Turning Biology Into Mathematics

Jean-Sebastien Roger

November 16, 2020

Abstract

In this experiment we examined the relationship between proteins and the primary structure sequences of amino acid which these proteins are constructed from. Our experiment confirmed that we are able to view the amino acid sequences within individual proteins and further confirmed our ability to manipulate the typical sequence of amino acids into lists of numbers which we can then perform additional calculations and further summaries.

1 Introduction

We begin this lab with a short introduction into amino acids and their one and three letter codes which we will be most interested in. For example, we have Alanine with a 3-letter code: Ala, and 1-letter code: A. Within the uniprot database that we will reference, we have access to thousands of proteins and their respective amino acid sequences.

2 Theory

We are interested in first enumerating amino acid sequences within proteins, as well as identifying and applying analytical methods to add insights into the sequences we are witnessing. This could potentially allow us to find patterns in the sequences that would normally be ignored when looking at a sequence of letters.

3 Procedures

We begin by reading in the uniprot data and creating two distinct lists of proteins, one related to antibodies and one not related to antibodies.

```
!pip install git+https://github.com/williamedwardhahn/mpcr
from mpcr import *

X, Y = get_uniprot_data('=antibody', '!antibody', 2000)
```

We convert our list of proteins with alphabetic sequences into lists of proteins with numerical amino acid sequences.

```
def process_strings(c):
    # Takes in a list of sequences 'c' and turns each one
    # into a list of numbers.

    X = []

    for  m, seq in enumerate(c):
        x = []
        for letter in seq:
            x.append(max(ord(letter)-97, 0))
            # In this line, ord() will take the unicode conversion
            # of the letter enumerated in seq, then will make sure
            # to take the max between 0 and the conversion

        X.append(x)

    return X

X = process_strings(X)
Y = process_strings(Y)
```

4 Analysis

In this section we will review the results of the enumeration of the amino acid sequences. We will also perform simple statistics to speak to the centrality and the levels of variation of the amino acids within each protein.

In order to take a look at the number of amino acids within each protein we may use the following code block:

```
X_lengths = [len(s) for s in X]

Y_lengths = [len(s) for s in Y]

X_lengths[0]
    369

Y_lengths[0]
    82
```

Now we can now proceed to looking at measures of centrality and variability

```
np.mean(X[0]),np.std(X[0])
    (10.987967914438503, 6.668917015482973)


chr(int(round(np.mean(X[0])+97)))
        'l'
```

4.1 Visualizations

When looking at data, a visual representation of how the values are distributed can often times lead to meaningful insights, in Figure 1 we see a typical way of presenting the lengths of amino acid chains within a list of proteins.

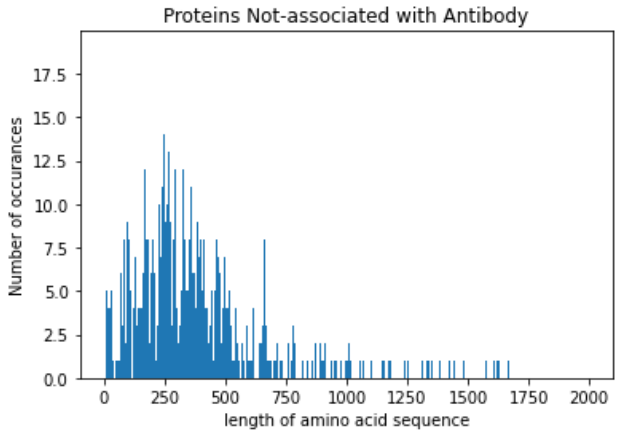

Figure 1: This here is a plot showing the length of amino acid chains for each protein in a list of 2,000 proteins not associated with the keyword 'antibody'.

In Figure 1 we can see that the majority of the proteins not associated with the keyword 'antibody' have a length less then 500 amino acids long.

5 Conclusions

In this lab, we were introduced to the potential of analyzing the make up and purpose of amino acids using mathematical methods. The process of converting the amino acid strings into numbers is a very useful technique where we are encoding alphabetic representations into numerical figures, this type of encoding arises in various other analyses such as encoding ordinal levels of a variable to use them to make predictions within typical regression models. We even see encoding when converting nominal levels of a variable using methods like one-hot encoding, to create a binary representation of our categorical variable. As a result of converting categorical attributes, like amino acid sequences, into numbers, we are able to now utilize methods like mean and standard deviation essentially allowing us to easily look at relationships between amino acid sequences rather than just the sequences themselves.

After getting a taste for encoding amino acids into numerical representations that we can mathematically measure, there's a whole new world of analysis that opens up to us. It would be interesting to revisit this lab utilizing other techniques of analysis. Perhaps using these numerical representations of amino acids we can train a classifier which could predict or classify unknown proteins. We could possibly use this information to distinguish between or cluster antibodies.

References

[1] StriverStriver(underscore)79 at Codechef and codeforces D, et al. "Ord() Function in Python." GeeksforGeeks, 7 Dec. 2017, www.geeksforgeeks.org/ord-function-python/.