ECS Assignment 4

Jai Kumar

29th October 2018

1 Introduction

For the problems given in the assignment, I have taken following four DNA sequences:

- NC000911
- NC000917
- NC000964
- NC003098

2 Problem 1

2.1 Statement

Given above four DNA sequences Draw its GC-skew and Correlation(C-G) plot.

Answer

First divide the DNA sequence into 500 parts and starting from every $\frac{N}{500}$th nucleotide take $\frac{N}{100}$ length of DNA sequence and find:

- GC-skew for each stand of DNA, which is given by:

$$S = \frac{N_G - N_C}{N_G - N_C}$$

  where, N_G is number of G in DNA and N_C is number of C in DNA.

- Correlation for each strand, which is given by: a=set of(i) such that i= 1 if 'G' else i= -1 for a given DNA sequence

$$C(k) = \frac{1}{N-k} \sum_{n=1}^{N-k} a_j a_{j-k}$$

$$C_G = \frac{1}{N-1} \sum_{n=1}^{N-1} |C(k)|$$

I have done summation for C_G from 1 to 20 otherwise the program will take a lot of time to run.

2.2 Results

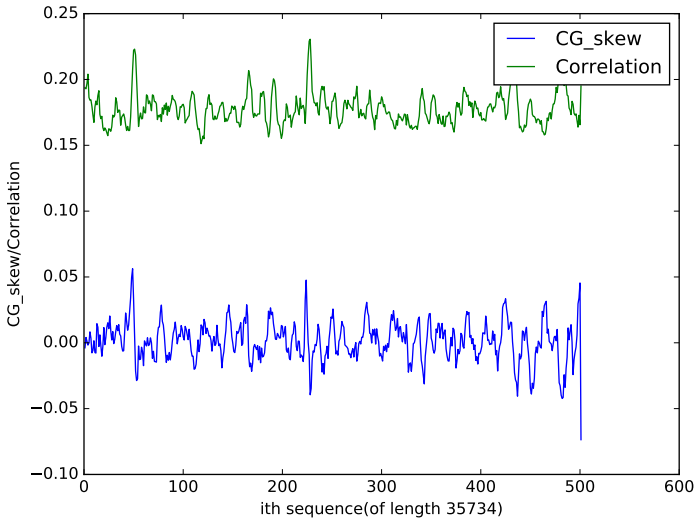

Figure 1: GC-skew and Correlation for NC000911

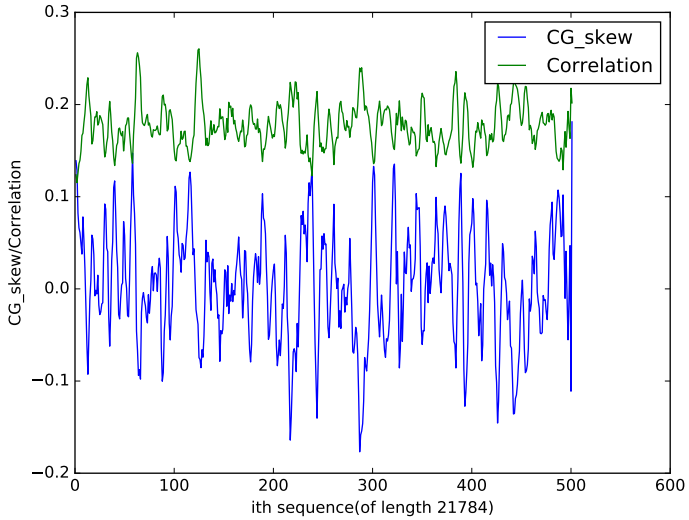

Figure 2: GC-skew and Correlation for NC000917

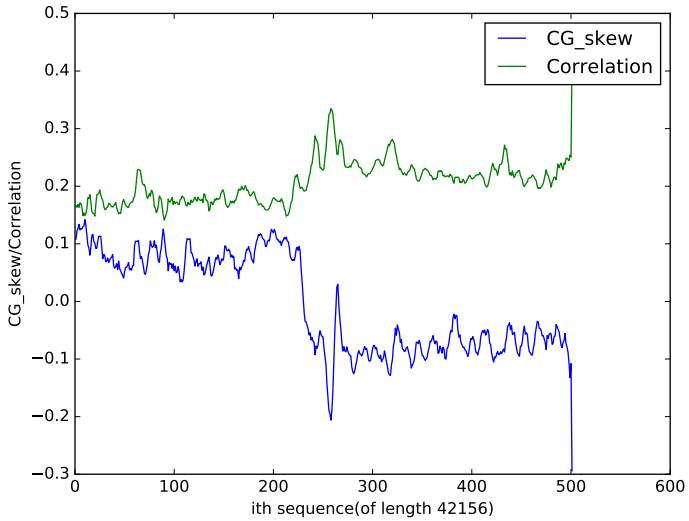

Figure 3: GC-skew and Correlation for NC000964

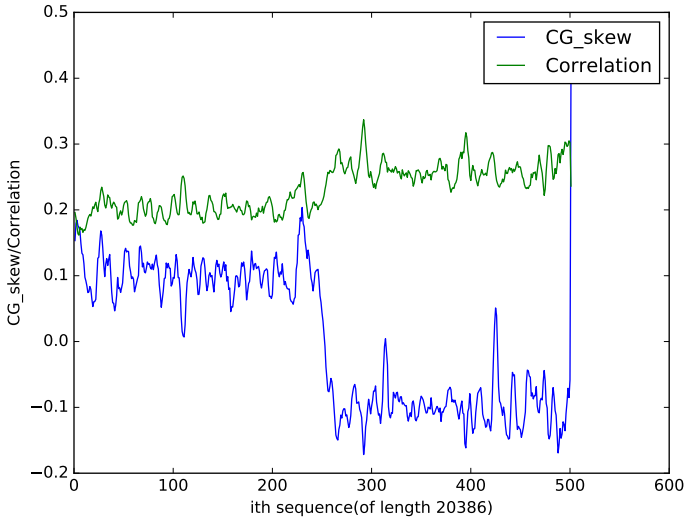

Figure 4: GC-skew and Correlation for NC000964

2.3 Conclusion

As we can see there is a sharp jump at the Origin of replication for NC000964 and NC003098 but nothing for other two which shows limitation of this method for finding origin of replication in different genomes.

3 Problem 2

3.1 Statement

From .ptt file (which has the data for coding and non-coding region of a genome) find P-value for coding and non-coding regions of above four DNA sequence.

Answer

Download .gff3 file from the 'ncbi' site for the 4 given DNA sequences. It has the following format.
genome type start end leading(+) or lagging(-)
'NC000911.1 RefSeq gene 1577 2098 . + . ID=gene'

So, the start and end given in the file are coding region and the one in between are non-coding region(I have taken those non-coding region who has length >= 50 nucleotide).

3.2 Results

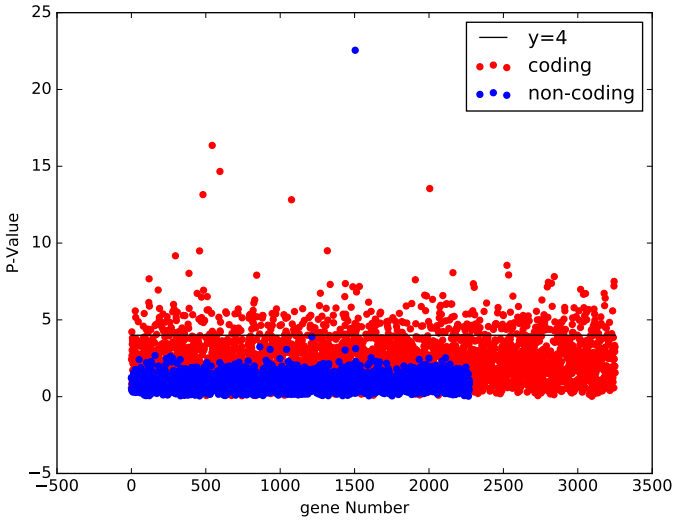

Figure 5: P-value for NC000911

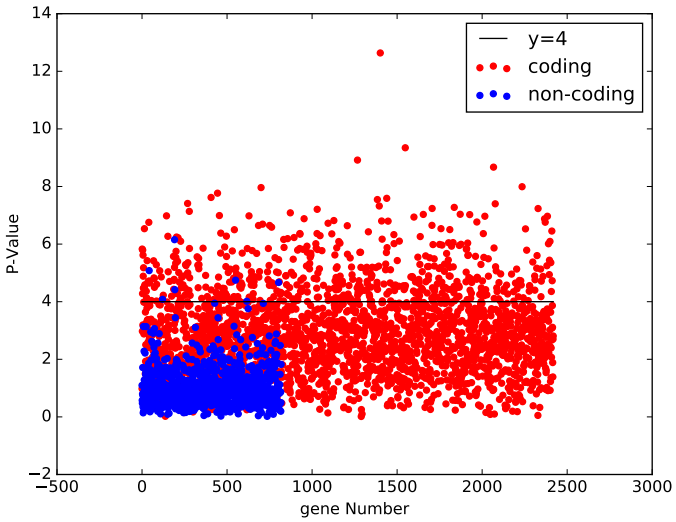

Figure 6: P-value for NC000917

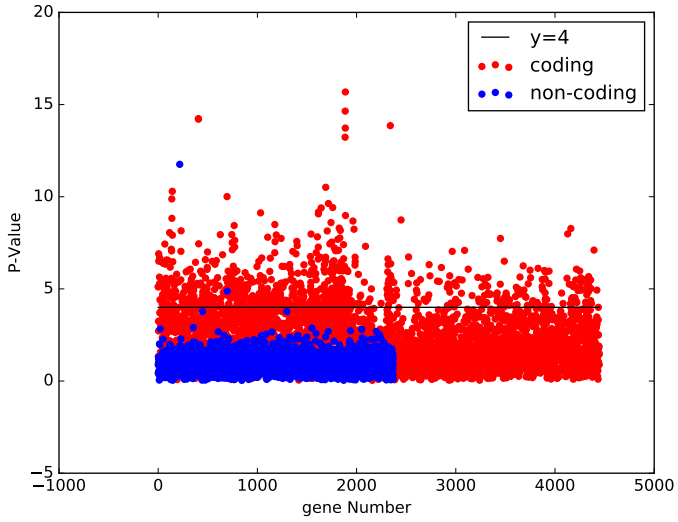

Figure 7: P-value for NC000964

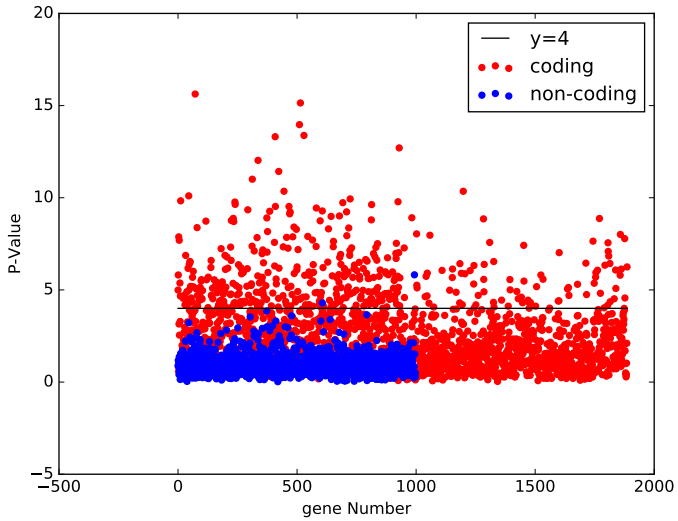

Figure 8: P-value for NC003098

3.3 Conclusion

From the graph it is clear that P-value is < 4 for non-coding regions but for coding regions it can go very high.

4 Problem 3

Draw a graph for the mean P-value of sequences vs. iteration/generation number in the Genetic Algorithm problem.

Answer

Take 1000 DNA sequences randomly defined as [1,0,1,0,0,1,0,0,0,0,0,,..] where 1 represent 'G' and 0 for other nucleaotides such that probability of '1' is 1/3. Mutate, arrange the sequence randomly and choose again 1000 sequence based on your fitness function(defined for P-value ¿4) and keep iterating it until P-value for all DNA sequences is ¿4 or desired no. of iterations (100 in my case). The graph is plotted between average P-value of 1000 sequence vs no. of iterations.

4.1 Results

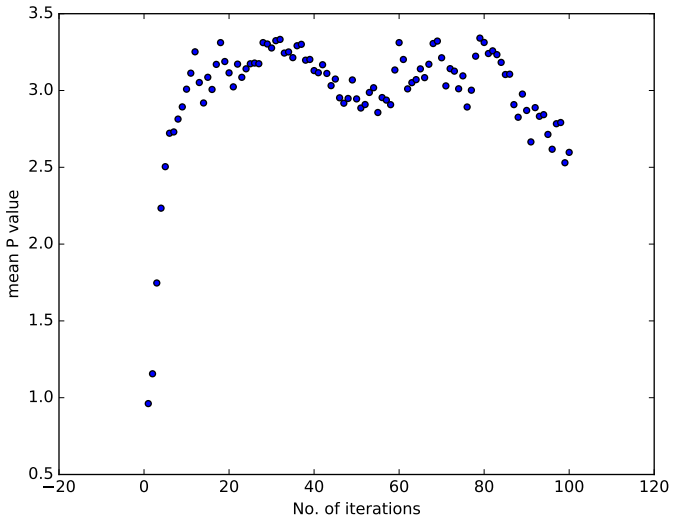

Figure 9: Genetic Algorithm

4.2 Conclusion

From the graph it is clear that P-value is < 4 for non-coding regions but for coding regions it can go very high.

5 Problem 4

Provide convincing justification that the frequency graph of inter-nucleotide distances in a DNA sequence is indeed exponential, through suitable linear least squares fit.

Answer

Calculate the value for the frequency of gap w.r.t any nucleotide (A in my case) take the log of those frequencies and linear fit it after taking a cut-off (50 in my case).

5.1 Results

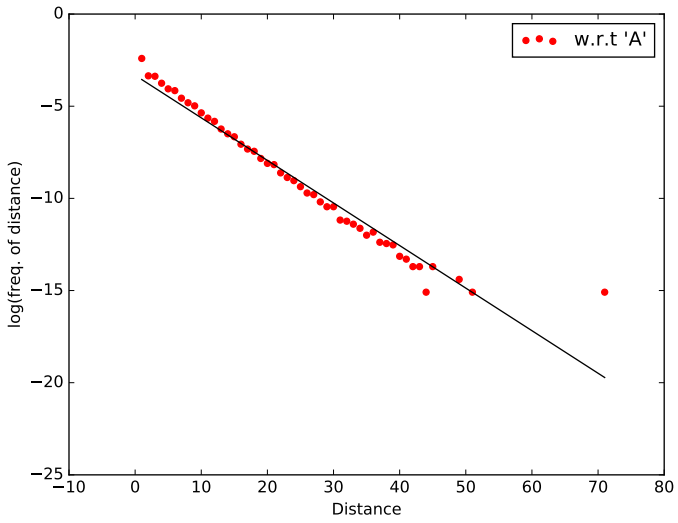

Figure 10: Linear fit for NC000911

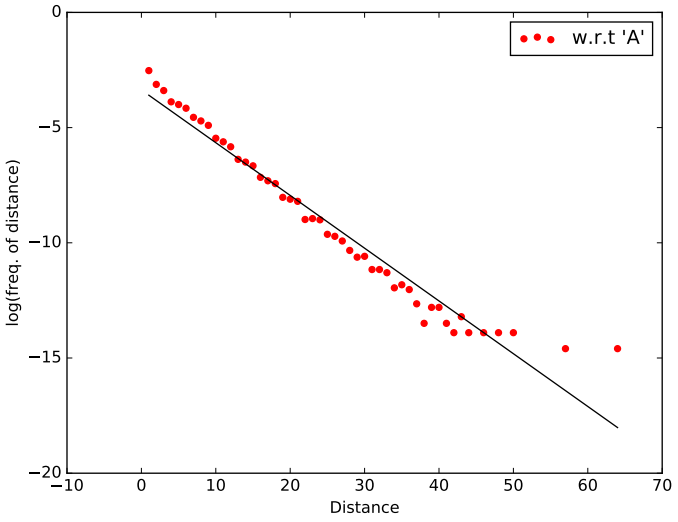

Figure 11: Linear fit for NC000917

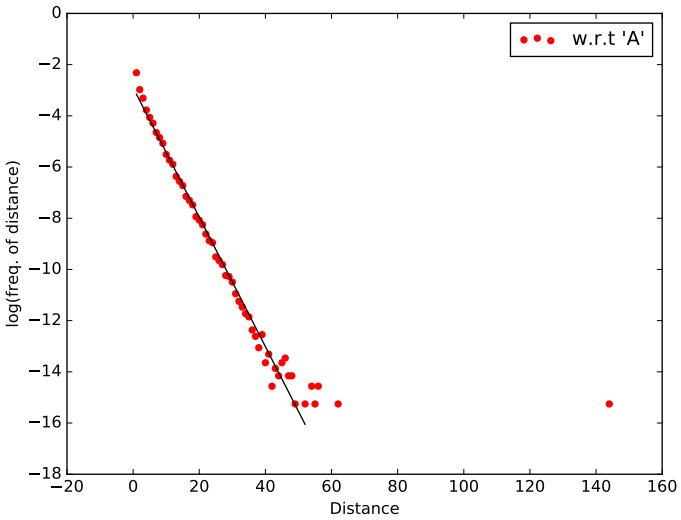

Figure 12: Linear fit for NC000964

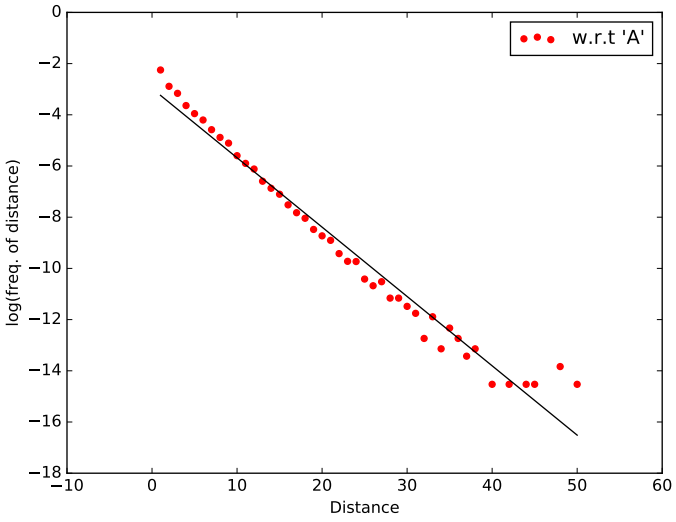

Figure 13: Linear fit for NC003098

5.2 Conclusion

From the graph it can be shown that all the DNA sequence follow exponential distribution.