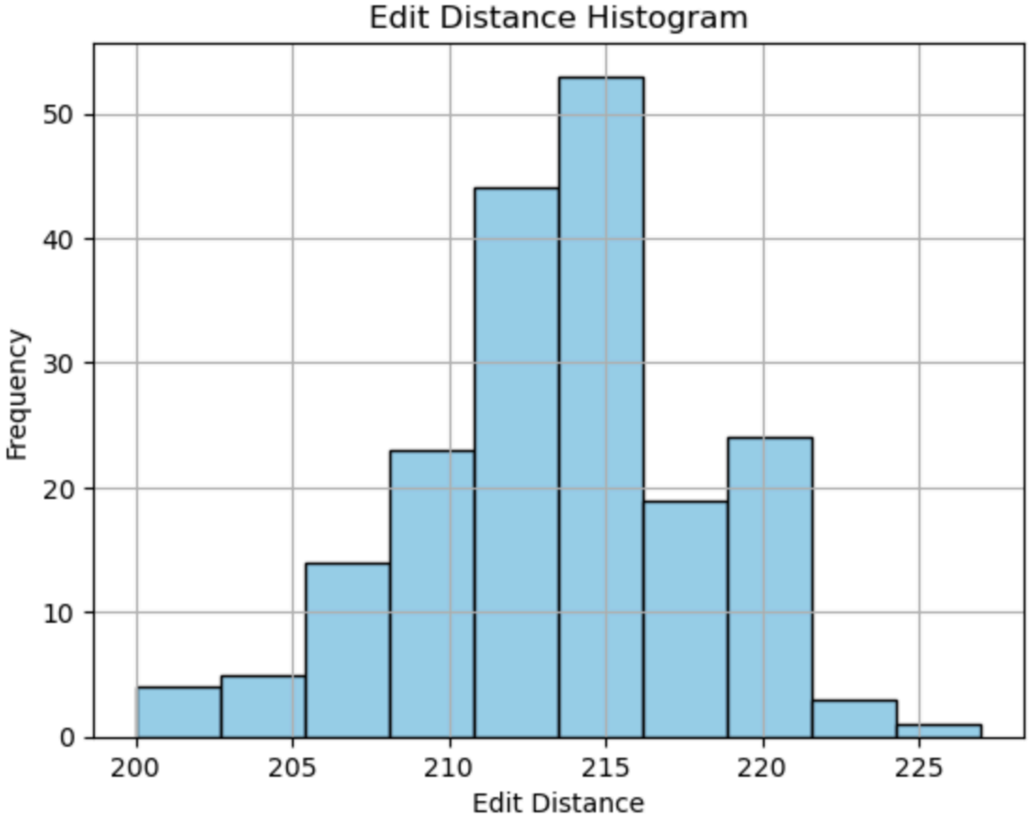
Libby Amir
920998496
ECS117
17 May 2024

Edit Distance in DNA Sequencing

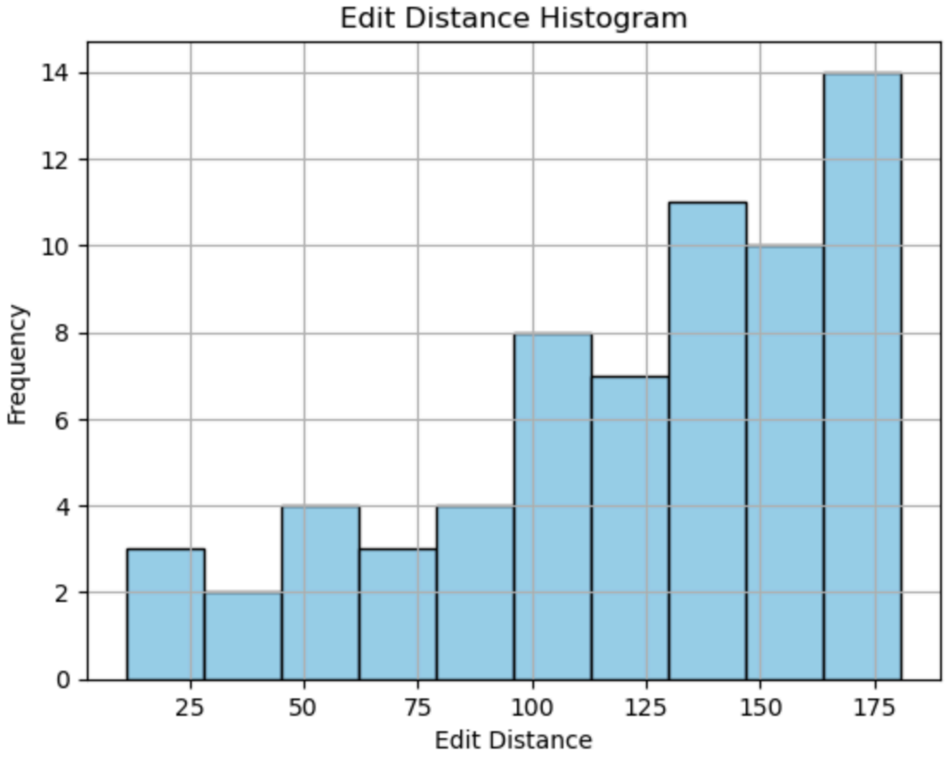
By generating random DNA sequences of substantial lengths, we can begin to analyze the nature of the edit distance program. After generating 20 random sequences each with a length of 400 nucleotides, we can test our dynamic edit distance program to see how much they differ from each other. Since the sequences are entirely random, we expect there to be an approximately normal distribution of distances. This is easy to interpret because there are not supposed to be any distinguishable patterns that link any two randomly generated sequences, so there should be some sequences that are coincidentally more similar and others that are more different. This also indicates that the mean, median, and mode are all relatively close to each other. If we were to have a larger sample size of randomly generated sequences, greater than just 20, we would expect the distribution to be even more normalized. As seen below, the bell curve of the histogram demonstrates the normal distribution, and the average is around the center of the graph. The average distance is quite high at 213.45, but this makes sense because there are no associations between random sequences.

```
Average Edit Distance: 213.45
```



In comparison, when we perform the same analysis on real DNA sequences we get very different results. In this case, we are specifically instructed to use data from a select set of species, and this implies a connection between some of the sequences. We can assume the biopython files contain several sequences for each species, and those sequences likely have commonalities because animals within the same species may share DNA subsequences. Thus, we would expect a lower average edit distance between DNA sequences within specific species due to their commonalities. This is confirmed in our analysis, with an average edit distance of 122.53 in the real sequences relative to the randomized average of 213.45. Another reason the mean distance is intuitively lower is that most of the species seem to be in the monkey family which would explain them having more shared DNA between them. Another core difference we witness in data is that the histogram of edit distances is very skewed and far from the normal distribution seen before. The edit distances calculated from the real sequences are left skewed, meaning there is a higher frequency of greater edit distances than lower ones. When analyzed, one can deduce this is likely a result of there being a couple DNA sequences from each of the species that are more similar to each other, but significantly more different from all the other species. For example, if there are 50 sequences and 5 of each species, then any given sequence is likely to have a low edit distance with the other 4 of the same species, and a higher edit distance with the other 45 sequences. The "other" greatly outnumbers the "alike," hence we can see below that there is a skew in favor of greater edit distances.

```
Average Edit Distance: 122.53
```


Edit Distance Histogram
Frequency
Edit Distance