Yana Slavcheva
COS 4091A
Fall 2025
Project: DNA Sequence Visualizer and Analyzer
Prof. Zachary Hutchinson

Progress Report 1

During our meeting this week, prof. Hutchinson suggested I need to consider the file size of a full human genome in FASTA format, so this is what I focused on first. I found that a single human genome contains about 3.2 billion base pairs and each base uses 1 byte of storage, meaning the raw sequence data for one full genome would be approximately 3.2GB in size. I realized that a 3.2GB file is too large for this project, as processing a file that big would be very slow, so I adjusted the project's scope to focus on much smaller sequences like multiple genes or chromosomes (for example, chromosome 21 has around 48 million base pairs – 48MB – which is a lot more feasible).

I also started researching computer science concepts related to my project and keeping track of the papers I read so I could use them as potential sources in the future. I spent time learning about efficient string searching algorithms since a DNA sequence is essentially a very long string. I read more about the Boyer-Moore algorithm and open reading frames, as well as start and stop codons (ATG, TAA, TAG, TGA). The core function of identifying start/stop codons is a pattern-matching problem on a very long string.

Furthermore, I began considering how to best present the results of the analysis. I'm not yet sure what would be the best way to create a clear visual representation of, for example, where the ORFs are located along the sequence. I plan to research more about this and possibly make a simple prototype to better understand how it can be done. Lastly, I drafted an initial UML diagram to map out the key classes in my project - their function, the hierarchy between them, and polymorphism.