METHODS

Genbank files

I was provided the link to the google drive which had the files that I had to analyze. I downloaded 21 genbank files (11 for gram-positive and 10 for gram-negative).

Calculations, Plotting and Packages information

Codes were written to extract information on location (start position, end position) and desirable qualifiers (Gene ID, Gene name, Locus tag, strand orientation) of the feature keys from the feature table of Genbank files for all the genes present in them. All the genes where Gene ID or Gene name were absent were marked as *unavailable* in our data. Gene sequences were also extracted along with upstream- and downstream- flanking regions of length 200 bp and 203 bp (to account for 3 positions, i.e., boundary cases, while looking for motifs in the next step where otherwise it'd give zero count for those positions).

The previously found upstream- and downstream- flanking regions of genes were searched for GAAG and GAAA motifs and the starting positions for these motifs relative to the gene boundary (In 5`-3` direction, gene's first base will have +1 position, so the upstream- flanking region will be from -200 to -1 and gene's last base will have, say *n* is the length of the gene, +*n* position, so the downstream- flanking region will be from *n*+1 to *n*+200) were stored. The counts of these motifs starting at each position in the flanking regions of 200 bp length were recorded.

These counts were then used to create line plots showing frequency of the motifs starting at a position relative to the start site, i.e., +1, of the gene.

For this part, the data analysis and visualizations were done using Python (v 3.8.12) and some of its libraries- pandas (v 1.3.3), BioPython (v 1.79), re module (v 2.2.1), and matplotlib (v 3.4.2), in Jupyter notebooks (v 6.4.3).