For my final project, I build a relatively simple website/script that allows one to submit entries for analysis on traits of the sequence, and then search for similar entries. In this write up, I'll cover why I choose this as a project, how it could be used, future uses, issues I had with implementation, and finally the results.

This idea came to mind because of the high usage of databases that come up in the bioinformatics courses. Most of the time, we're looking through NCBI's huge amount of data and entries, and thought it would be nice to find similar entries based on sequence content. The semester before this, I was working a lot with BLAST to find novel proteins, and thought that this could be something analogous.

The main usage as it is now is to find entries with similar GC content, Purine content, or Pyrimidine content. GC content in particular is useful for finding a signature profile of a sequence or organism, which will have different GC percentages, and may help with identifying origins, or may help in identifying potential errors in sequencing. Although calculating this content isn't too difficult, the tool here provides an advantage of grouping them together for easy identification.

Unfortunately my programming skills are pretty basic still, but I could imagine some simple extension but maybe difficult implementation for this tool. For example, I think this tool could be really useful if you could search by motifs or group by motifs, or other traits of sequencings, such as high consecutive bases or statistically strange patterns. In addition to this, having location information would be useful as well, by finding similar structures in the same region, you might better identify paralogs and orthologs. Some minor quality of life improvement I noticed were things like names of the accession themselves, as searching by accession number by itself is confusing and unintuitive.

In general the main issue I had building this project was the troubleshooting process. I could test the scripts and sql queries easily by just running them directly from command line, but I had major issues figuring out the web connections. For example, at one point I wasn't pulling in any data from my form, and I struggled to find out what was causing the issue. I resolved it by going through the statements in my code and trying to traceback what was happening as I read through it line by line and found an incorrect line that was pulling data from the wrong form. In general although this project was conceptually easier, I found myself having a harder time putting everything together. The individual parts were simple, but the interactions was where I was having difficulty. What helped was physically drawing a map of how each part worked together so that I could visually see how each file acted together.

As expected similar genes provided similar results when searching by profile. For example AB086379.1 and AB085628.1, both isoforms of hTERT for telomerase show extremely similar profiles. What was surprising was the similarity to AVP or ADH. Another surprise was how varied the populations could be. For example, I put in an initial dataset of about 10 homo sapien genes, but typically only 3 genes show up when searching for GC content values, suggesting that the GC content is much more varied than I initially thought, and falls outside a +/- of 5.

Overall I thought the project was a good summary of the skills I learned throughout this course, and could easily be a project for my own self improvement by adding new features, like mentioned earlier. I've been studying some D3.js on my own recently, and I'd like to be able to add in barcharts and other visual content that is reflective of the results, and I'm hoping that I'll be able to do more as I grow.