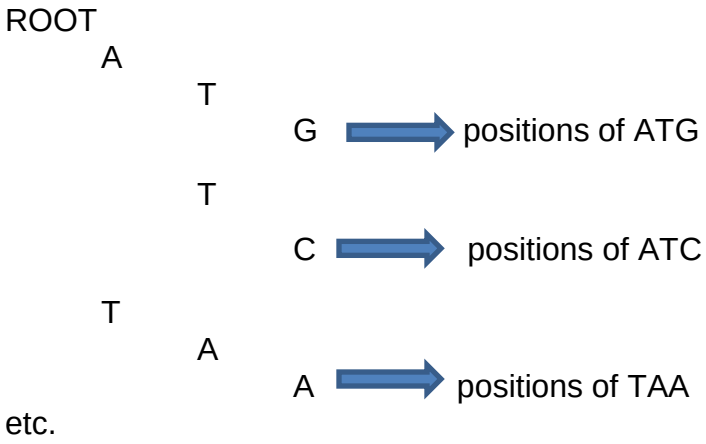
Yana Slavcheva
COS 4091A
Fall 2025
Project: DNA Sequence Visualizer and Analyzer
Prof. Zachary Hutchinson

Progress Report 7

This week I worked on implementing the Trie and ORF finding. They're connected but still separate structures. The main idea is that the Trie will find the start and stop codons quickly and will keep track of their positions in the DNA. In this way, the actual ORF finding algorithm will be able to look up all the start and stop codons instantly and use the results to identify the genes instead of scanning the entire sequence multiple times.

The Trie won't hold the entire DNA sequence, only the important codons. This means each path from the root will represents one codon. Each node will store one letter (A, T, C, or G) and then each leaf node will store a list with positions where that codon occurs in the DNA string. To visualize:



If the codon ATC is found at positions 15 and 37, the node should contain a vector of positions = {15, 37}

I started working on the Trie and I had several issues, mainly with managing the different frames and overlaps. Each reading frame shifts the starting index by one and at first I was looping incorrectly and was either missing codons or counting some twice. I also noticed that the Trie could miss overlapping codons or stop early. Lastly, handling

the reverse strand properly took me some time. At first, the coordinates were not mapped correctly, so the ORF positions on the reverse strand didn't match their real locations in the original sequence. I found that the correct solution is to keep a single coordinate system for both strands and just record which strand (+ or –) each ORF belongs to.

My advisor suggested reading each ORF as an independent string which was a valid solution and helped avoid the coordinate and frame-handling issues.

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