**Biol 798 Project Update**

1. Matt Heffel

Cameron Amos

1. Open Reading Frame Predictor

(title is not necessarily set)

1. The data set can be any X genes from an organism. For now we will be working with the entire genbank dataset for PRRSV, Porcine Reproductive and Respiratory Syndrome Virus.

I choose this set because it is one stand of RNA that has 8+ open reading frames.

1. The problem that we are addressing with this dataset is why certain start codons are being selected to start an ORF and other start codons are just treated as normal codons.
2. Our workflow broken into steps,
3. Pull all of the files for species X from GenBank and remove incorrect pulls
4. Parse the annotated genomes and gather the Y nucleotides up and downstream from each start codon. Annotate these short sequences as positive or negative referring to ORF beginning.
5. Run these sequences through a supervised machine learning algorithm
6. Develop consensus sequence and scoring for start codon sequences
7. Be able to input an unannotated genome of the species into our algorithm developed from the consensus and scoring and output the genome annotated with its correctly identified ORFs.
8. https://www.ncbi.nlm.nih.gov/genbank/