This program analyzes FASTA sequences to find regions of chemical and structural distinctness. It takes in two FASTA files, a name for an output file, and a desired Kmer length via the command line. It scores them on simple scales based one structurally complex residues and hydrophilicity given their frequent use as predictors for antigenicity. More on the thought process behind the ASH scale can be found on the ASH github. The intent is that this tool can help users analyze protein sequence for epitope/antigen selection by automating the search for epitopes that are distinct to their protein or, conversely, find regions that are well conserved in both. The output is a csv file containing information on all the epitopes in the first protein, and the following pieces of information:

Position in the sequence
Hydrophilicity distinctness rating
Percentage of Hydrophilic residues
Structural complexity distinctness rating
Percentage of structurally distinct residues
Analog sequences, ie the kmer it is aligned to in sequence 2

The script was written using Python 3.5, with the additional dependency of scikit-bio.

The test_package includes two FASTA files for testing purposes, so main the program, **run_ash.py**, could be run as follows:

$ python3 run_ash.py ENV_HV1MN.fasta ENV_HV1VI.fasta test_out 16

The first two arguments are the sequence to align, test_out will be the name of the csv file, and the kmers will be 16 residues long.

For those wish to script with the tools themselves, the package also includes **ash.py** which is a version of the ASH class without the reporting or error control, and the inclusion of a “getter” method (get_entries) to allow users to access the the entries themselves. **ash_example_script.py** demonstrates how this might be used,