* 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

Hepatitis C virus (HCV) is a single-stranded RNA virus from the *Flaviviridae* family that is currently causing a pandemic affecting approximately 120 million people worldwide. In 2017 alone, roughly 400,000 infected individuals were killed due to diseases caused by the Hepatitis C virus. Infection with HCV is strongly linked to the development of severe liver diseases, which include acute and chronic hepatitis, cirrhosis, and hepatocellular carcinoma.

The diversity of circulating hepatitis C viruses throughout the world has been classified into seven main genotypes and additional more closely related subtypes. Past studies have demonstrated that patient prognoses and progression of disease can differ between HCV genotypes and subtypes.

* What is the hepatitis C virus
* How does it affect the human population
  + Causes both acute and chronic hepatitis
  + An estimated 120 million people are living with it and 400,000 people die of it each year
* Subtypes
  + Why are they important
  + How do they differ
* Scope / aim of the study
  + My study aims to analyze the HCV sequence

Data

* **Source:** HCV Database at Los Alamos National Laboratory
* **Format:** Nucleotide sequences in FASTA format
* **Target Gene:** Protein F
* **Subtypes:**
  + 1a
  + 1b
* **Geographical Regions:**
  + North America
  + Europe
* **Additional Information:** Reference sequence of HCV (NC\_004102)

Objectives

1. To compare the quality and availability of Hepatitis C sequence data in Europe and North America
2. To analyse the genetic relatedness of HCV sequences within each continent, and between the two different subtypes 1a and 1b