**Student**: Trinity Cheng, trinity.cheng@fda.hhs.gov (CDRH/OSEL/DIDSR)

**Mentor**: Dr. Mike Mikailov, mike.mikailov@fda.hhs.gov (CDRH/OSEL/DIDSR)

**Poster Title**: Optimal Data Parallelization of the BLAST Bioinformatics Application on HPC clusters

**Abstract**:

The Basic Local Alignment Search Tool (BLAST) is an application widely used in the bioinformatics field for matching a list of unknown sequences--the query--with known, archived sequences—the database. In 2017, Mikailov et al. developed the “dual segmentation” method to decrease the runtime of BLAST by dividing the database and query into m and n fragments, respectively, that were then processed in parallel using p=m\*n nodes. In this work, we investigate the optimal partitioning (m,n) of any given database and query to result in the greatest speedup. Our approach estimates the overall runtime by measuring the runtime of a sub-job running on only one node, which performs 1/(m\*n) of the total job. The sub-job with the fastest runtime determines the optimal solution (m,n), which is then used to run the full dataset in parallel on p nodes.

Using a 2.5 GB database and a 15 GB query on the CDRH HPC Betsy cluster, the optimal solution was found to be (m,n)=(2,2048). The sub-job running on a single node ran 3.67x faster than the previous 2017 configuration of (m,n)=(27,152). To verify our findings, both solutions were tested with the full database and query by running BLAST in parallel. The (m,n)=(2,2048) configuration ran 4.56x faster than the original (m,n)=(27,152) configuration.

In this study, we learned that BLAST users may benefit from dividing the query rather than dividing the database. Although the effectiveness of this partitioning may vary based on different database and query sizes, we propose the method of estimating the overall runtime by measuring the runtime of a sub-job.

**Category**: Unleashing the Power of Data

**Poster Synopsis:**

The widely used bioinformatics application Basic Local Alignment Search Tool (BLAST) was improved by optimally distributing the workload among thousands of computing nodes in the CDRH HPC clusters to gain a speedup of ~200x when comparing a 15 GB query against a 2.5 GB nucleotide sequence database.