Data Analytics - Assignment 2

Data Visualization Techniques on Bio-Informatics Data

R Mukesh (CED15I002), Kiran Robert (EVD15I007), Gajaraj G. (MPD15I011)

IIITDM, Kancheepuram

October 10, 2018

Abstract

With the advent of the **Human Genome Project**, there has been an explosion of genomics data in the public domain. Data Visualization have become very crucial in comprehending and assimilating the growing pile of bio-informatics data. The article demonstrates some of the data visualization techniques on genomics data.

1 Introduction

The DNA (Deoxyribonucleic Acid) is a molecule composed of two strands, each built from the four chemical building blocks (Cytosine[**C**], Guanine[**G**], Adenine[**A**] and Thymine[**T**]) called "bases". DNA sequencing is the process of determing the order of these "bases" in a DNA molecule.

The DNA sequence dataset is a set strings composed of the four characters 'A', 'G', 'C' and 'T', each representing the four bases that make up a DNA molecule. The article makes use of one such DNA sequence dataset stored in the FASTA format.

2 Percentages of A, G, C, T in DNA sequences

The percentage of A,G,C,T and other unidentified bases in the DNA sequences are summarized and visually represented using the following charts.

2.1 Pie Chart

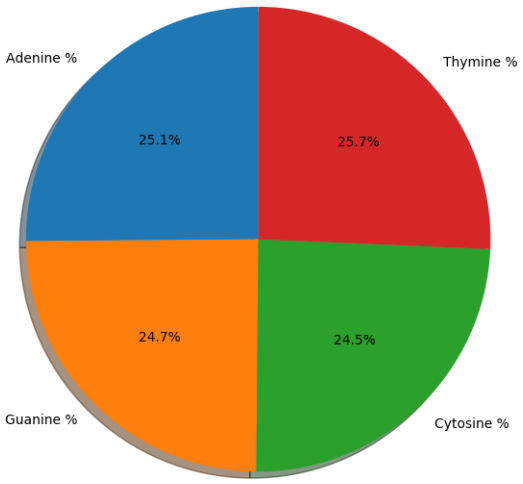

Figure 1: The average percentages of A,G,C,T across all DNA sequences.

Inference: The average percentage of A,G,C,T computed across all DNA sequences are approximately equal (about 25%).

2.2 Box-Whisker Plot

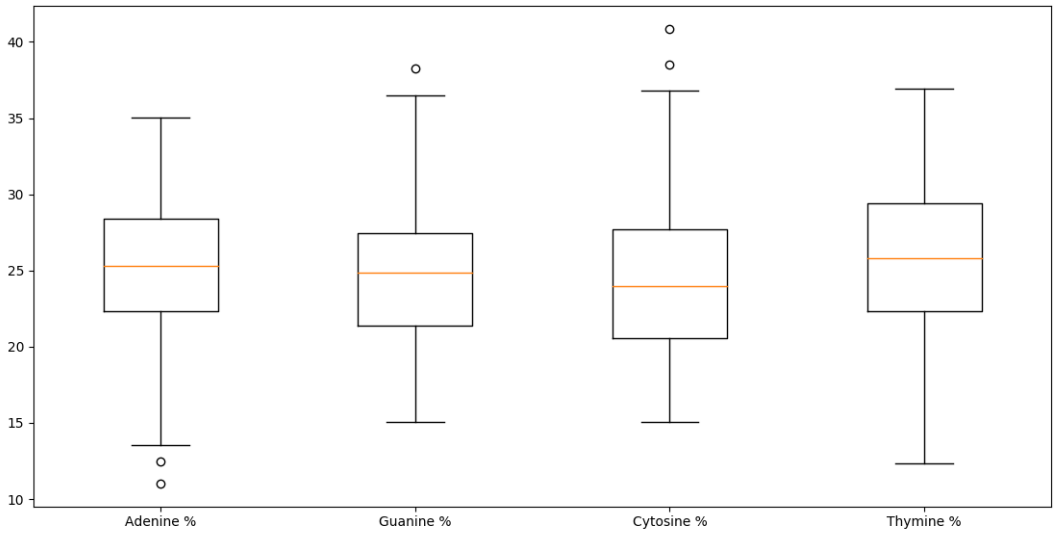

Figure 2: The Box-Whisker plot for percentages of each A,G,C,T for all DNA sequences.

Inferences:

- The percentages of each A,G,C,T have approximately the same distribution with minor variations. The meadian of the percentages of each A,G,C,T are approximately equal (about 25 %).
- A few DNA sequences lie below the lower whisker or above the upper whisker indicating that they vary drastically from their respective distributions (outliers).

2.3 Histogram

The histograms represent the count of DNA sequence with the different percentage ranges for 'A', 'G', 'C' and 'T'.

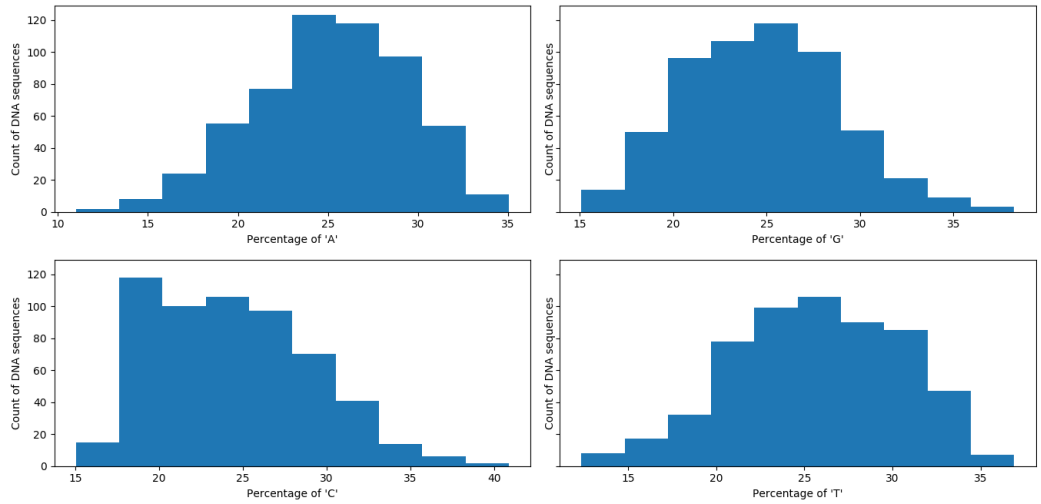

Figure 3: The histograms for the percentages of each A,G,C,T for all DNA sequences.

Inference: Majority of DNA sequences have percentages of 'A','G','C' and 'T' in ranges 20-30%, 20-29%, 18-28% and 20-32% respectively.

2.4 Violin Plot

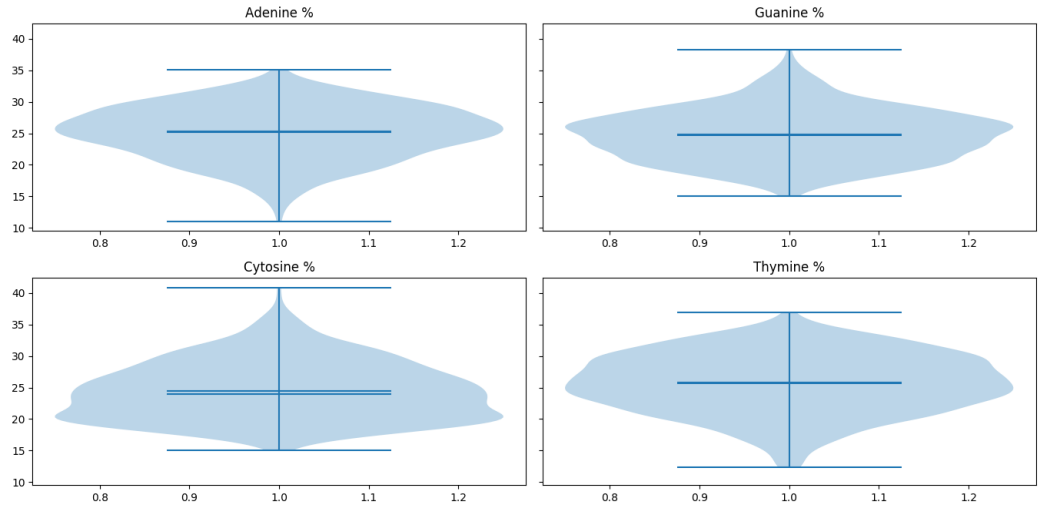

Figure 4: Violin plot for the percentages of each A,G,C,T for all DNA sequences.

3 Number of occurrences of specific sub-sequences

The number of occurrences of the sub-sequences 'AC', 'CAG', 'TTAGGG' in the DNA sequences are summarized and visually represented using the following charts.

3.1 Bar Chart

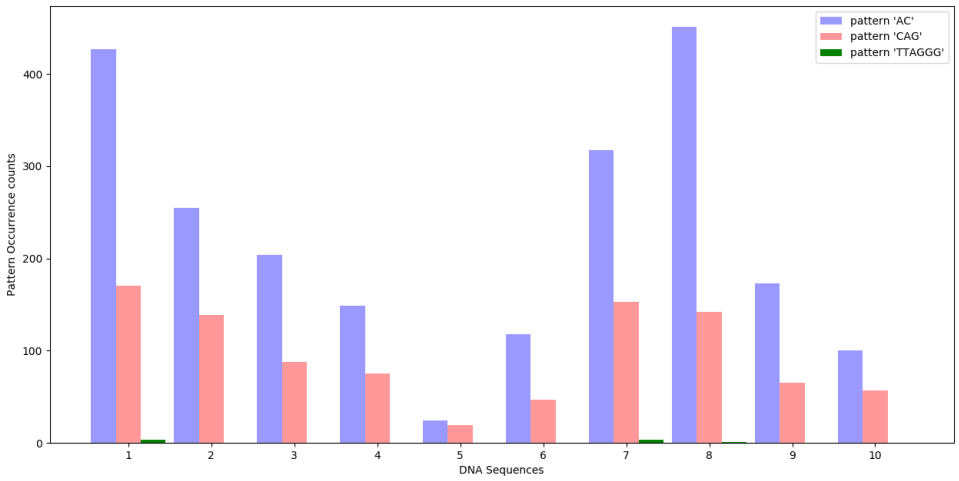

Figure 5: Bar chart for the count of different sub-sequence patterns in 10 randomly selected DNA sequences.

Inference: The number of occurrences of pattern 'AC' is much higher than number of occurrences of pattern 'CAG' for the DNA sequences. The number of occurrences of pattern 'TTAGGG' is negligible.

3.2 Histogram

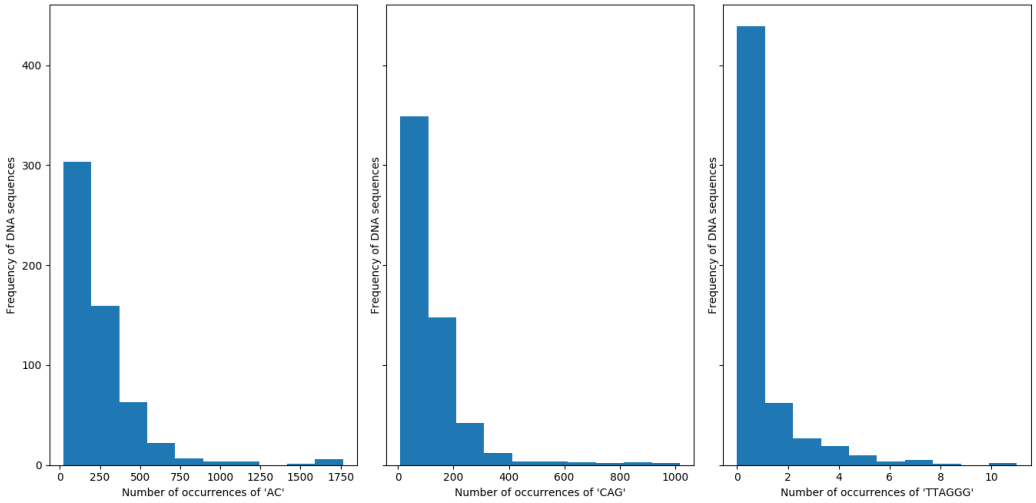

Figure 6: Histograms representing the frequency of DNA sequences for varying number of occurrences of the patterns 'AC', 'CAG' and 'TTAGGG' in the DNA sequence.

Inferences:

- About 300 DNA sequences have the number of occurrence of pattern 'AC' in range 0 to 170. About 350 DNA sequences have number of occurrences of pattern 'CAG' in range 0 to 100. About 440 DNA sequences have number of occurrences of pattern 'TTAGGG' in range 0 to 1.
- Negligible number of DNA sequences have number of occurrences of pattern 'AC' above 700, number of occurrences of pattern 'CAG' above 400 and number of occurrences of pattern 'TTAGG' above 5.

3.3 Box-Whisker Plot

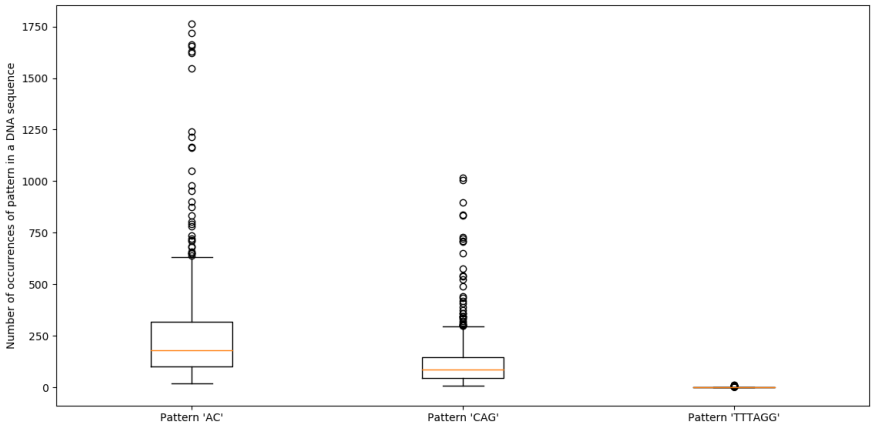

Figure 7: Box-Whisker plot for the number of occurrences of pattern 'AC', 'CAG', 'TTAGGG' for each DNA sequence.

Inferences:

- The distribution of number of occurrence of pattern 'AC' has more variation than that for 'CAG'. The distribution of number of occurrences of pattern 'TTAGGG' has least variation.
- The distribution number of occurrences of pattern 'CAG' has lower median than distribution number of occurrences of pattern 'AC', which is inturn smaller than distribution number of occurrences of pattern 'TTAGGG'.

3.4 Violin Plot

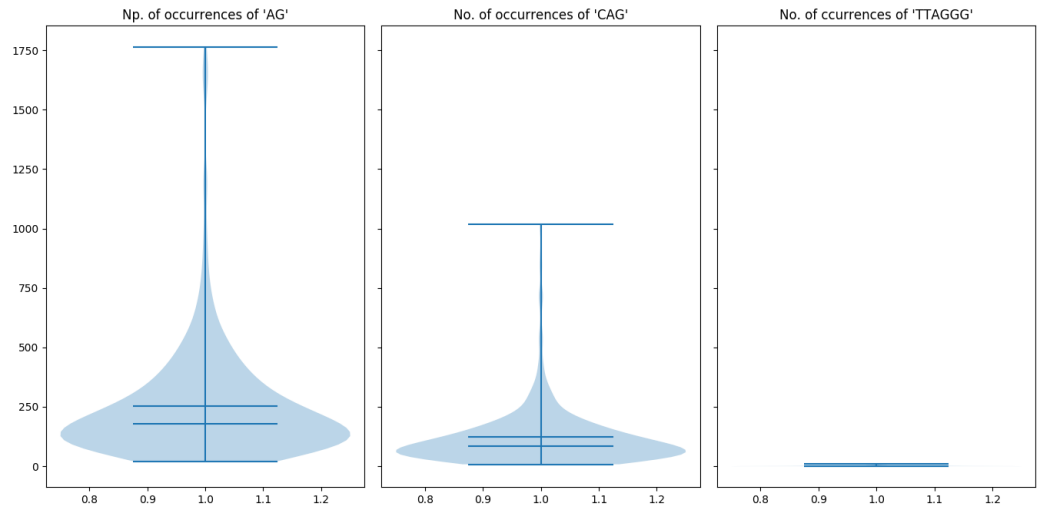

Figure 8: Violin plot for the number of occurrences of pattern 'AC', 'CAG', 'TTAGGG' for each DNA sequence.

3.5 Scatter Plot

The scatter plot seeks to find the correlation(if any) between the values of number of occurrences of pattern 'AC' and 'CAG' for the DNA sequences.

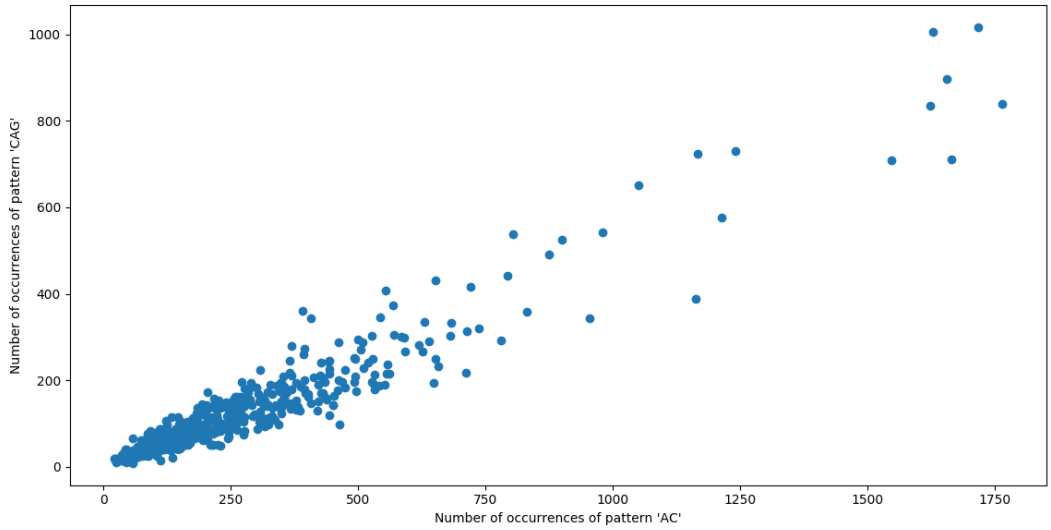

Figure 9: Scatter plot between the number of occurrences of pattern 'AC' and 'CAG', 'TTAGGG' for the DNA sequences.

4 Length of DNA sequence

The length of the DNA sequences are summarized and visually represented using the following charts.

4.1 Box-Whisker Plot

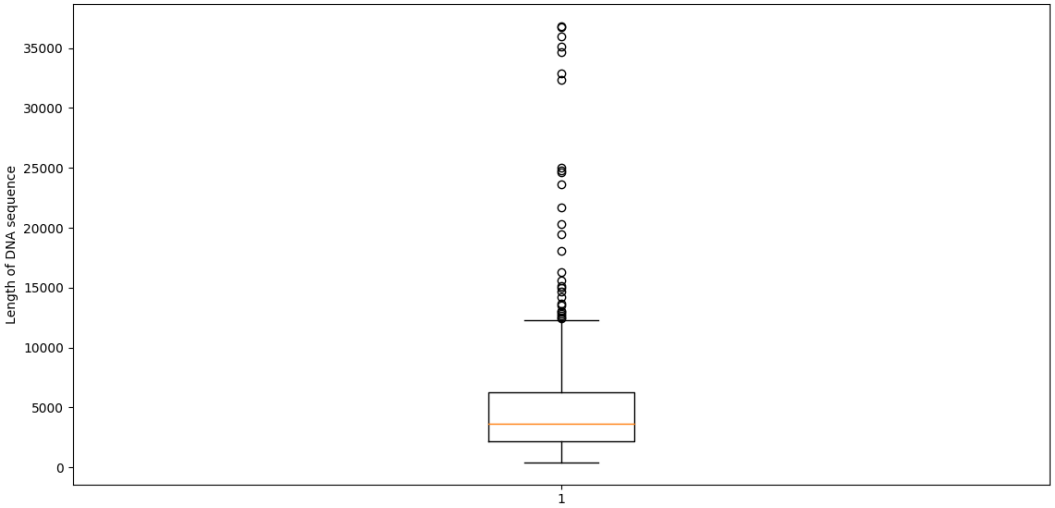

Figure 10: Box-Whisker plot representing the lengths of the DNA sequences.

Inference:

- The distribution of length of all the DNA sequences has median of approximately 2500.
- The DNA sequences that lie above the upper whisker have lengths much higher than the distribution median (outliers).

4.2 Histogram

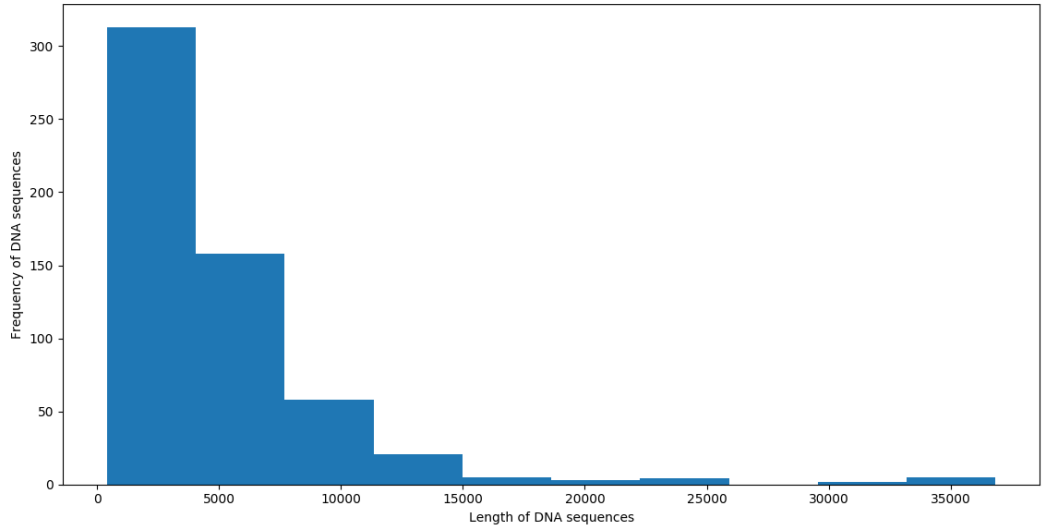

Figure 11: Histogram plot representing the frequency of DNA sequences with various lengths.

4.3 Violin Plot

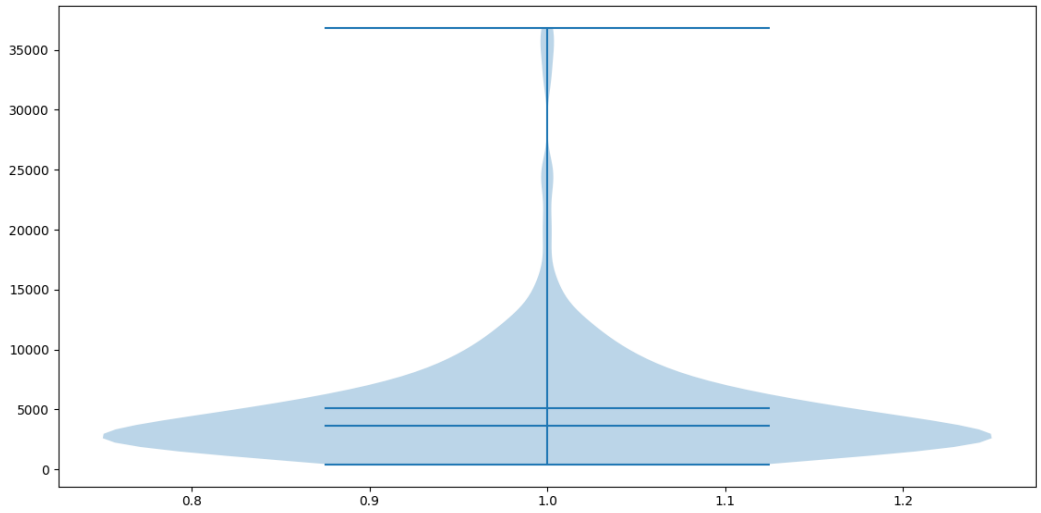

Figure 12: Violin plot for the lengths of the DNA sequences.