Kyle Mitra

**Problem 1:**

Whole-genome shotgun sequencing:

* Several copies of a genome of length G
  + Broken at many points uniformly at random
* Location of each fragment produced is independent of the locations of other fragments
* Edge effects can be ignored
  + When object has different behavior near the boundary (beginning and end of genome)
    - The first nucleotide will be sequenced only when the region from 1 to L is sequenced
    - The probability of covering the first position in genome is smaller than the probability of covering the ith position
* The process of cloning the fragments into vectors is perfect
* From each cloned fragment, exactly L nucleotides are sequenced from one end of the fragment, yielding a total of R sequence reads of length L (where L << G)
* A contig is a set of contiguously assembled nucleotides.
  + Assume oracular assembly algorithm which can perfectly assemble the reads it is given, without needing to satisfy any minimal length overlap requirements
  + Knows precisely where every read should go in the genome
* R = number of sequence reads
* L = number of nucleotides in a read (the length)
* G = size of genome

1. C = RL/G
2. Probabilty of not covering a specific location = [1- C/R]^R or e^-C

^^^^\*Check

Number of unsequenced nucleotides = G \* exp(-C)

1. The expected number of contigs is Ne^(-a) **Re^(-C)**

The expected length of each contig is [g(1-e^(-a))]/(Ne^(-a))

**(G(1-e^-C))/(Re^(-C))**

N = number of fragments chosen at random, cloned, and sequenced

G = length of DNA segment in number of bases

a is the number of fragments that cover the point x such that a = NL/g

\*N = R

\*a = C

d) See Code

e) See Table

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!.

i) It would take about (R! / 50 million) seconds.

j) When the assembler is finished,

h) Use values from part G

**Problem 2:**

\*See lecture videos if necessary

* Burrows-Wheeler Transform (BWT) also referred to as block-sorting compression
  + Short-read mapping
  + Data compression
* Essentially, the BWT manipulates a block of input text using a reversible transformation that does not itself compress the text, but reorders it in a manner that facilitates compression
* For DNA sequences, they are rearranged so that the output of the transformation is a sequence that contains long runs of similar characters
  + Ex: TATCGTACACTACGTACGA$ goes to AGTTTCTAATAACCCGGC$A
  + $ is the End of File character

1. T$CGTTGCA
2. Use append and sort method

**Problem 3:**

* Your policy whenever a patient comes to see you is to order a routine analysis of their gut microbiome.
* Stool sample analyzed by purifying the genomic DNA, randomly fragmenting all that DNA, and generating short reads from the fragments
* Three patients show up
  + Patient 1 and 2 are asymptomatic
  + Patient 3 is complaining about abdominal pain and severe diarrhea causing dehydration and electrolyte imbalance

1. *Bifidobacterium longum* is a non-pathogenic, mutualistic bacterium that is believed to prevent growth of pathogenic organisms via production of lactic acid. Although it is not especially prevalent in the human adult gut, it is considered to be one of the earliest colonizers of the infant GI tract.
2. *Eubacterium rectale* is a normal component of the human gut microbiota, thought to produce butyrate, which is an important energy source for cells in the colonic epithelium. Its prevalence in the gut appears to be similar to that of individual species from the Bacteroides genus.
3. *Roseburia intestinalis* breaks down sugars in order to produce butyrate, thus possessing a protective effect against colon disease by nourishing colonocytes (similar to E. rectale). Depressed levels of R. intestinalis are associated with various bowel conditions, while elevated levels may provide protection against atherosclerosis (at least in mice) and may promote weight loss.
4. *Baceroides ovatus* is an anaerobic, non-spore-forming, nonmotile, and Gram-negative rod bacterium. It is the predominant commensal instinal microbe which causes a systemic antibody response in inflammatory bowel disease.

*https://www.ncbi.nlm.nih.gov/pmc/articles/PMC119885/*

1. *Bacteroides thetaiotaomicron* is a Gram-negative anaerobe. It is one of the most common components of the human gut flora. It specializes in the digestion of polysaccharides and provides food for other components of the microbiome. As an opportunistic pathogen, it also has extreme diseases causing potential as well as antibiotic resistance.

*https://en.wikipedia.org/wiki/Bacteroides\_thetaiotaomicron*

1. *Lactobacillus acidophilus* is a Gram-positive bacterium which grows rapidly at low pH values and around 37 degrees Celsius. Some strains are considered to have probiotic characteristics and are able to make lactic acid by breaking down carbohydrates. This bacterium also has medical uses and is sometimes prescribed to treat diarrhea.

*https://www.urmc.rochester.edu/encyclopedia/content.aspx?contenttypeid=19&contentid=Lactobacillus*

1. *Peptoniphilus timonensis*  is a Gram-positive and anaerobic bacterium. Its presence is sometimes reported in diabetic skin and soft tissue infections, bone and join infections, surgical site infections, chorioamnionitis, and bloodstream infections.

*https://en.wikipedia.org/wiki/Peptoniphilus*

1. *Prevotella copri* is a gram-negative bacteria which is anaerobic and does not form spores. The cells are rod shaped and grow best at 37 degrees Celsius. Some variants of the species may likely be involved in the onset of Rheumatoid Arthritis. Some studies also show that the presence of Prevotella copri causes a reduction of other beneficial microbes.

*https://microbewiki.kenyon.edu/index.php/Prevotella\_copri*

1. *Ruminococcus bromii* is a dominant member of the human gut microbiota that plays a key role in releasing energy from dietary starches that escape digestion by host enzymes via its exceptional activity against particulate “resistant” starches. It is a keystone species in the breakdown of resistant starch in the human large intestine. It is a specialized amylolytic bacterium known for its ability to degrade cellulose.

*https://mbio.asm.org/content/6/5/e01058-15*

1. *Vibrio cholera* is a bacterium commonly found in saltwater. Some strains of this bacterium can cause cholera, an infection of the small intestine which can induce diarrhea which can lead to severe dehydration and electrolyte imbalance, vomiting, and muscle cramps. It is a facultative anaerobe and can undergo respiratory and fermentative metabolism.

[*https://en.wikipedia.org/wiki/Vibrio\_cholerae*](https://en.wikipedia.org/wiki/Vibrio_cholerae)

Notes:

* Exact read-matching algorithm discussed in class which iterates through a given query sequence in reverse and repeatedly calculates the ever-shrinking range of rows and prefixes match successively longer suffixes of the query.
* While there are still characters in the query sequence to check, we update a pair of indices demarcating the beginning and end of the range of sorted suffixes that start with the suffix of the query corresponding to this iteration.
* After the final step of the iteration (in which we have finished considering the largest possible query suffix), if we find that the range of sorted suffixes contains at least one row with a match to the query, we can use the range indices to obtain the actual locations of those matches in the reference genome, locations which are themselves stored in the suffix array.

Write program which takes a query sequence and the necessary data structures containing various information about the reference genome, and returns a list containing all locations of the query in the original reference genome.