Assignment 1

Mytraya Gattu, 180050032

22 August 2020

All files pertaining to the assignment can be found at `https://github.com/LordThunder333/PH549-assignment1.git`

1 Problem 1

In what follows, r is the radius of the hemispherical caps,l the length and m of the cell. It is given that,

$$r = 0.5\mu\text{m}$$

$$l = 2\mu\text{m}$$

(a) Volume is given by

$$V = \pi r^2 l + \frac{4}{3}\pi r^3 = 2.094\mu\text{m}^3$$

Typical cell volume is $0.6 - 0.7\mu\text{m}^3$
`https://en.wikipedia.org/wiki/Escherichia_coli`

(b) m is the sum of the mass of water contained and that of the other substances. Since, density of water is

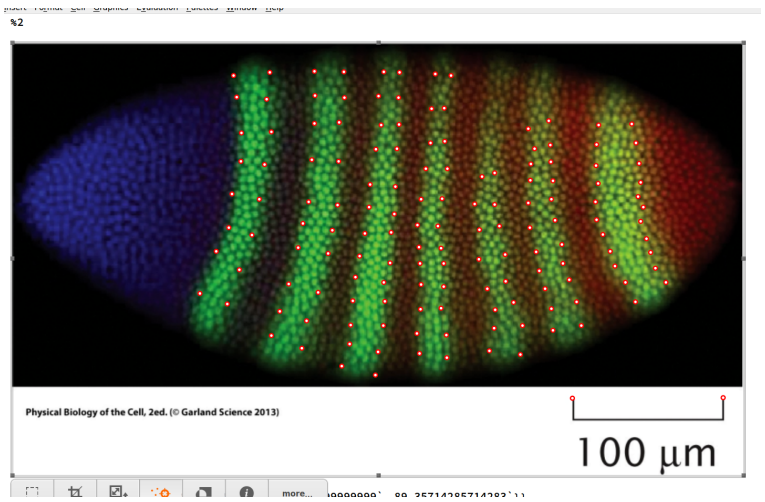
$$10^3 \frac{\text{kg}}{\text{m}^3} = 10^{-6} \frac{\text{pg}}{\mu\text{m}^3}$$

We have

$$m = 10^{-6} \cdot \left(\frac{2}{3} + 1.3 \cdot \frac{1}{3}\right) \cdot V\text{pg} = 2.304 \cdot 10^{-3}\text{pg}$$

Typical cell weight is $1 \cdot 10^{-3}$pg
`https://ecmdb.ca/e_coli_stats)`

It is evident from the above calculations, that the it is the assumed shape of *E. coli* which deviates greatly from what is true.



2 Problem 2

Let at the end of 9^{th} cycle there be N nuclei at the surface. At the end of 13^{th} cycle, therefore there are $16N$ nuclei, which is given as ≈ 6000. Hence,

$$N \approx 375$$

Since, every embryo must start from a single fertilized nucleus, at the end of 9^{th} cycle, there must be $2^9 = 512$ nuclei in total. Therefore, the fraction of nuclei which migrated to the surface is

$$\frac{375}{512} \approx 0.7324$$

Following the notation of Problem 1, surface area of the spherocylinder is

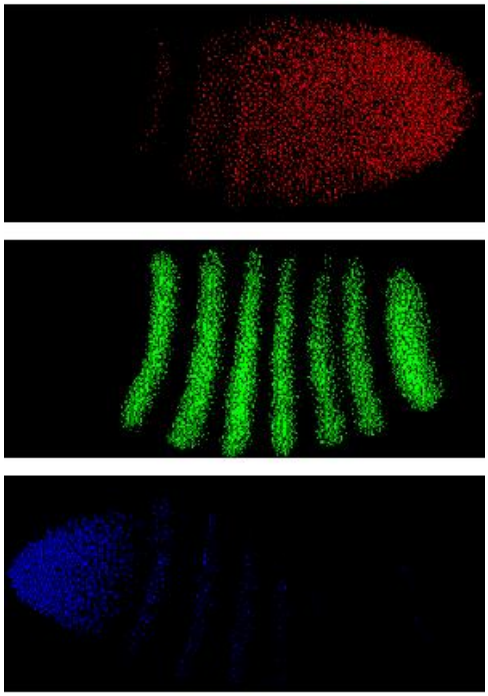
$$A = 4\pi r^2 + 2\pi rl = 4.3982 \cdot 10^{-7}\text{m}^2$$

Therefore, the areal density is

$$1.3642 \cdot 10^{10} \frac{\text{nuclei}}{\text{m}^2}$$

(For the next part, I have used Mathematica.) To calculate the length of the green strips, I marked the ends of widths along the lenghts of each strip, and then used the given scale.
Length of the green strip is: $18.37 \pm 4.99\mu$m To evaluate the number of nuclei in each strip, I isolated the colors, and counted the number of pixels.
Therefore, the number of nuclei corresponding to each dye(I'm considering only a single surface – In the diagram, the surface of the cylinder has been projected onto a rectangle)

- Red: 932
- Green: 1320
- Blue: 364



3 Problem 3

3.1 E. coli

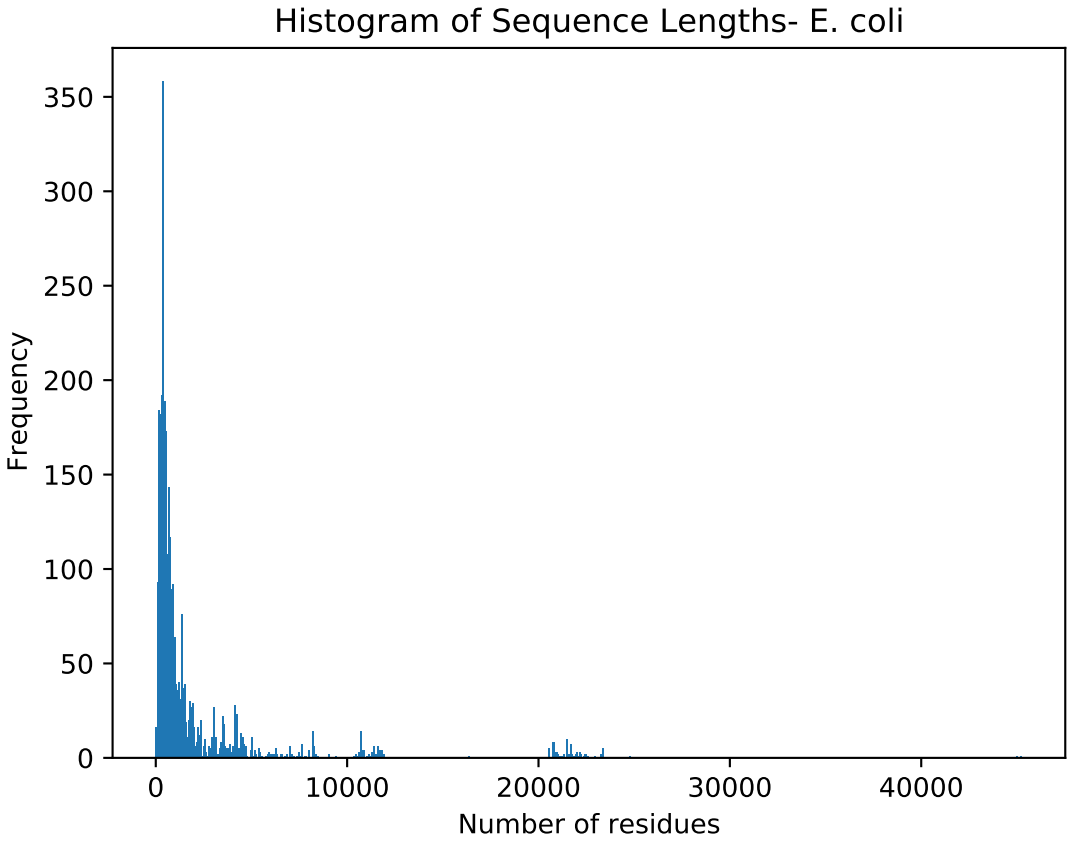
Mean of sequence length: 1892.00
Standard deviation of sequence length: 3872.73
Mean of molecular mass: 289.03
Standard deviation of molecular mass: 775.76

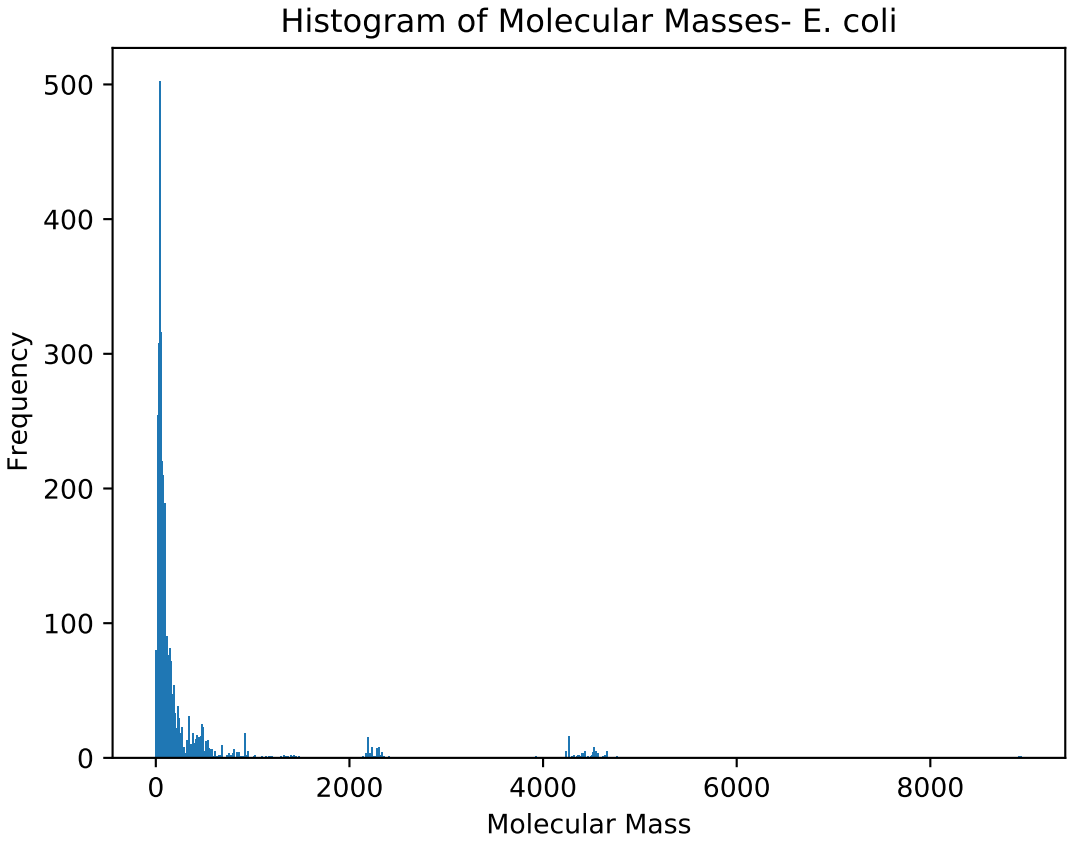
3.2 S. cerevisiae

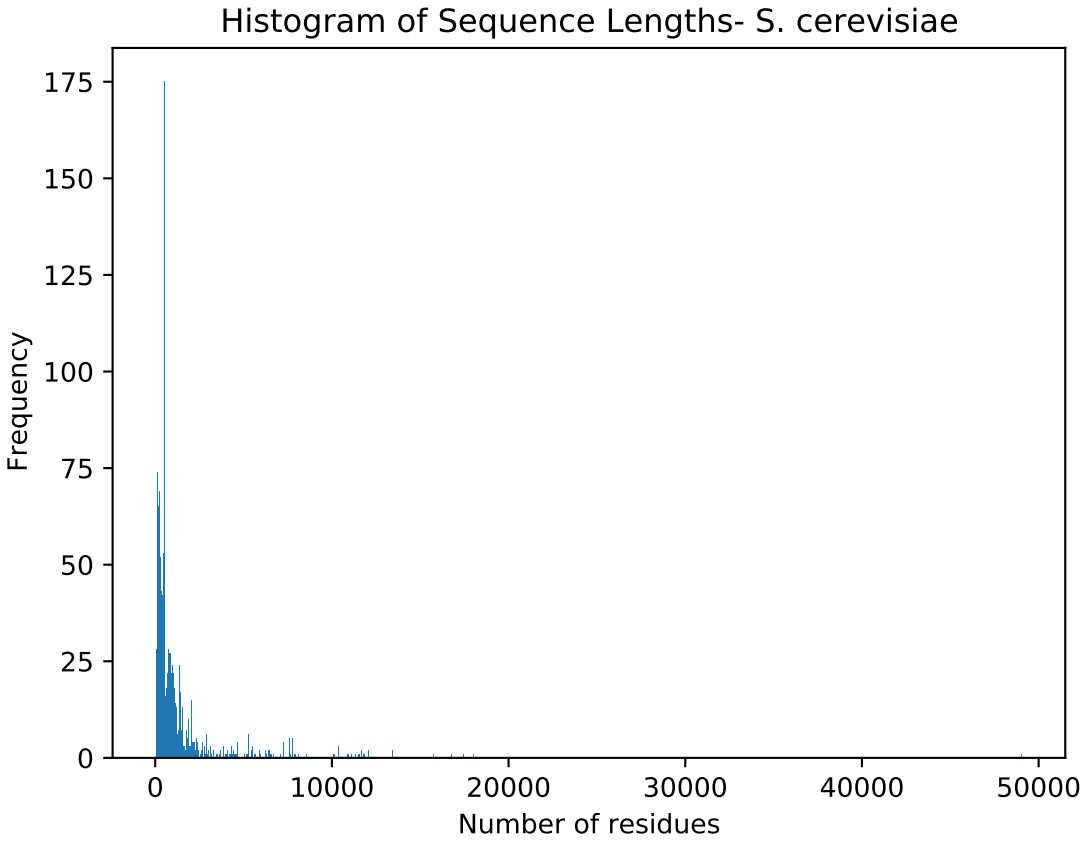
Mean of sequence length: 1287.90
Standard deviation of sequence length: 2444.49
Mean of molecular mass: 154.78
Standard deviation of molecular mass: 313.31

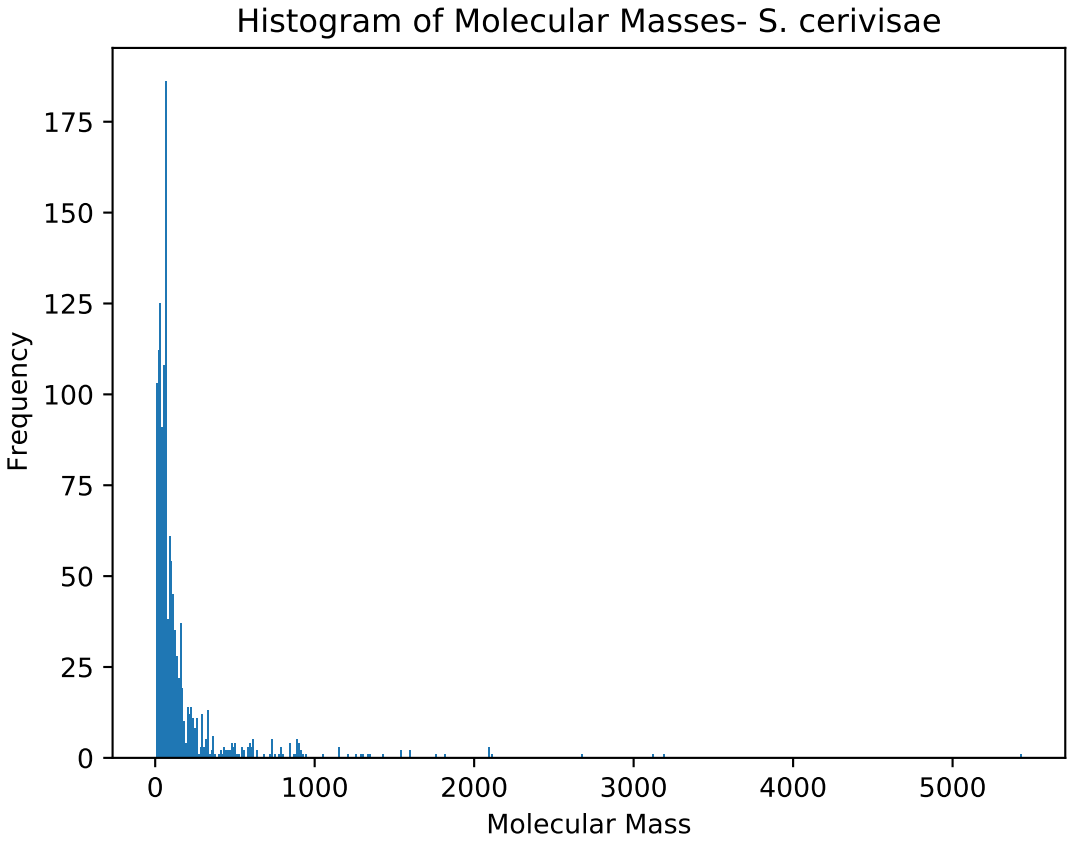
3.3 Humans

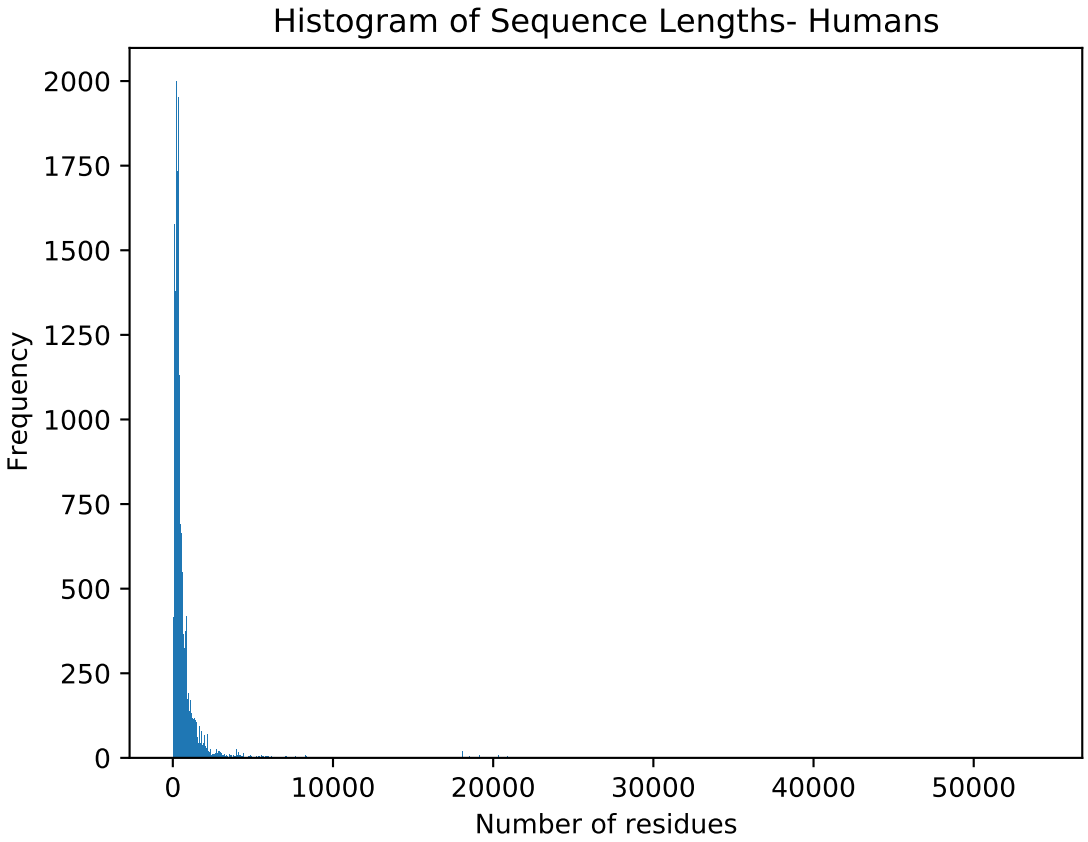
Mean of sequence length: 630.74
Standard deviation of sequence length: 1263.08
Mean of molecular mass: 74.21
Standard deviation of molecular mass:181.93

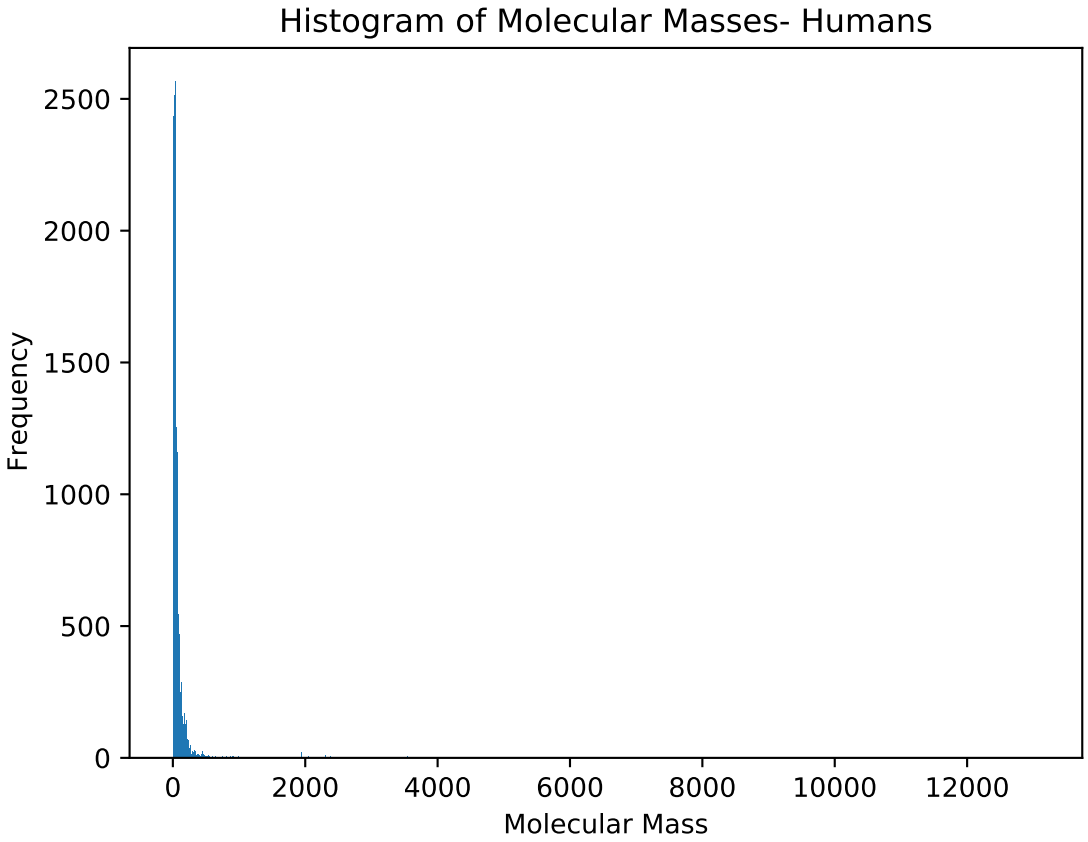

Histogram of Sequence Lengths- E. coli
Frequency
Number of residues
0
50
100
150
200
250
300
350
0
10000
20000
30000
40000


Histogram of Molecular Masses- E. coli
Frequency
Molecular Mass
0
100
200
300
400
500
0
2000
4000
6000
8000


Histogram of Sequence Lengths- S. cerevisiae
Frequency
175
150
125
100
75
50
25
0
0
10000
20000
30000
40000
50000
Number of residues


Histogram of Molecular Masses- S. cerivisae
Frequency
175
150
125
100
75
50
25
0
0
1000
2000
3000
4000
5000
Molecular Mass


Histogram of Sequence Lengths- Humans
Frequency
2000
1750
1500
1250
1000
750
500
250
0
0
10000
20000
30000
40000
50000
Number of residues


Histogram of Molecular Masses- Humans
Frequency
2500
2000
1500
1000
500
0
0
2000
4000
6000
8000
10000
12000
Molecular Mass

4 Problem 4

The probability of having 0 and 6 is

$$p^3q^3$$

. The probability of having 1 and 5 is

$$\binom{3}{1}p^4q^2 + \binom{3}{1}p^2q^4$$

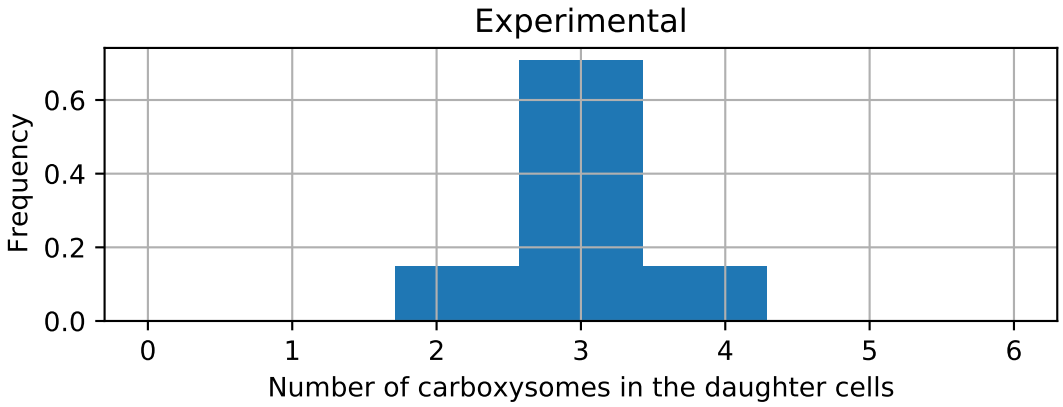
The probability of having 2 and 4 in one cell is

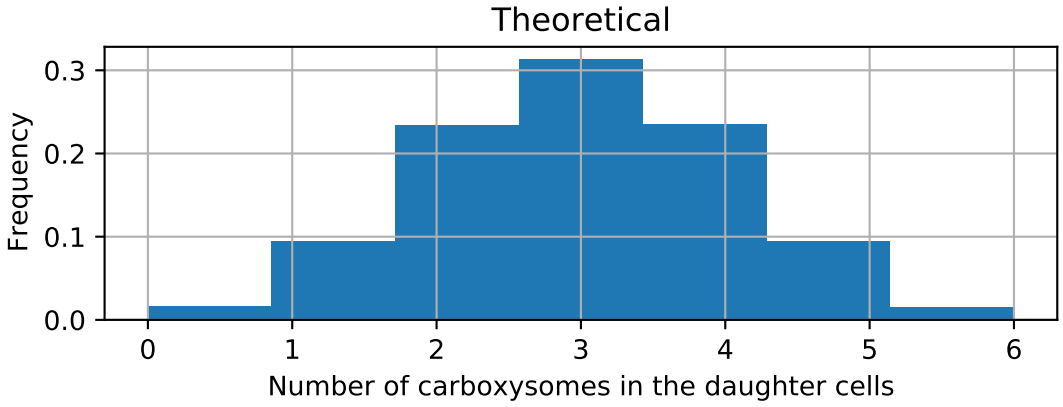
$$\binom{3}{2}p^5q^1 + \binom{3}{2}p^1q^5 + \binom{3}{1}\binom{3}{1}p^3q^3$$

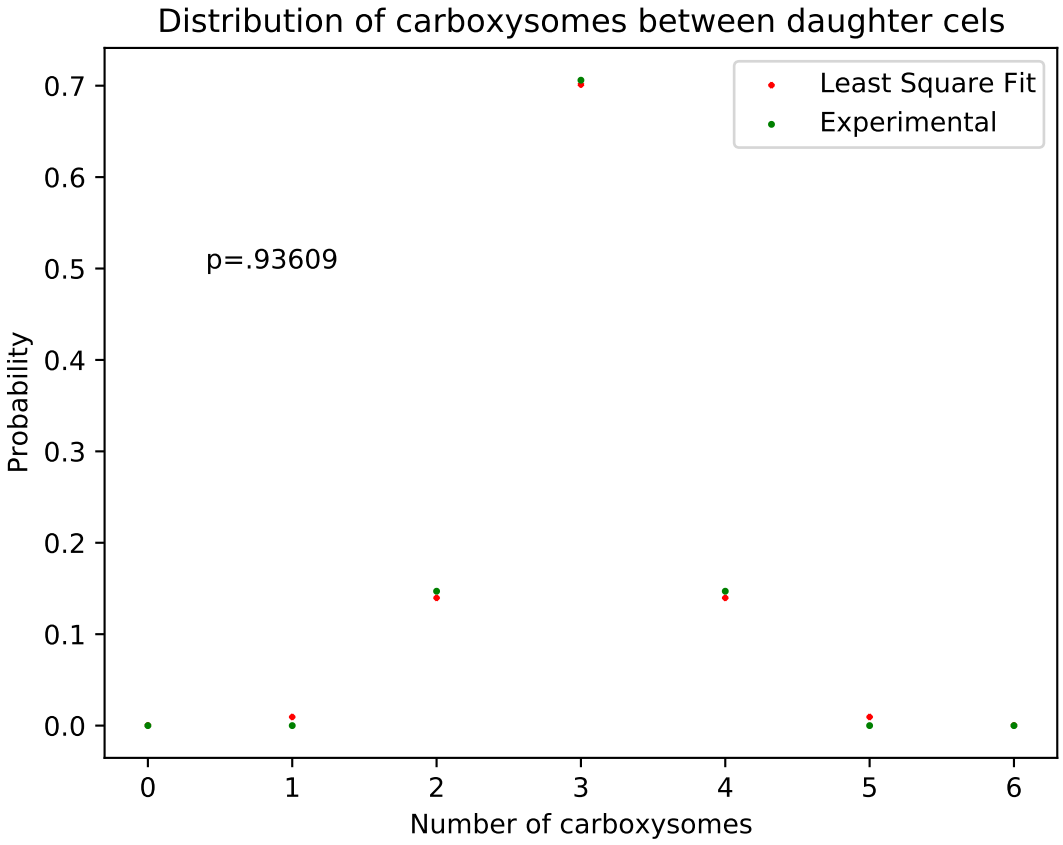
The probability of having 3 in one cell is

$$p^6 + q^6 + \binom{3}{1}\binom{3}{2}p^2q^4 + \binom{3}{2}\binom{3}{1}p^4q^2$$

By substituting $p = 0.5 + x$, and minimizing using the least-squares method with respect to the experimental distribution, obtained value for p is 0.936093. Attached, is a scatter plot showing the deviation. Given that, there are only 6 molecules, and the error is about 0.007 ± 0.001 (for the non-zero values), I believe it's a good model.


Experimental
Frequency
0.0
0.2
0.4
0.6
0
1
2
3
4
5
6
Number of carboxysomes in the daughter cells


Theoretical
Frequency
0.0
0.1
0.2
0.3
0
1
2
3
4
5
6
Number of carboxysomes in the daughter cells


Distribution of carboxysomes between daughter cels
Least Square Fit
Experimental
p=.93609
Probability
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
1
2
3
4
5
6
Number of carboxysomes

5 Problem 5

(a) There are 4^3 possible ways of choosing the bases to form a triplet, each with the same probability. Therefore,

$$p_s = \frac{3}{64}$$

(b) There are 64 possible codons. Three of these are not allowed(The stop codons), to form an ORF.Therefore, the number of acceptable possibilities is $(64-3)^N$.
Required probability is $\left(\frac{61}{64}\right)^N$.

(c) Codons are of length 3 bases. Since, the DNA is circular, let us pick an arbitray base and call it the starting point, and label the bases as

$$\dots x_{N-1}x_N x_0 x_1 x_2 \dots$$

Looking at the above sequence, we can see that the codons containing x_0 are

$$x_{N-1}x_N x_0$$
$$x_N x_0 x_1$$
$$x_0 x_1 x_2$$

and their reverse(corresponding to either a clockwise read or an anti-clockwise read – Also respectively, each codon corresponds to reading frames: +2,+1 and +0). Therefore, there are 6 reading frames in total.

(d) Exact length of the E. coli genome is 4639675 (As per the data in U00096.fna) base pairs, whose remainder with respect to 3 is 1.
Since, when chosen at random, choice of reading frame doesn't matter, let's suppose that I have picked one already. In any reading frame, because of the non-zero remainder, when codons are selected, one base pair will remain ungrouped. I select this as my origin and group bases accordingly. The sequence will look like

$$\dots |XXX|X|XXX| \dots$$

Total number of codons formed is 1546558. The above problem is extremely complex when evaluated directly. Therefore, I did the following: Generate a random genome: Construct a frequency table for ORF lengths.(I performed 1000 iterations)

By looking at the data, I concluded that the distribution is exponential. By fitting the curve, I obtained the following:

$$p(x) = 0.041491476584857405 * 0.9540938153337749^x$$

. Here x is the length of the ORF and $p(x)$ it's probability of occurence. (Attached are images, which depict the distribution obtained from the random genome generation and the comparision between the chosen distribution and the former.) Using this, the number of ORFs of length 600 that can be expected to occur on the basis of chance is $3.647e-08$, for the given length of the genome.

(e) Take a look at the attached ipython notebook.

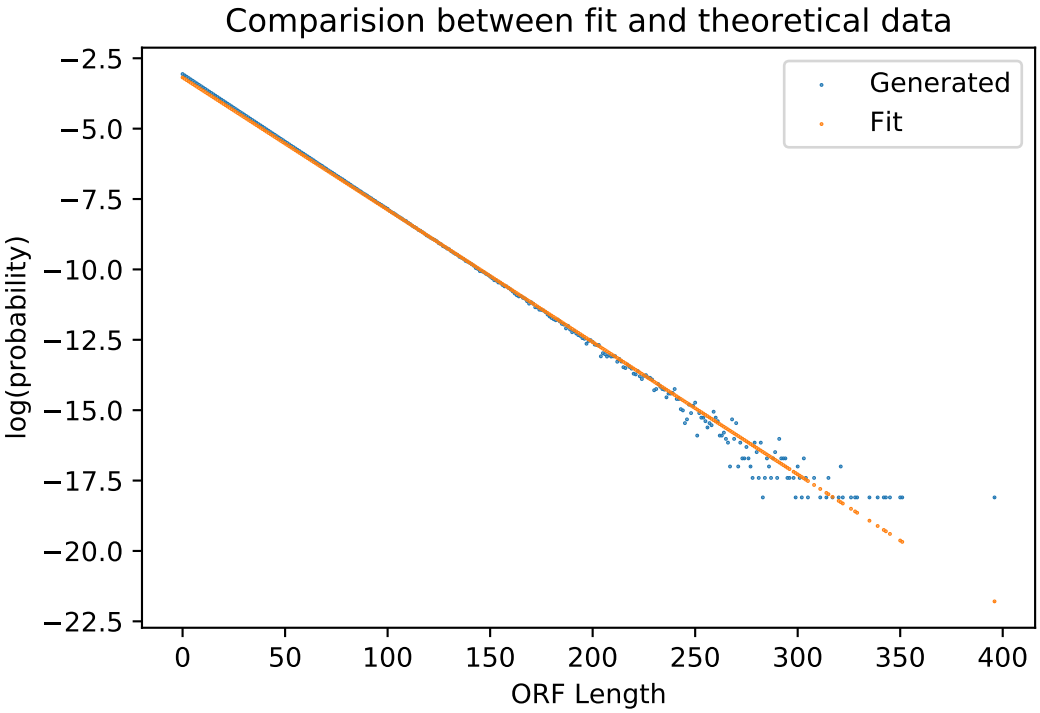
(f) The figures are placed at the end.

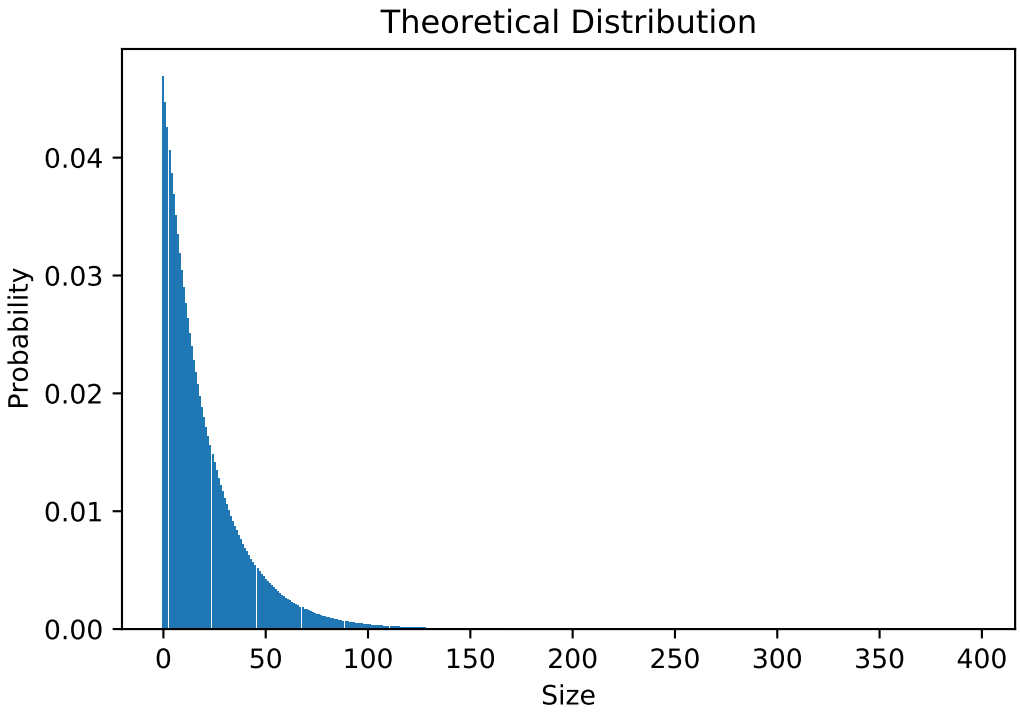
(g) By looking at the scatter plots(look at the attached plots), for each reading frame,

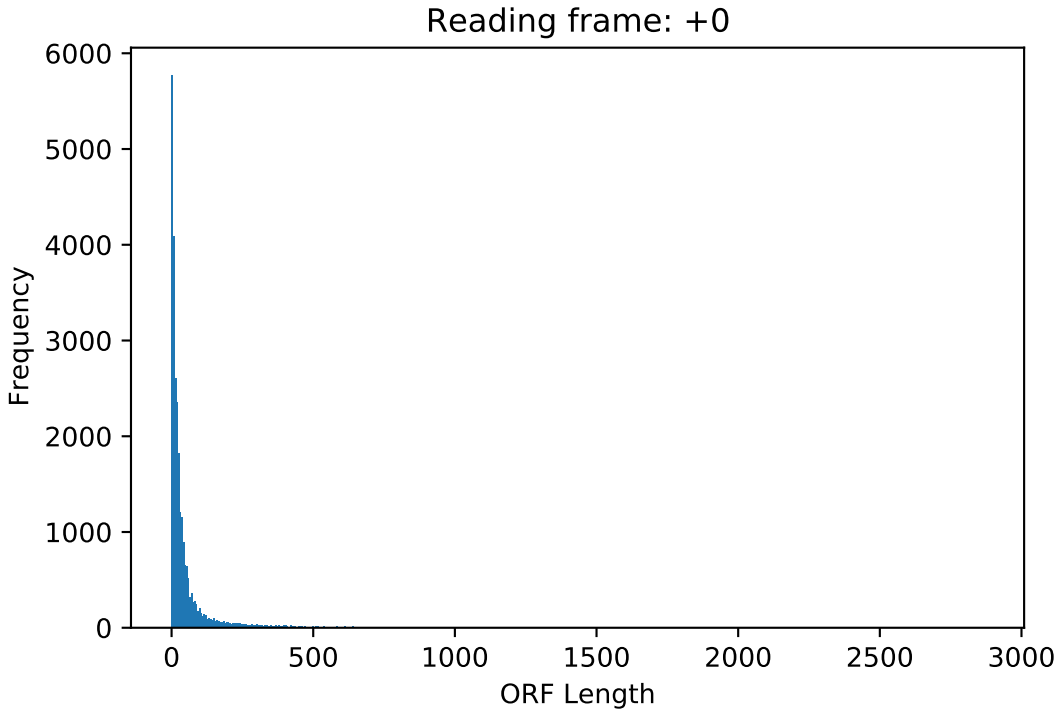
$$L_{cut} \approx 60$$

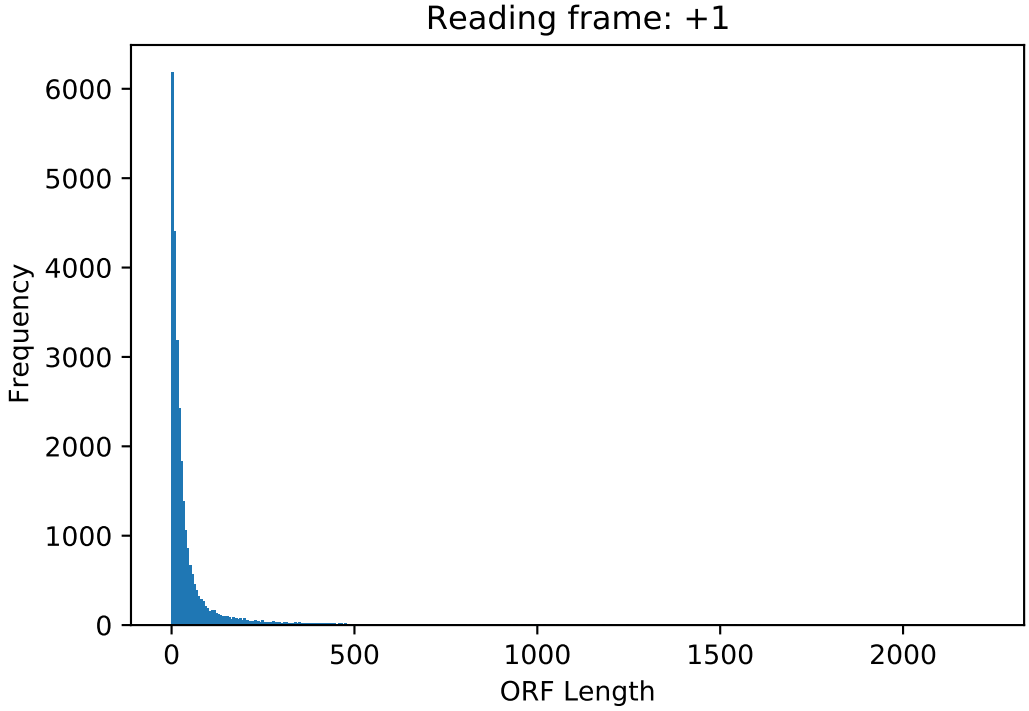
.

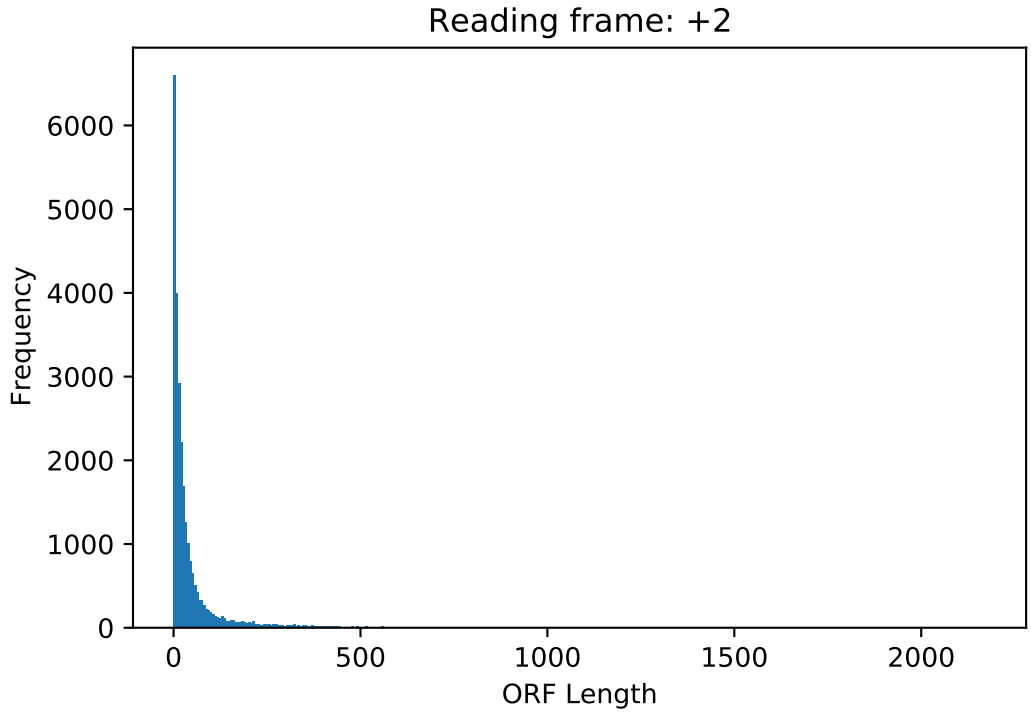
(h) Plot is placed at the end.

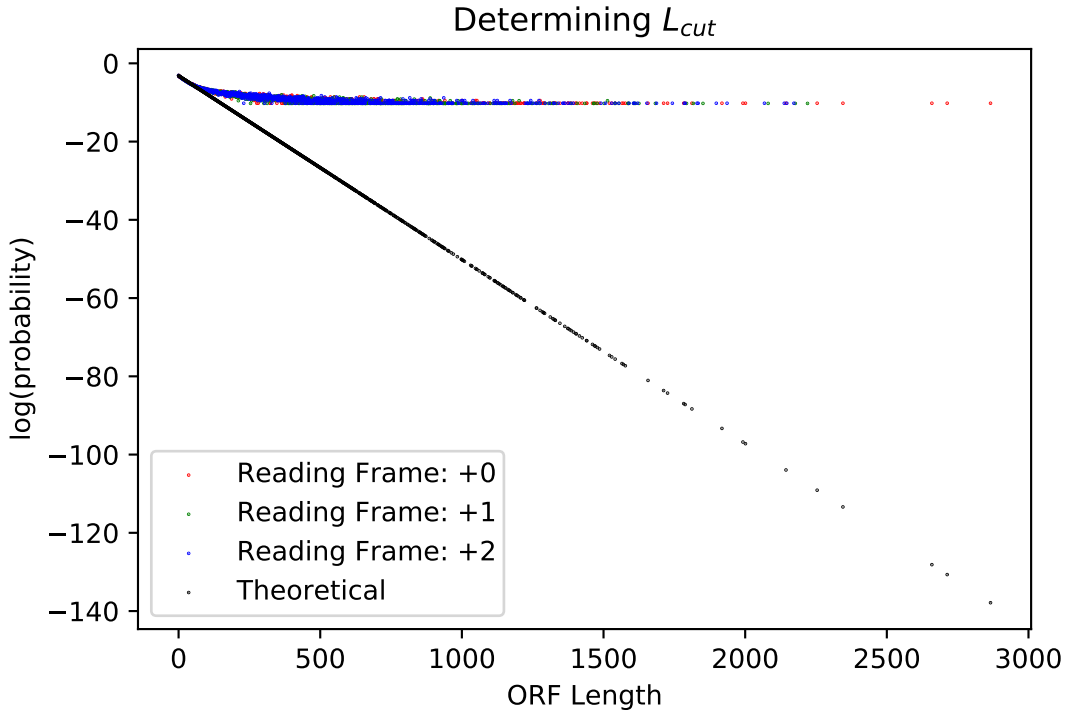

Comparision between fit and theoretical data
Generated
Fit
log(probability)
−2.5
−5.0
−7.5
−10.0
−12.5
−15.0
−17.5
−20.0
−22.5
0
50
100
150
200
250
300
350
400
ORF Length

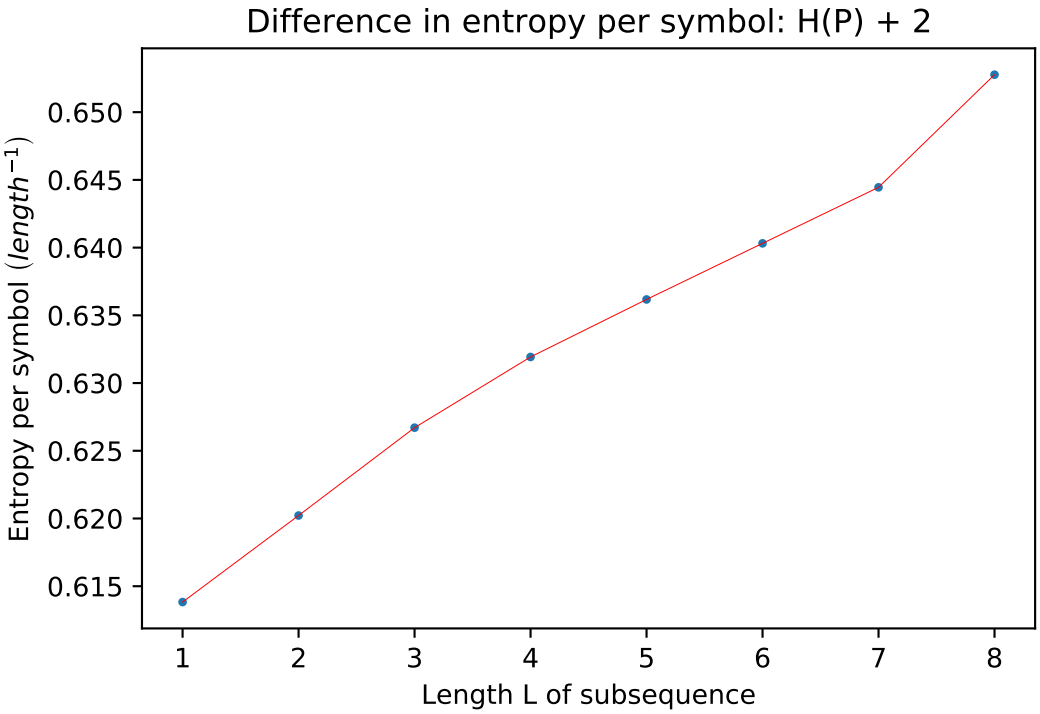

Theoretical Distribution
Probability
0.04
0.03
0.02
0.01
0.00
0
50
100
150
200
250
300
350
400
Size


Reading frame: +0
Frequency
ORF Length
0
1000
2000
3000
4000
5000
6000
0
500
1000
1500
2000
2500
3000


Reading frame: +1
Frequency
ORF Length
0
1000
2000
3000
4000
5000
6000
0
500
1000
1500
2000


Reading frame: +2
Frequency
6000
5000
4000
3000
2000
1000
0
0
500
1000
1500
2000
ORF Length


Determining L_{cut}
log(probability)
0
−20
−40
−60
−80
−100
−120
−140
Reading Frame: +0
Reading Frame: +1
Reading Frame: +2
Theoretical
0
500
1000
1500
2000
2500
3000
ORF Length


Difference in entropy per symbol: H(P) + 2
Entropy per symbol (length^{-1})
0.650
0.645
0.640
0.635
0.630
0.625
0.620
0.615
1
2
3
4
5
6
7
8
Length L of subsequence