* retrieve lots of protein F sequences of the hepatitis C virus
  + retrieve subtype 1a sequences from North American (dominant form)
  + retrieve subtype 1b sequences from Europe (dominant form)
* parse these sequences using software packages such as Biopython
* filter, sort, and reorganize
  + remove any sequences that do not have the proper fields in the FASTA header
  + sort the sequences by country, report the number of sequences in each
* Sampling
  + Report the range of collection years for each country
* Analyse the genetic relatedness of the data
  + generate a multiple sequence alignment of all sequences in each continent
    - possibly generate a phylogenetic tree
  + pairwise genetic distance measures between
    - countries within each continent
    - continents

- read in the fasta files (using biopython NOT seqUtils)

- pairwise align the sequences to E1 reference sequence from NC\_004102 reference using gotoh2

- translate nucleotide sequences into amino acid sequences using a MANUALLY CODED translation dictionary

Project Proposal

John Palmer

**Background:**

Hepatitis C virus (HCV) is a single-stranded RNA virus from the *Flaviviridae* family that is currently causing a pandemic affecting approximately 185 million people worldwide (Messina et al., 2015). In 2017 alone, roughly 400,000 infected individuals were killed due to diseases caused by HCV (WHO). Chronic infection with HCV is linked to the development of severe liver diseases, which include acute and chronic hepatitis, cirrhosis, and hepatocellular carcinoma (Messina et al., 2015).

The diversity of circulating hepatitis C viruses throughout the world has been classified into seven main genotypes and additional more closely related subtypes (Messina et al., 2015). Past studies have demonstrated that patient prognoses and overall progression of disease can differ between HCV genotypes and subtypes (Schröter et al., 1999). It is therefore, important to track the prevalence of different HCV subtypes in all regions of the world so that public health systems can respond accordingly.

The Los Alamos National Laboratory has curated a database of HCV sequences collected from infected patients around the world (https://hcv.lanl.gov/content/index). For this project, I plan to query this database for patient HCV sequence data to analyze the prevalence and relatedness of infections found on two different continents: Asia and South America. I will focus this analysis on the core protein of HCV as it is a well populated region within the database and plays a crucial functional role in the formation of new HCV particles.

**Data:**

* **Source:** HCV Database at Los Alamos National Laboratory
* **Format:** Nucleotide sequences in FASTA format
* **Necessary Metadata:** Subtype, patient ID, accession number
* **Target Gene:** Core protein
* **Geographical Regions:**
  + South America
  + Asia
* **Additional Information:** Reference sequence of HCV (NC\_004102)

**Objectives:**

1. To sort sequences by country and find the most prevalent HCV subtype in each country and continent
2. To sort sequences by subtype and compute genetic distances between these different subtypes by comparing random samples of sequences to each other
3. To translate nucleotide sequences into amino acid sequences and determine whether there are differences in hydrophobicity between subtypes
4. To find instances of patients who have been sequenced more than once