The motivation for my project was to try to better understand the Evening Element, a regulatory element that is influenced by the presence and absence of CCA1. I wanted to learn more about where in the promoter region this motif showed up, and how consistent this location was over multiple different genes with the same regulatory element. To gather my data, I went to the Eukaryotic Promoter Database and got 1000 base pairs of the promoter regions from each gene and copy and pasted them each into a .txt file in my repl. My program starts by constructing a list that contains all of the promoter sequences as strings. Then it starts to build the graph by graphing ten (number of genes) straight lines on a graph to represent each gene. Then the program goes through three different hamming distances (2, 1, and 0) and does the following process for each one. First, it finds all of the kmers that have exactly the given hamming distance and puts them in a dictionary, keeping track of the gene's name and the position of the kmer in its promoter region. Then, it assigns x and y coordinates using the gene name and the position of each kmer to graph all of the kmers with the same hamming distance from the motif.

What I found was that for a few of the genes (about four out of the 10), the position of the motif in the promoter region was conserved (it tended to be about -900 bps away from the start of transcription), but for other genes it was harder to tell. There were some genes that had a cluster of the less high affinity matches in the 900 bp region, but again it was difficult to evaluate these from the graph alone. I did some research about conservation of motif locations, and found that the conservation of a motif's location can be an indication of its biological significance. I think would have to look at a lot more genes than just these ten in order to make a claim about the biological significance of this Evening Element.