1. Give a clear statement of the problem.

The main problem in this homework assignment was to find the frequencies of all nucleotides in the Enterobacteria phage lambda genome at different sliding window sizes. The problem extended to include the frequencies of both A-T and C-G frequencies together in a given sliding window.

2. Give a brief description of the method of your procedure. This should be a few sentences describing the problem, solution, and results. It should be similar to an abstract in a scientific paper.

The process by which these frequencies were calculated were through the input, parsing and manipulation of a data file containing the Enterobacteria phage lambda genome. Using the Python programming language along with various modules (matplotlib, biopython, numpy, scipy), we were able to parse the data into readable sequence data structure. This structure was then split according to window sizes to count the frequencies of each of the nucleotides and nucleotide pairs. My results are displayed below in the graphs. When looking at the A-T and C-G Pairs graphs, we can see a distinct index of 20,000 where the genome switches from C-G rich to A-T rich.

3. Present the procedure in detail, including any relevant mathematics. If the solution involves programming, give an overview in words of how the program operates. The presentation should contain sufficient detail so that a reader can follow each step without difficulty. \*

Pseudo Code:

import all modules needed (numpy, scipy, math, matplotlib/pyplot, and Bio)

parse file into record data structure

set window size variables (500, 1000 and len(record.seq)/20)

calculate length of total sequence

find the number of points that will be plotted for each window size

initialize all lists that will hold the lists of nucleotides and x\_values

for each window size:

initialize list to track freq

for each start\_index in seq:

record beginning of next window in seq

find freq of each nucleotide/pairs in that window

set new start of window as the end + 1

for each window:

plot the nucleotides/pairs frequencies

4. If the solution involves programming code, output must be clearly labeled with each item having a descriptive title. Programming code must be documented with comments describing each major step in the program and all relevant variables. Source code should be submitted in a separate file. \*

\*\*Source code attached\*\*

5. Graphs must be titled and have each axis labeled clearly



