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

Progress Report 8

After completing the trie data structure, I implemented the ORF finding algorithm. It scans all six reading frames of a DNA sequence. It treats each reading frame as an independent sequence, as per my advisor's suggestion. Reading frames under 100 bases are discarded. The initial version had incorrect position calculations for reverse strand ORFs which I fixed and now the coordinate mapping to properly show positions relative to the original sequence. I also improved the algorithm to check that ORFs don't contain any internal stop codons, which would make them biologically invalid. Finally, I added filtering functions:

- By minimum length
- By specific reading frame
- For removing overlapping ORFs (keeping the longest one)

These also make the display of the found ORFs clearer.

Currently, I'm almost finished with the implementation of the last algorithm - the comparison one. It can find the best alignment between two sequences by building a scoring matrix and reconstructing the actual alignment from the matrix. My next objective is to test this algorithm in the main and make it show some statistics about the two sequences, like a percentage of similar regions over 50 bases.

https://github.com/yYana33/senior-project