* 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