COMPSCI 260 - Problem Set 3, Problem 1

Due: Fri 11 Oct 2019, 5pm

Name: Kyle Mitra

NetID: km423

Statement of collaboration and resources used (put None if you worked

entirely without collaboration or resources; otherwise cite carefully): Office Hours / StackOverflow

My solutions and comments for this problem are below.

-----------------------------------------------------

a) C =

b) The probability of not a specific location in the genome not being covered is

[1 - ]R  which simplifies to e-C

From this probability, the expected number of unsequenced nucleotides is

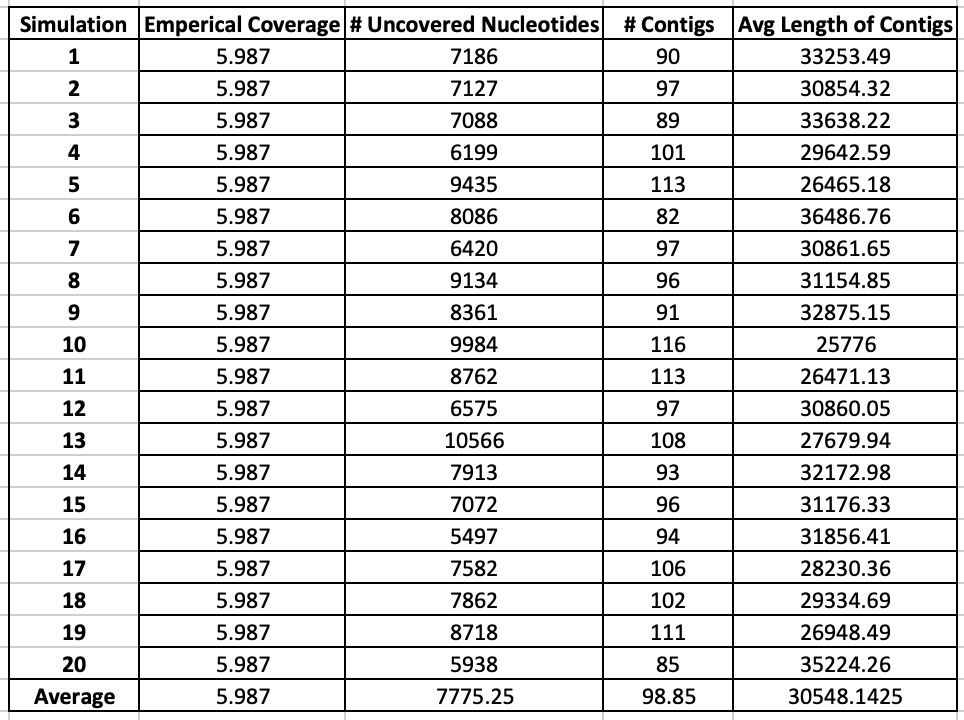
G \* e-C

c) The expected number of contigs is

R \* e-C

The expected length of each contain is

d) See code

e) 

f) The average values obtained from the 20 simulations is similar to the expected values obtained from the equations derived in parts (a), (b), and (c). The slight discrepancies between the numbers may be a result of an insufficient number of trials. Since there is large variation in some values between trials, many simulations would have to be run in order to get a value even closer to the expected value.

g) Based on the previously derived equation for the expected number of unsequenced

nucleotides is (G\*e^(-C)), we would expect about 1659253.11 unsequenced nucleotides. Based on the previously derived equation C = RL/G, the equation for R would be CG/L. Based on these values, we would require 3.75 x 10^7 reads.

h) The total number of read comparisons the assembler will need to undertake is R \* (R-1) or (3.75E7)\*((3.75E7)-1)

i) seconds or

j) Using the formulas we previously determined above, we are able to calculate these values using the newly provided information from part g.

When the assembler is finished, there will be about 20,741 contigs.

The average length of a contig will be about 144,563 nucleotides.

Using the formula above, there are expected to be approximately 1659253 unsequenced nucleotides. If we divide this number by the number of contigs, we get that there should be approximately 80 unsequenced nucleotides between adjacent contigs.

k) Based on the set of values I computed in this subproblem, I believe that these numbers are reasonable. One of the striking things about this task is that relative simplicity of the formulas. Once deriving the formulas, they are easily used to solve a very complex problem. When initially faced with the task at hand, this problem seemed extremely daunting but simply derivations of formulas allow you to break down this large task into smaller feasible tasks which when brought together allow you to solve this problem.