MATHEMATICS DEPARTMENT

BIOSTATISTICS

FALL 2020

Notebook - Lab 2

Kerry-Ann Bartley Donaldson
Email: kbartleydona2019@fau.edu

2nd October 2020

Contents

1 Introduction

Turning Biology into Mathematics

In this lab demonstration we will access the uniprot database and create a new dataset. Our dataset will consist of four thousand proteins, half associated with the keyword antibody and the other half not related to the keyword. The proteins are represented by their primary structure sequence of amino acids, in other words each protein is a string of letters representing each amino acid in the sequence.

2 The Dataset

The one-letter and three-letter codes for amino acids used in the knowledgebase are those adopted by the commission on Biochemical Nomenclature of the IUPAC-IUB

One-letter code Three-letter code Amino-acid name
A Ala Alanine
R Arg Arginine
N Asn Asparagine
D Asp Aspartic acid
C Cys Cysteine
Q Gln Glutamine
E Glu Glutamic acid
G Gly Glycine
H His Histidine
I Ile Isoleucine
L Leu Leucine
K Lys Lysine
M Met Methionine
F Phe Phenylalanine
P Pro Proline
S Ser Serine
T Thr Threonine
W Trp Tryptophan
Y Tyr Tyrosine
V Val Valine
O Pyl Pyrrolysine
U Sec Selenocysteine
B Asx Aspartic acid or Asparagine
Z Glx Glutamic acid or Glutamine
X Xaa Any amino acid

3 Python Codes for Analysis

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

```python
# This code will create a dataset from the uniprot database
X, Y = get_uniprot_data('=antibody', '!antibody', 2000)
# create dataset with 2000 samples
```

```python
number_X = len(X)          #Assigning name to the length of X and Y from the dataset
number_Y = len(Y)
```

```python
print(number_X)            #printing out the length
print(number_Y)
```

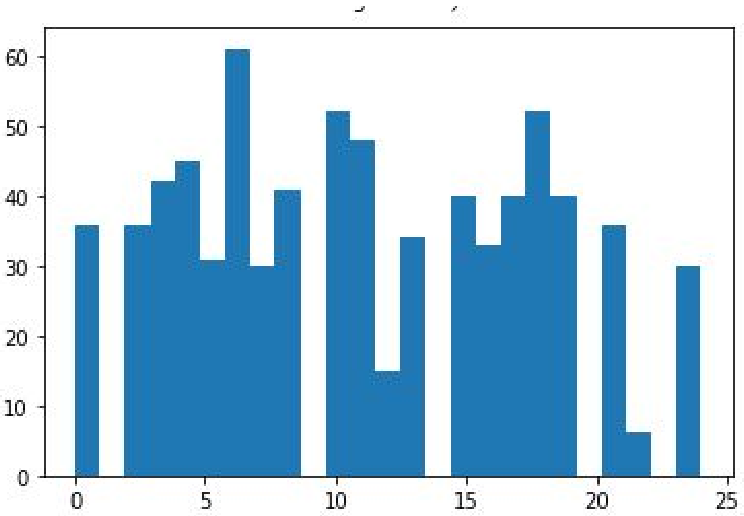
```python
print(len(X[0])),print(len(Y[0]))    #printing out the length of the 1 st protein in X (antibodies) and Y (antibodies)
print(len(X[1])),print(len(Y[1]))
print(len(X[2])),print(len(Y[2]))
print(len(X[3])),print(len(Y[3]))
print(len(X[1999])),print(len(Y[1999]))
```

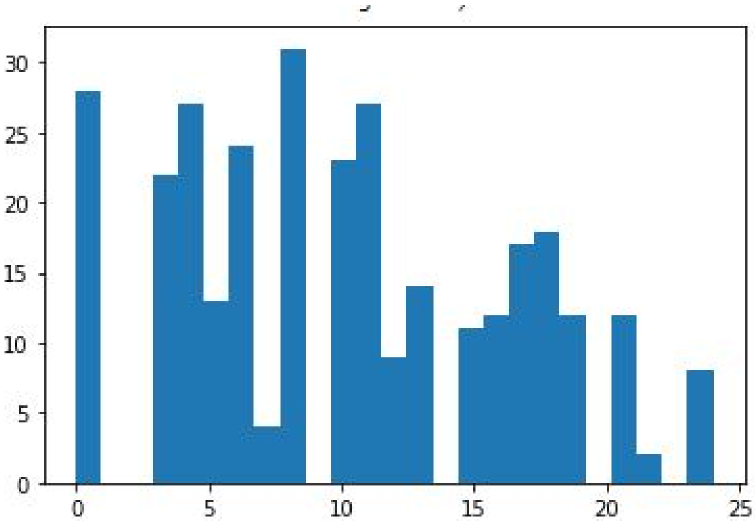
```python
X[0] #Amino acid sequence of the first protein on the list of proteins associated with 'antibody'
```

```python
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))

        X.append(x)

    return X
```

```python
X = process_strings(X)     #Assigning names
Y = process_strings(Y)
```

```python
plt.hist(X[0],25)   #plotting histogram of the 1st X protein, 25 of them
plt.hist(Y[0],25)  #plotting histogram of the 1st Y protein, 25 of them
```





```python
np.mean(X[0]),np.std(X[0]) #meaningful? Finding the mean and standard deviation of the 1st proteins
```

```python
np.array(X[0]).shape           #size of the first protein in X
```

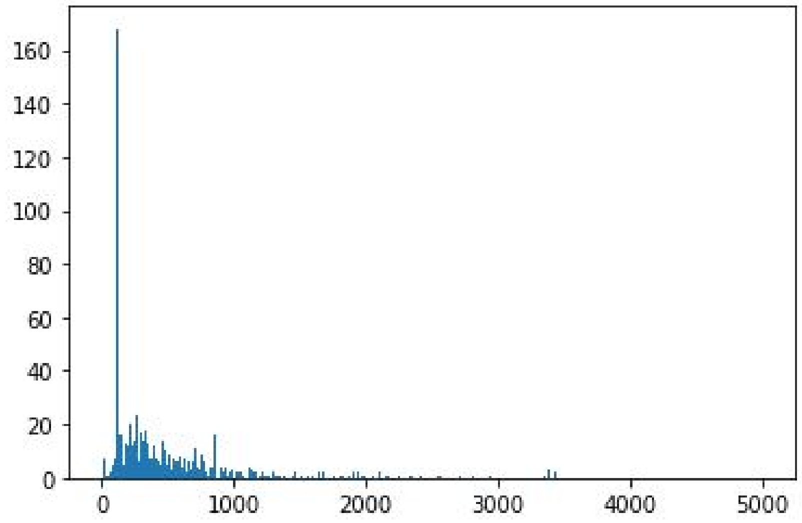
Find lengths of all proteins:

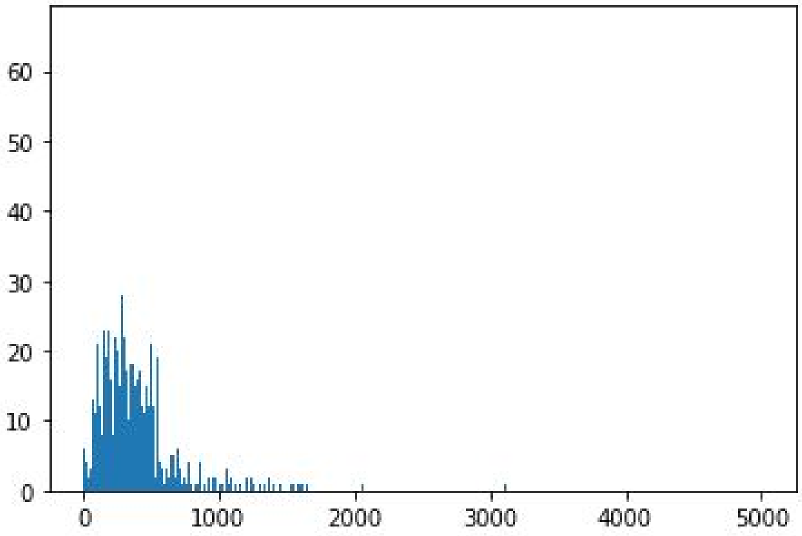
```python
X_lengths = [len(s) for s in X]
Y_lengths = [len(s) for s in Y]
```

```python
np.max(X_lengths)  #Max length in X
np.max(Y_lengths)     #Max length in Y
```

```python
np.min(X_lengths)     #min length in X
np.min(Y_lengths)     #min length in Y
```

```python
plt.hist(X_lengths,bins=1000,range=(0,5000));       #plot a histogram of X lengths with intervals 1000 from 0-5000
plt.hist(Y_lengths,bins=1000,range=(0,5000));       #plot a histogram of Y lengths with intervals 1000 from 0-5000
```





K. Bartley Donaldson