Problem Set 2, Problem 1 **Jihong Tang**
COMPSCI 260 NetID: jt290
Work Date: 09/22/2018 Due Date: Fri 09/28/2018 5pm

Introduction to the problem

Problem 1 from problem set 2 ask us to consider a sequence mapping problem.

For question 1a, we need to work out the expected number of occurences of a read of length m in a genome of size n assuming a uniform nucleotide distribution.

For question 1b, we need to work out a brute force algorithm to solve this problem and analyze its worst case running time.

For question 1c, we need to compare the brute froce algorithm with the given preprocessing approach and find the treshold number of reads for choosing the better approach.

Problem 1a

To slve this question, I divided the calculation process to get the number of occurence of a given m length sequence into two parts:

1. Calculate the total number of the sequence of length m in the genome of size n.
2. Calculate the probability that a random sequence of length m is the same as the given read sequence.

Therefore, my answer to this problem is shown as following:

$$N_{occurence} = (n - m + 1) \cdot (\frac{1}{4})^m$$

Problem 1b

Algorithm pseudo-code

To use brute force algorithm to solve the mapping problem, I decided to check whether the m length sequence in the n length genome is the same as the given one by one from the first site. To check their similarity, I still check whether the nucleotides are the same one by one from the first site. The detailed pseudo-code is shown as Algorithm1 and the algorithm will return the position index of the m length sequence in the genome when the mapping is successful.

Algorithm 1 Brute Force Algorithm

```
for i = 1 to n - m + 1 do
  index = i
  for j= 1 to m do
    if Genome[ index ] == Sequence[ j ] then
      index = index + 1
    else
      break
    end if
    if index - i == m then
      return i
    end if
  end for
end for
```

Algorithm analysis

Firstly, I believe my algorithm above can solve the question correctly because all possible subarraies have been checked in the algorithm. It may cost more time, but the correctness can be guaranteed.

The worst case in the brute force algorithm will be to check all possible m length sequence in the genome and each one checked $\Theta(m)$ times. Therefore, the time total will be $(n - m + 1) \times \Theta(m)$. To be more exact, it will be $\Theta(mn - m^2)$. Since m is less than or equal to n, it can be simplified to $\Theta(mn)$.

Problem 1c

From the given information, I can work out the total time for mapping k reads using the pre-processes approach:

$$\Theta(n \log n) + k \cdot \Theta(m \log n)$$

Also, the total time using brute force algorithm will be:

$$k \cdot \Theta(mn)$$

I put these two time into comparison and find the trade-off k value as following:

$$k = \lceil \frac{n \log n}{m(n - \log n)} \rceil$$

Therefore, when $k \geqslant \lceil \frac{n \log n}{m(n - \log n)} \rceil$, it is worth using the pre-process approach to map k reads. And the trade-off number of reads is $\lceil \frac{n \log n}{m(n - \log n)} \rceil$.