Logan Brown
PI: Krisztina Varga
09/22/2024
NASA EPSCoR: NASA Space Grant Graduate Fellowships at UNH
Summer Research Report: Antifreeze Protein Database and Soil Sample Analysis

During the summer, my research focused on establishing a comprehensive database of known antifreeze proteins, with the goal of identifying potential antifreeze proteins from soil samples collected from cold-weather climates. This effort aims to advance our understanding of natural antifreeze proteins, which have significant potential in biotechnology due to their ability to inhibit ice crystal formation and preserve cold-labile cells and tissues during stresses from cryopreservation.

The project's initial objective was to create a robust database of antifreeze proteins, against which genome sequences from environmental samples could be screened. Hits from these screenings would then be used to clone and express candidate antifreeze proteins for further characterization through assays measuring antifreeze capabilities, such as thermal hysteresis, ice recrystallization inhibition, and ice crystal formation studies.

Over the summer, I took a week-long genetic programming workshop, which provided a foundation in computational biology techniques. This training proved essential, as I am relatively new to genomics and bioinformatics. Together with the Hubbard Center for Genome Studies, we created an on-site antifreeze protein database by conducting searches on publicly available databases, such as UniProt and NCBI, and gathering all the relevant proteins into one location on a UNH server. I generated a database of ~23000 proteins form NCBI and ~5300 proteins from UniProt. After filtering proteins out for duplications and hypothesized proteins, these numbers were cut to ~6800 and ~4900, respectively. This was done to make the database robust but also cut down on matches generating multiple hits on the same protein.

While significant progress was made in database development, my work on analyzing genomic data from cold-weather soil samples is ongoing. BLASTp is being used to compare the metagenomic data of around 50 unique sequences against the database. I will need to wait to recombinantly express any proteins until after all the data has been analyzed and a small list of target proteins are determined. This step, along with the experimental characterization of antifreeze activity, will form the next phase of the project. Delays in progress were largely due to the learning curve associated with acquiring new skills in genomics and bioinformatics.

As the data from samples are compared to the on-site database, further curation of the database will be necessary, for example we are expanding the database to include ice nucleation proteins. One aspect currently missing is the structural data for known antifreeze proteins. While many entries include accession numbers for protein databank structures, these files are larger and more complex than the sequences that currently populate the database. Ensuring that the database is well-curated and robust is crucial, as it must reliably guide the selection of relevant sequences. This is important because expressing and purifying recombinant proteins is resource intensive. Once the database is sufficiently curated, we can be more confident in selecting proteins that are likely to exhibit antifreeze capabilities.

Despite the slower-than-anticipated pace, the work completed this summer lays a strong foundation for the project's continuation. Furthermore, I am continuing to work with the Hubbard Center for Genome Studies to continue this effort during the school year. The database is a critical resource that will facilitate the identification of novel antifreeze proteins, and ongoing genomic analysis will likely yield promising candidates for future experimental validation for many years extending much further than my personal involvement.