Turning Biology into Mathematics Presentation

A short synopsis of Lab 2

Jean-Sebastien Roger[1]

[1]Florida Atlantic University

Biostatistics, Fall 2020

Abstract

This experiment was aimed at systematically re-coding the amino acid sequences that makeup proteins, into number sequences to gather meaningful empirical data from our otherwise categorical dataset.
Our experiment was successful in that we were able to construct a loop which takes in a string of characters and outputs a sequence of number conversions. We then used these numbers to calculate a mean, standard deviation, and a few plots which all provide insight into to the distribution of amino acids within particular proteins

Procedure

Using python code, we:

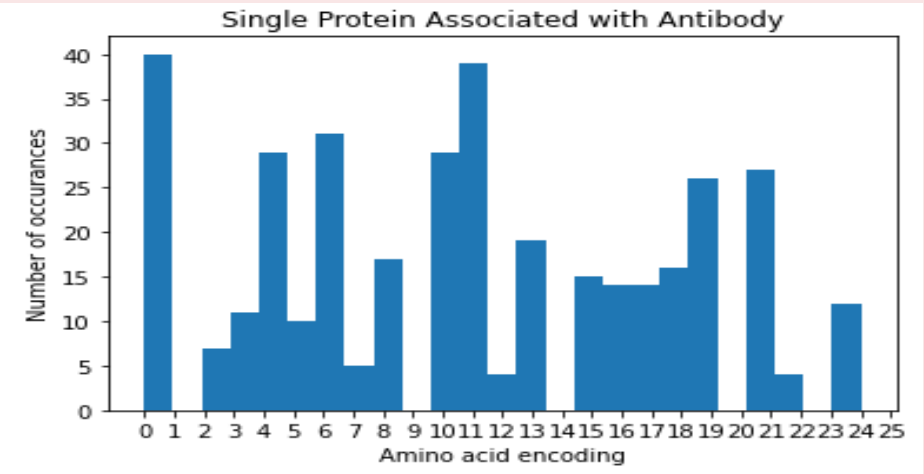
1. Read in our data consisting of 4,000 proteins
2. Transform the amino acid sequences which make up each protein into a sequence of numbers
3. Calculate the mean and standard deviation of the amino acids being studied
4. Plot the numerical distribution of a single protein
5. Calculated various measurements of distributions for the lengths and means of the sample of proteins
6. Plot the distributions of protein lengths and compare

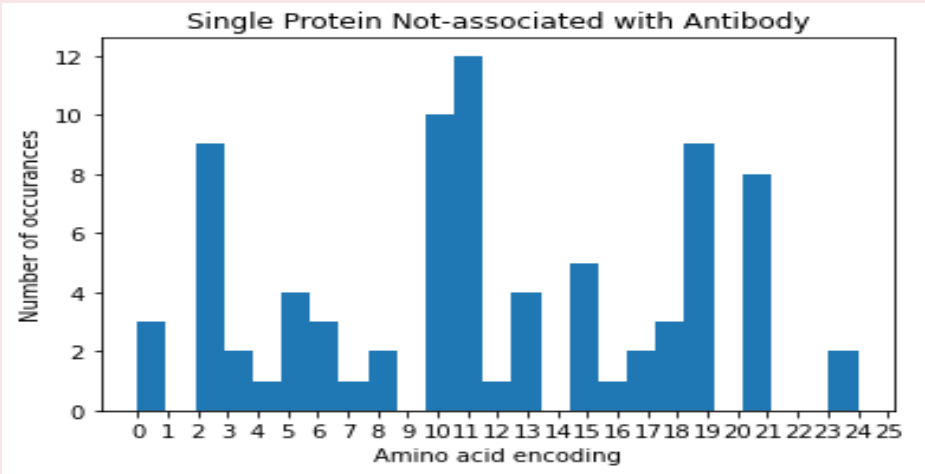
Transformation function call

$$X = process\_strings(ProteinList) \tag{1}$$

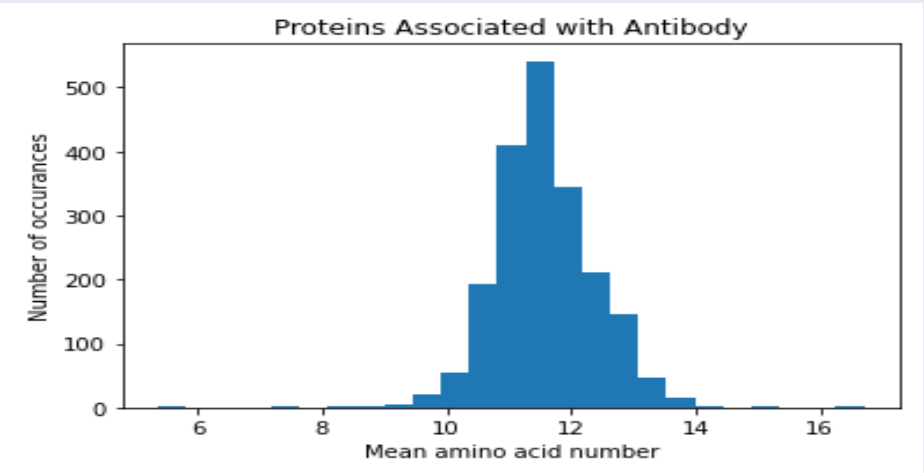
Comparisons between protein groups

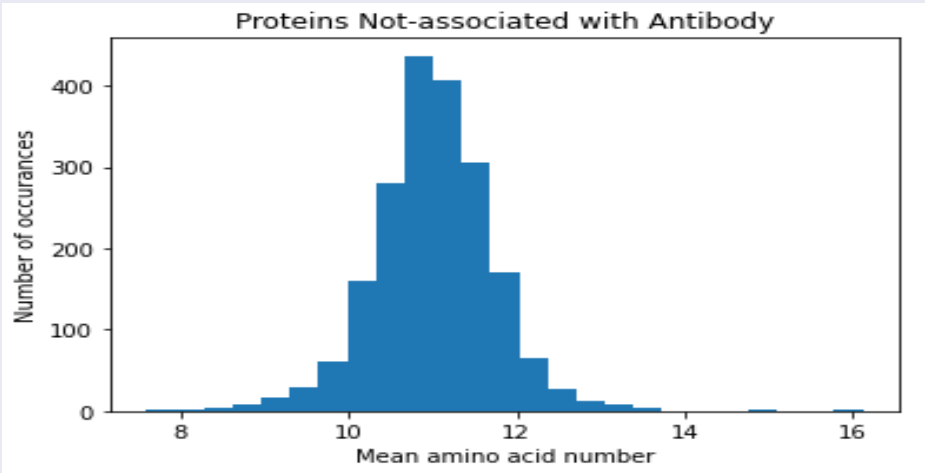
Encoding distribution for a single protein





Distribution of means from all sampled proteins





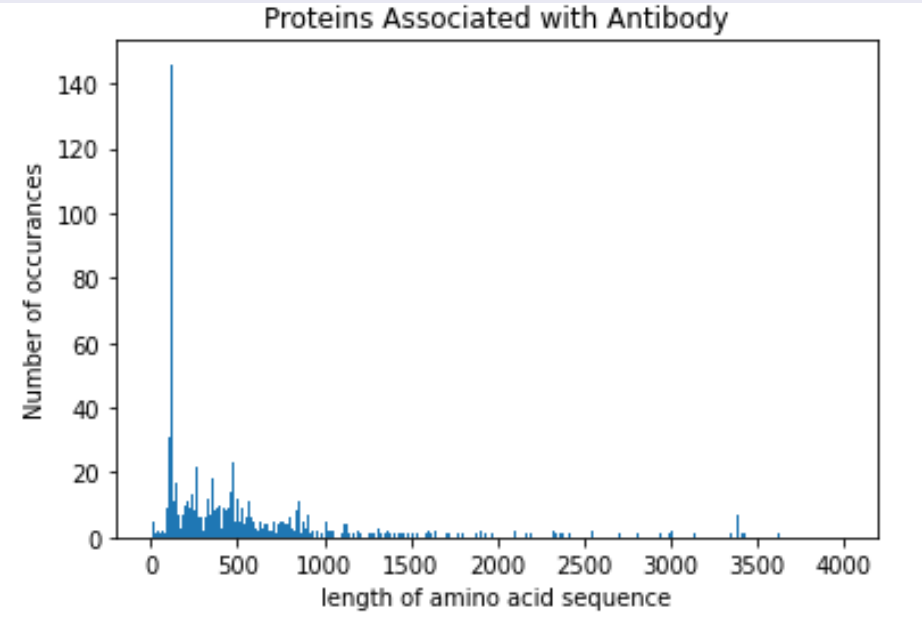
Metrics

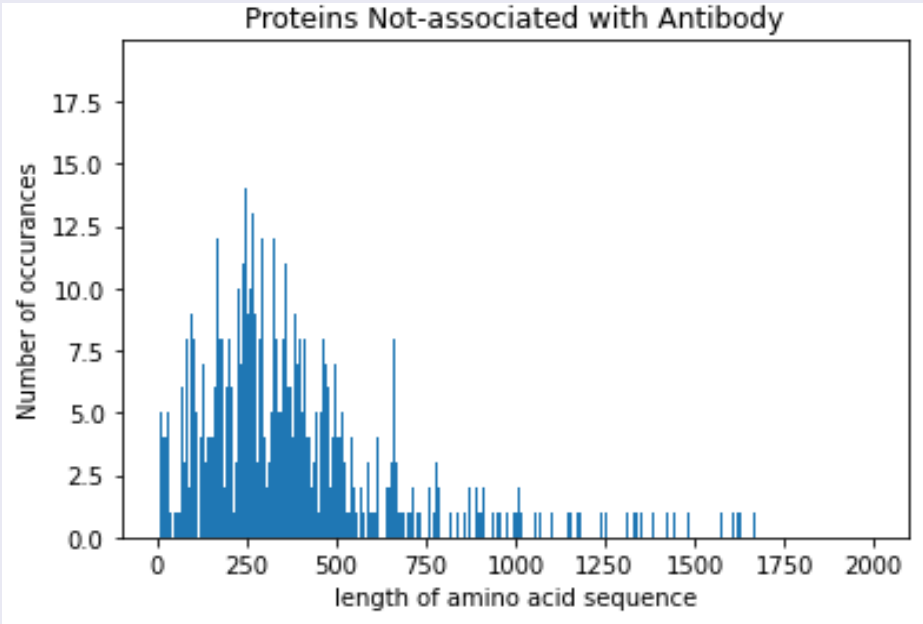
From the above slide we can see how there are possibilities for comparison once the amino acids have been converted to sequences of numbers. The function we created for this transformation is located in the lab report for reference.
Also, this transformation allows us to perform statistical tests on aspects of the distribution, for example, we can take a look at the distribution of means and compare the means from proteins associated with antibody against proteins not associated with antibody. In doing this, we find that these two distributions of means are significantly different.
ks_2samp(Means_X, Means_Y)
Ks_2sampResult(statistic=0.3155, pvalue=2.3415567591665026e-88)

Protein Lengths

We also took a look at the lengths of the proteins between each group. After plotting, we can see that the lengths of proteins not associated with the word antibody, tend to be shorter than the proteins associated with the word antibody.

Protein lengths





Conclusion

The main goal of this experiment was to measure and analyze two sets of protein groups. We were able to perform some preliminary comparisons between the two groups after first encoding the amino acid sequences into numerical representations. Calculating measurements such as mean and standard deviation of the length of amino acid sequences allowed us to realize how the proteins associated with antibody tend to larger and have twice the amino acid variation as compared to the proteins not associated with antibodies. Further, we were able to conclude that the two distributions of amino acid sequences were in fact significantly different at a 99% significance level with a K-S statistic p-value of:

$$p - value = 2.34x10^{-88} < 0.01 \tag{2}$$

.

In all, this lab was a good introduction into the world of encoding, with respect to amino acids and proteins. Further investigation should be done to find additional ways to compare the numerical distributions of the amino acid sequences which we have encoded.