished: went through all of the code and added comments so that each part of the code's purpose was clear. Went through and compiled all code and files used into one folder.