Iain Hennington
4/9/2024
ECS 117
Frid

Programming Assignment: Analysis of DNA Sequences and Edit Distance

Observations: After making it, I noticed that there was an almost normal distribution for the randomly generated DNA sequences (my data). There were outliers at 0 (shown on histogram) which was when the edit distance of DNA sequence was being compared to itself, thus making the edit distance 0. The mean for my data was around 225-240 edits made. For the real data, the data was left skewed on the histogram, but had outliers at 0 as well, for when the edit distance of the DNA sequence was being compared to itself. The mean for the real data I estimated to be about 100-125 edits made.

Insights/Conclusions: Based on the data collected, it does make sense that the randomly generated (my data) DNA sequences were gonna have an edit distance that was very high and very close together. Because it was completely random, the chance of the edit distance being very low would be difficult because the DNA sequences would have to be very similar (and randomly generating and getting that is really difficult). The real data is on biological data of gorillas, orangutans, humans, etc, so the edit distance will vary a lot between all of the real data. This is because orangutans are gonna be very different from us DNA wise as compared to a neanderthal. So that is why the variation in the edit distance on the histogram is very dispersed, the DNA sequences are gonna vary a lot compared to each other. From this, we can conclude that there are other biological beings that are very similar to us DNA-wise, as the chance of having similar DNA sequences is very rare (shown in the random data).