Nucleotide Frequencies in the Enterobacteria phage lambda genome

Assignment 1

Caitlin Ross
Computer Science Department, Rensselaer Polytechnic Institute
rossc3@rpi.edu

ABSTRACT

In order to understand the composition of genomes, the frequencies of each nucelotide in windows can be found. The frequencies can then be plotted to show areas that are rich in AT or GC content. This work explores the nucleotide frequences of the Enterobacteria phage lambda genome using three window sizes. The frequencies of the nucleotides are then plotted for each window. The frequencies of GC and AT content are also determined for these windows. The results show that approximately the first 20,000 base pairs are more GC rich, while the rest of the genome tends to be more AT rich.

1. PROBLEM STATEMENT

The problem is to examine nucleotide frequencies in the Enterobacteria phage lambda genome using various window sizes. This can be used to look at the composition of the genome and determine which areas are more AT rich and which are more GC rich.

2. METHODS

First I downloaded the Enterobacteria phage lambda complete genome from BLAST. I selected 'Nucleotide' from the database and searched for 'NC_001416', which is the accession number for this genome. I downloaded the genome in FASTA format and saved it as lambda.fasta.

Then I wrote a python program to examine the genome composition. The program takes in the lambda.fasta file and then calculates the frequencies of each nucleotide (A, C, G, and T) using different window sizes. It also calculates the frequency of GC content and AT content for each window. For each window size, the program outputs a plot that shows the frequency of each window at the starting position. The program contains two functions. One is readfile(), which simply takes the file name, reads in the DNA sequence from file, and returns the DNA sequence as a string.

The other function is calcfreq(), which takes the sequence, window size, and starting index for the window. This function totals the number of each nucleotide and divides each of the totals by the window size. It does the same for AT content and GC content. The function returns the frequencies of each nucleotide, AT content, and GC content.

For each window size, the program collects the frequency data, then plots it. The x-axis is the starting position of each window, while the y-axis is the frequency. The results from the program are shown in section 3.

3. RESULTS

The window sizes used in this work are 500 base pairs, 1000 base pairs, and the length of the sequence/20, which results in a window size of 2425 base pairs. The total length of the genome is 48502 base pairs.

Figures 1-3 show the nucleotide frequences for each window size. Figure 1 shows frequences for a window size of 500 base pairs. The frequency of T makes a sharp increase to about 40% around position 20,000, while G and C both drop down to about 15%. Using larger window sizes of 1000 and 2425 provide much clearer graphs. Using these two graphs make it easier to see how the composition of the genome changes throughout the genome. The first 20,000 base pairs appear to be more GC rich, while the rest appears to be more AT rich. This composition becomes even clearer in Figures 4-6, which plot the frequencies of AT content and GC content. In these graphs it is also easier to see that there are a few points where the frequencies are almost equal, which happen around approximately positions 30,000 and 40,000.

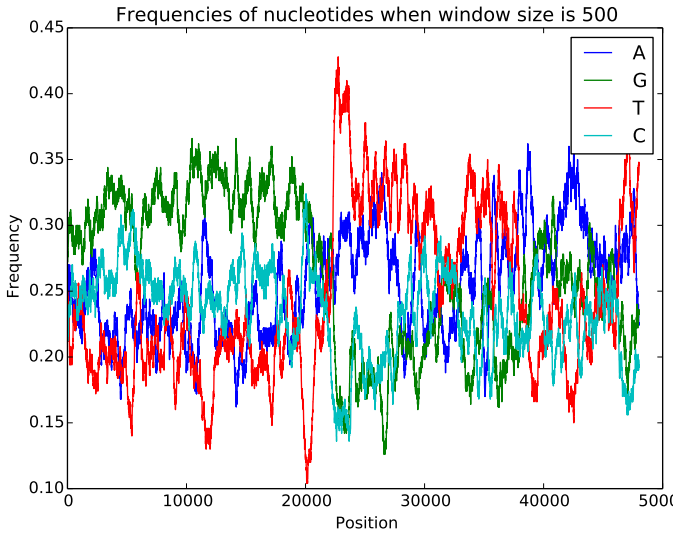

Figure 1: Nucleotide frequences when window size is 500

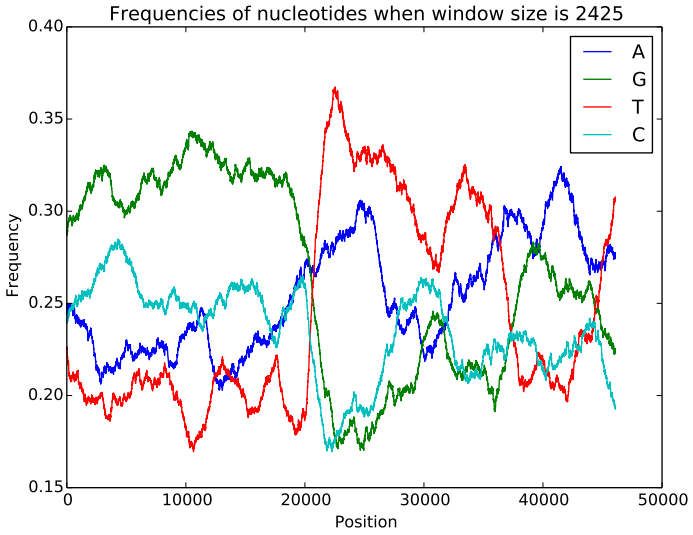

Figure 3: Nucleotide frequences when window size is 2425

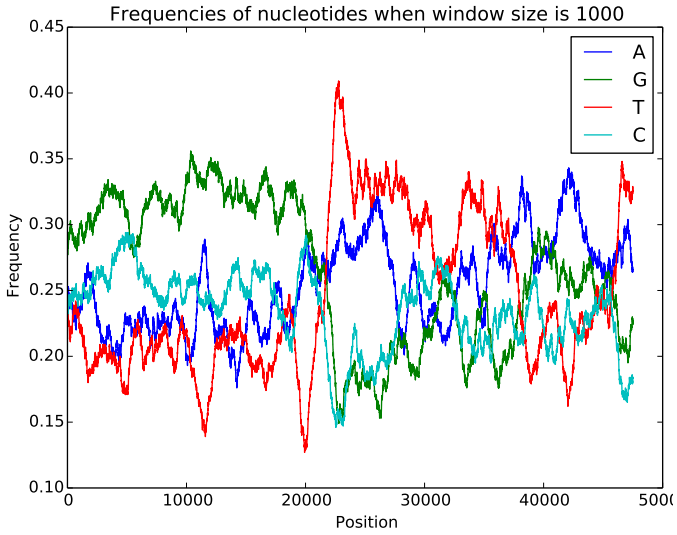

Figure 2: Nucleotide frequences when window size is 1000

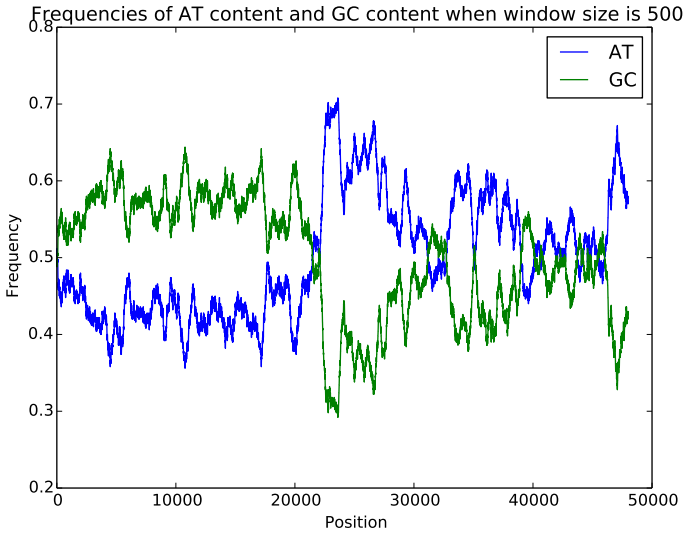

Figure 4: AT and CG content frequences when window size is 500

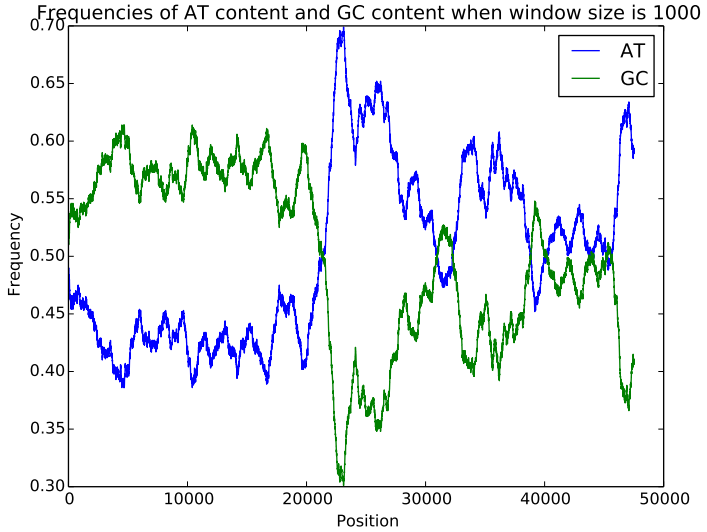

Figure 5: AT and GC content frequences when window size is 1000

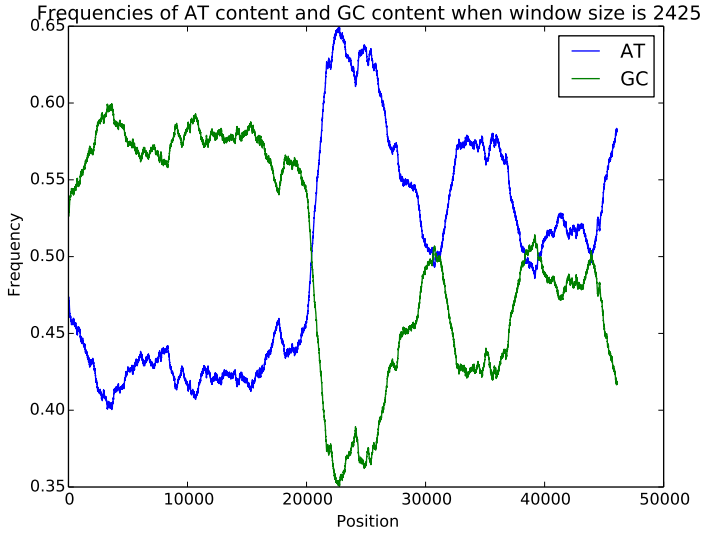

Figure 6: AT and GC content frequences when window size is 2425