Bioinformatics

Lab 3

A.

The conditional probability of AC is signicantly higher than most others, at 42.2%, with the next highest being nearly 8% less than it. The lowest is GT with a probability of only 9.7%.

B.

It turns out that the total probabilities were becoming too small to be dealt with by the default float class implemented in numPy so I had to use the “decimal’ module’s ‘Decimal’ class. This rendered very small numbers (thousands of decimal places in). I then took the natural (base e) logarithm to render a number a bit easier to comprehend. The probability for testN sequence, 1.44E-1030, was much smaller relative to the probability for test1 sequence, 4.08E-990. The ratio of the probabilities test1/testN produced a number much greater than 1 (93.145 to be exact) supporting this observation I made.

The test1 sequence appears to belonged to *Influenza A virus (A/Charlottesville/28/95(H1N1)) hemagglutinin (HA) gene, partial cdsi* as a perfect match with 100% coverage.

The testN sequence appears to belong to no known entity according to the BLAST database.

That having been said, it makes prefect sense that the sequence test1 was much more probable than the testN sequence given that the original DNA sequence which we trained the markov chain with was an Influenza A virus, and the test1 sequence was a specific strain of that influenza virus. It’s pretty near that we are able to train this matrix and obtain likelyhood’s and using that decide whether a given sequence was similar or not or likely given the training from the other sequence. Pretty neat lab.