Problem Set 3, Problem 1 **Jihong Tang**
COMPSCI 260 NetID: jt290
Work Date: 10/05/2018 Due Date: Fri 10/12/2018 5pm
Extension Due Date: Sun 10/14/2018 5pm

Introduction to the problem

Problem 1 in problem set 3 is based on the whole-genome shotgun sequencing, covering several topics about the theory in this fasinating sequencing technology.

For question 1a, we need to provide the expression of the coverage C based on the given information about the length of genome G, number of sequence reads R, length of each reads L.

For question 1b, we need to write expressions for the probability that a specific location in the genome will not be covered by any of the R reads and the expected number of nucleotides in the genome that remain unsequences during this procedure respectively.

For question 1c, we should first understand the concept of contigs, which are the output of an assembly algorithm. we are asked to provide the expected number and length of contigs reproted by such an algorithm based on our understanding.

Problem 1a

I can proveide the expression for C from the given information as following:

$$C = \frac{L \cdot R}{G} \tag{1}$$

Problem 1b

Question 1

To work out the probability that a specific location will not be covered by any of the R reads, I first consider the probability that the specific location will not be covered by one read. Since the probability of the covered condition is $\frac{L}{G}$, the probability of not covered will be:

$$1 - \frac{L}{G} \tag{2}$$

Therefore, the probability that a specific location will not be covered by any of the R reads will be shown as following:

$$P_1 = \left(1 - \frac{L}{G}\right)^R \tag{3}$$

Substitute Equation1 into Equation3, we can get the function of R and C as Equation4.

$$P_1 = \left(1 - \frac{C}{R}\right)^R \tag{4}$$

Question 2

Using the probability worked out in Question 1, we cuold easily find the expression for the expected number of nucleotides in the genome that remain unsequenced as Equation5.

$$N_{not} = G \cdot P_1 \tag{5}$$

Substitute Equation1 and Equation3 into Equation5, we could get the expression as function of G, C and L as shown in Equation6.

$$N_{not} = G \cdot \left(1 - \frac{L}{G}\right)^{\frac{G \cdot C}{L}} \tag{6}$$

Since we konw the Equation7 from calculus and $L \ll G$, we could calculate part of Equation6 as shown in Equation8.

$$\lim_{x \to \infty} \left(1 - \frac{a}{x}\right)^x = e^{-a} \tag{7}$$

$$\lim_{\frac{G}{L} \to \infty} \left(1 - \frac{1}{\frac{G}{L}}\right)^{\frac{G}{L}} = e^{-1} \tag{8}$$

Substitute Equation8 into Equation6, we could get the expression as function of G and C in Equation9

$$N_{not} = G \cdot e^{-C} \tag{9}$$

Problem 1c

To work out the expressions in Problem 1c, I first consider the way to form contigs. I realized that there is one ”most right” read in each contig, therefore, the number of contigs equal to the number of the ”most right” reads. In addition, the probability that finding one ”most right” read is the same as that no read will start in the read of length L, which is the same as the probability that given position will not be covered during the procedure. Thanks to the work in problem 1b, we have found that probability as e^{-C}. Therefore, the expressions for the expected number of contigs and expected length of each contig will be shown as Equation10 and Equation11 respectively.

$$N_{contig} = R \cdot e^{-C} \tag{10}$$

$$L_{contig} = \frac{G \cdot \left(1 - e^{-C}\right)}{R \cdot e^{-C}} \tag{11}$$