CSC2417: Assignment 3

Name: Juliana De La Vega Fernández
Student id: 1003092468
Email: j.delavegafernandez@mail.utoronto.ca

1.a) Using the probability that one of the N reads start at any nucleotide is:

$$p = \frac{N}{G}$$

The number of trials within the interval of length L would correspond to:

$$n = L$$

The number of successful events for no reads starting at the interval would be:

$$k = 0$$

Then the binomial distribution would be:

$$P(no\ reads\ in\ L) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$P(no\ reads\ in\ L) = \frac{n!}{k!\,(n-k)!} p^k (1-p)^{n-k}$$

Replacing n, k and p:

$$P(no\ reads\ in\ L) = \frac{L!}{0!\,(L-0)!}\left(\frac{N}{G}\right)^0 \left(1-\left(\frac{N}{G}\right)\right)^{L-0}$$

$$P(no\ reads\ in\ L) = \frac{L!}{L!} \cdot 1 \cdot \left(1-\left(\frac{N}{G}\right)\right)^{L}$$

obtaining:

$$\boxed{P(no\ reads\ in\ L) = \left(1-\frac{N}{G}\right)^{L}}$$

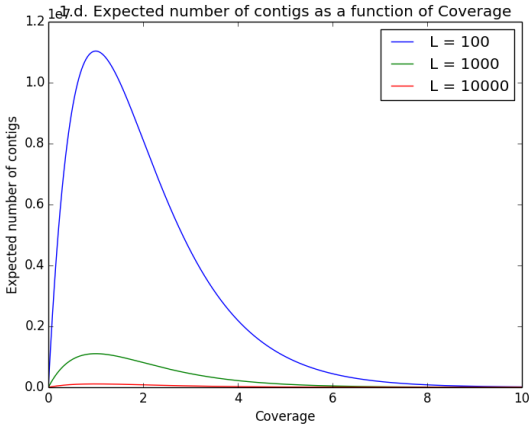
1.b) The percentage of the genome not sequenced is e^{-a}
The percentage of the genome we expect is 99.99%, which gives:

$$0.0001 = e^{-a}$$
$$\ln(0.001) = -a$$
$$a = -\ln(0.001)$$

1.c) The percentage of the genome not covered was 0.000083%, which is less than 0.0001%, obtaining more coverage than expected. Poisson's approximation was very precise. As G (the genome size) gets larger than L, the N reads will cover more of the complete genome(G). The N reads were calculated using the coverage formula, which would increment proportionally to the size of the genome. As G is incremented, the obtained percentage of uncovered nucleotides is closer to the poisson model's approximation.

```
# File:
./shotgun_sequencing.py
```

1.d) Plot of the Expected number of Contigs as a function of coverage:



#file
./expected_contigs.py

2.a) Conditional probability is described as the probability that A occurred under the condition B. Mathematically, it is the joint probability divided over the probability of the condition:

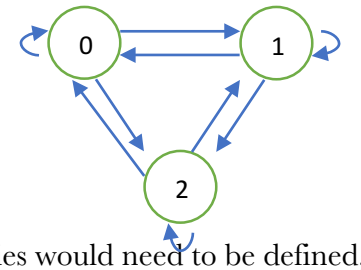
$$P(A|B) = \frac{P(A,B)}{P(B)}$$

The conditional probability is directly proportional to the joint probability, so maximizing $P(A,B)$ would maximize $P(A|B)$. So:

$$p^* = \arg\max P(p|x) = \arg\max P(p,x)$$

2.b)

i. The HMM model would contain three states described by: 0 (deletion), 1 copy (no change), 2 copies (duplication). The emission would correspond to the number of sequences mapped. For the emission distribution, the conditional distribution of the observed amounts of reads from a specific state are captured. We would need to determine in a given window how many reads are required for there to be a deletion, how many reads would be the normal amount, and how many would account for a duplication. Assuming we have normal distribution, we could take the number of reads in each window and obtain the standard deviation to determine which counts would be in a certain state. It could be that everything below -1 standard deviations counts as a deletion, everything between -1 and +1 standard deviations represents no change, and everything above +1 standard deviations is a duplicate. The state to state transitions would be:



The transition probabilities would need to be defined, probably from the data observation to determine how likely it is to change from one state to another, or to remain in the same state. In other words, the probabilities for each of the arrows would need to be defined.

ii. With the labeled data, I would get the average of the number of reads for every state to state transition using part of the labeled data to determine the transition probability. I would also average the number of reads that are expected for a deletion, for 1 copy and for 2 copies to determine the expected mapped reads. With the second part of the labeled data, I would validate how well the average estimates the number of reads correspond to a deletion or a copy, and how probable it is for it to transition from a deletion, or a single copy, or a duplicate, to a new state or the same state as before. I would then adjust these probabilities using the first set and test them again in the second set of the labeled data to ensure that it is obtaining the desired result. Once the model has been appropriately adjusted, it can now test unlabeled data.

iii. To train the parameters of the model without knowing their labels we could use the forward-backward algorithm, where $\alpha(z_n)$ is the joint probability of observing $x_1, \dots, x_n$ and being in state z_n:

$$\alpha(z_n) \equiv p(x_1, \dots, x_n, z_n)$$

Likewise, $\beta(z_n)$ is the conditional probability of future observation $x_{n+1}, \dots, x_N$ assuming being in state z_n

$$\beta(z_n) \equiv p(x_{n+1}, \dots, x_N | z_n)$$

Using the joint and the conditional probability, the likelihood of the observations can be obtained as:

$$p(X) = \sum_{z_n} \alpha(z_n)\beta(z_n)$$

$$p(X) = \sum_{z_N} \alpha(z_N)$$

In this case, the joint probability is the forward algorithm, and the conditional probability is the backward algorithm. The forward-backward algorithm is useful for determining the likelihood of any sequence of observations and predicting the next observation in that given sequence.

iv. It is expected that reads will map to the regions with a deletion, just in a lower than expected amount. This is because even if a deletion is present, we would have two strands of the sequence, and one of the strands may not have that deletion. In the case where the deletion is present in both strands, we could still have a copy of this region in the other chromosome. The other chromosome would account for some copies of the region, but the count would be less than the usual.

v. We would have to make a more complex model in which we consider the regions in which the GC content is higher to down regulate the sampling size in these zones. We could set the probability of getting into these GC state in such a way that the over-represented data is considered and the number of mapped reads in the region requires is to be higher than the expected to determine each state. We could also detect the regions with over-represented data and perform a down-sampling of the read quantity so that it represents more accurately the actual expectation of sampling in the region. This would allow to determine the state of the read with less bias.