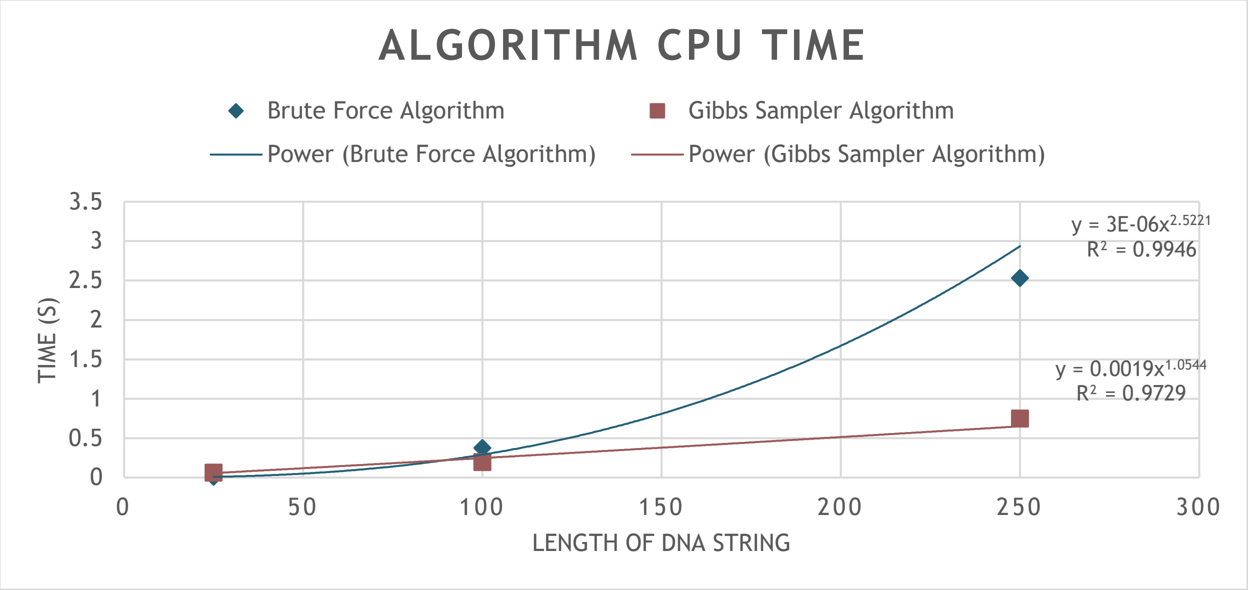
Assignment 2: Motif Finding – Write Up

Question 4 – Part A



Length	Brute Force	Gibbs Sampler
25	0.008 s	0.063 s
100	0.373 s	0.196 s
250	2.531 s	0.749 s

Both algorithms show an increasing graph as the length of the DNA string increases. However, while the Brute Force algorithm is a little quicker to compute in small strings, as the length n of the string increases, the Gibbs Sampler rapidly overtakes the other. Calculating a best line fit for each algorithm shows that both Brute Force and Gibbs Sampler follow a power equation trend. Using the equations from each:

Brute Force: $y = (3 \times 10^{-6})x^{2.5221}$

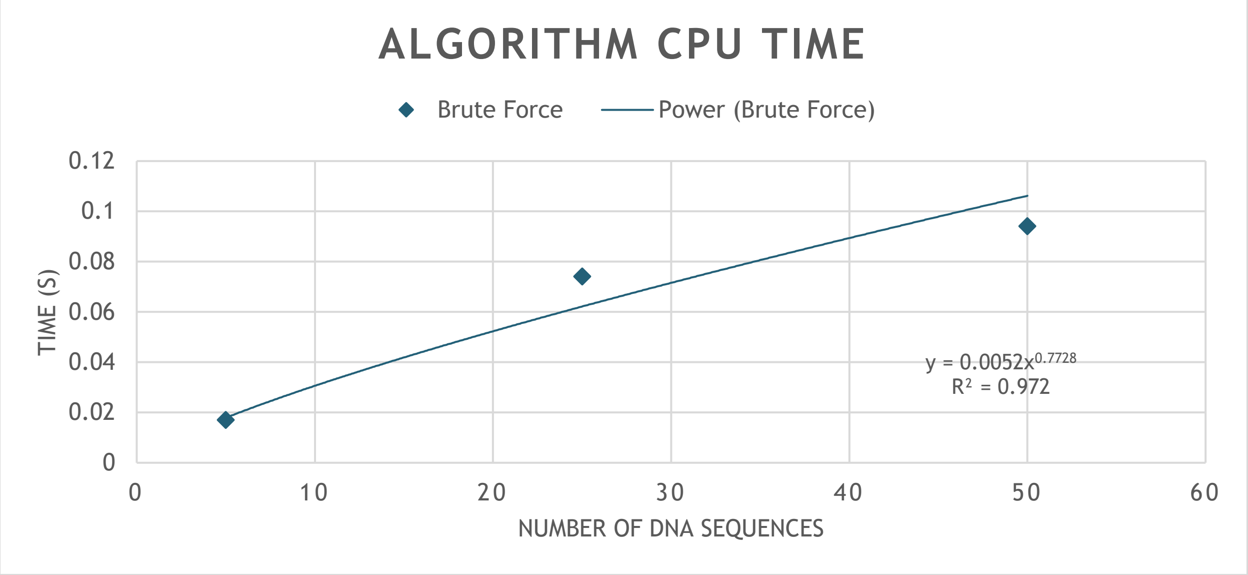
Gibbs Sampler: $y = (0.0019)x^{1.0544}$

Where y is the time in seconds, and x is the string length. It can then be extrapolated that to calculate a string which would take a week of CPU time, for

Brute Force, n ~ 30,351
Gibbs Sampler, n ~ 115,922,428

Which highlights how the Gibbs Sampler is a much faster algorithm

Question 4 – Part B



# of Sequences	Brute Force	Gibbs Sampler
5	0.017 s	N/A
25	0.074 s	N/A
50	0.094 s	N/A

Unfortunately due to some issues with the looping of the Gibbs Sampler and how it interacts with the number of sequences and length of DNA, data on its CPU time could not be collected for how it varies with sequences of DNA. However, based off the previous part it can be assumed that Gibbs Sampler would perform much better over time for longer sequences than the Brute Force Algorithm. A power equation fits the spread fairly well as before, and using its equation:

Brute Force: $y = 0.0052x^{0.7728}$

It can be calculated that a week long CPU calculation would be required when the number of sequences, t, is approximately 27,344,118,151 long.