Problem Set 3, Problem 2 **Jihong Tang**
COMPSCI 260 NetID: jt290
Work Date: 10/09/2018 Due Date: Fri 10/12/2018 5pm
Extension Due Date: Sun 10/14/2018 5pm

Introduction to the problem

Problem 2 from problem set 3 is based on the simulation of the sequencing process mentioned in problem 1.

For question 2a, we need to simulate the process with the given parameters in python and compute the different number from the previous problem. After the simulation, we are asked to compare the results to the results come from the expression we worked out in problem one.

For question 2b, we are asked to take a look at the situation faced by Celera in the late 1990s to get more insight into the development of genome sequencing technology.

Problem 2a

Implement consideration

Before the python implement of the simulation process, I consider the way to address edge effects and make my own choice.

In my implement process, I deleted the first and last $L-1$ elelment before I compute the several parameters. After the simulation, the first and last part of the result list will lead to the edge effect. Since $L \ll G$, I believe it will not effect the whole result after my deletion operation.

Implement results

The code solution of the simulation implement in python can be found in **simulate.py**. The results from my code can be illustrated as following:

1. The empirical coverage across each of my 20 simulations:

$$[6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6]$$

2. The number of nucleotides which are not cover by any read across each of my 20 simulations:

$$[7718, 7758, 6239, 7092, 5702, 7365, 8621, 8772, 6749, 7536]$$

$$[10962, 9312, 7364, 7509, 7229, 6629, 7951, 8399, 6614, 6885]$$

3. The number of contigs in each of my 20 simulations:

$$[103, 98, 87, 92, 84, 91, 104, 109, 105, 92, 120, 122, 92, 101, 112, 100, 103, 101, 99, 102]$$

4. The average length of contigs in each of my 20 simulations:

$$[29042, 30523, 34400, 32521, 35635, 32876, 28754, 27434, 28498, 32517]$$

$$[24901, 24506, 32518, 29619, 26713, 29924, 29040, 29610, 30227, 29335]$$

The compare between the computation value from the expression in probelm 1 and the simulation value after the averaging operation can be shown as Table1. The equations used to compute the variables can be found as Equation1 to Equation4.

Table 1: Compare between computation and simulation results

Variable	Computation Result	Simulation Result
empirical coverage(C)	6	6
number of missed nucleotides(N_{not})	7436	7620
number of contigs(N_{contig})	99	100
average length of contigs(L_{contig})	30182	29929

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

$$N_{not} = G \cdot e^{-C} \tag{2}$$

$$N_{contig} = R \cdot e^{-C} \tag{3}$$

$$L_{contig} = \frac{G \cdot \left(1 - e^{-C}\right)}{R \cdot e^{-C}} \tag{4}$$

Implement analysis

Firstly, the code analysis: the runtime for my whole program is about 65 seconds. I used regular expressions to find all zeros in my result list and do the following computation.

Secondly, the result analysis: as can be seen in Table1, I got the similar numbers using my simulation procedure. I have run my code for several times, and the simulation results all locate near the computation result, some are bigger and some are smaller. Therefore, I think the little descrepancies between the computation and simulation results come from the random simulation process.

Problem 2b

After considering the condition that $G = 3 \times 10^9, \quad C = 7.5, \quad L = 600$, I computed the results as shown in Table2.

Table 2: Problem 2b results

Variable	Result
number of nucleoides remaining unsequenced	1659253
numer of reads	3.75×10^7
number of read comparisons	2.81249×10^{15}
time of read comparisons	$5.62498 \times 10^7 s$
number of contigs	20740
average length of a contig	144563
average length of unsequenced region	80

Table2 contains the results for problem 2b. When I look at the set of values, I found that these values are all very huge compared to the simulation process we made. However, I think it is reasonable to some extent because we are focusing on the whole human genome, it is true that such things are huge.

In addition, I am stoken by the time of the first step read comparison since it is almost two years' work if all reads being compared to each other. I think it is a huge work and it must exist some methods for scientists to speed up this process.

About the whole task done by Celera in the late 1990s, I believe it is such a fantastic work that all bioinformatics scientists should remember them and take some valuable methods for future use from this procedure.