

```
In [ ]: ##This code allows you to take a DNA string and convert it to the corresponding
##It then outputs information on the amino acid sequence including sequence length
##Finally, this code outputs a pie chart of the amino acids in the sequence based on their frequency
##Refer to the README.md file for additional information on required inputs and outputs
##Requires entire DNA sequence to be on a single line in the txt file.
##Author:Michaela Lindemann
```

```
In [28]: import os
import numpy as np
#Open DNA.txt file and saved data to a numpy array called DNA_sequence.
#This subroutine reads a txt and puts the data into a numpy array of string.
#It also prints the DNA sequence found in the input text file.
#INPUT:
    #Prompt for location of DNA txt file
#OUTPUT:
    #1)DNA_sequence:numpy array of string
    #2)Prints DNA_sequence
data_location = input('Enter path to DNA file:')
DNA_sequence_text= np.genfromtxt(fname=data_location, dtype='unicode')
DNA_sequence=str(DNA_sequence_text)
print(DNA_sequence)
```

```
Enter path to DNA file:/Users/michaelalindemann/Prog. Data Analysis/DNA.txt
TGGGTTGATTCCACACCCCCGCCCCGGCACCCGCGTCCGCGCCGTGGCCATCTACAAGCAGTCACAGCACATGACGGAG
GTTGTGAGGCGCTGCCCCCACCATGAGCGCTGCTCAGARAGCGAT
```

```
In [29]: import os
#Save amino acid key data in key_data to be used later to convert the DNA sequence to amino acid sequence
#This subroutine reads a txt and outputs the data as a list.
#INPUT:
    #Prompt for location of Key txt file

key_file_location = input('Enter path to Key file:')
key_file=open(key_file_location,'r')
key_data = key_file.readlines()
key_file.close()
```

```
Enter path to Key file:/Users/michaelalindemann/Prog. Data Analysis/Key.txt
```

```
In [30]: #Translate the DNA sequence to the corresponding one letter amino acid codes and
#This subroutine takes the DNA sequence and converts it to an amino acid sequence
#INPUT:
    #DNA_sequence data in string form inputted from txt file
#OUTPUT
    #Show on the screen the amino acid sequence in list format

amino_acid_decode=[]
codon_array=[]
#3 letter codons
for i in range(0, len(DNA_sequence), 3):
    codon_1=DNA_sequence[i:i+3]
    codon_array.append(codon_1)
for codon in codon_array:
    for line in key_data:
        if codon in line:
            codon_line_split=line.split()
            amino_acid_decode.append(codon_line_split[-1])
print(amino_acid_decode)
```

```
['W', 'V', 'D', 'S', 'T', 'P', 'P', 'P', 'G', 'T', 'R', 'V', 'R', 'A', 'V',
'A', 'I', 'Y', 'K', 'Q', 'S', 'Q', 'H', 'M', 'T', 'E', 'V', 'V', 'R', 'R',
'C', 'P', 'H', 'H', 'E', 'R', 'C', 'S', 'S', 'D']
```

```
In [31]: #Determine the amino acid sequence length.
#This subroutine shows the length of the amino acid sequence.
#INPUT:
#The amino acid sequence in list form
#OUTPUT:
#On the screen, the length of the amino acid sequence as an integer

amino_acids_list = set(amino_acid_decode)
sequence_length=len(amino_acid_decode)
print (sequence_length)
```

40

```
In [32]: #Determine the percentage of each amino acid in the total sequence.
#This subroutine takes shows each amino acid and its percentage of the total sequence
#INPUT
#The amino acid in list form
#OUTPUT
#The one letter amino acid codes present in the list followed by their frequency

percent=[]
for amino_acid in amino_acids_list:
    amino_acid_count = amino_acid_decode.count(amino_acid)
    amino_acid_percent = (amino_acid_count/sequence_length) * 100
    percent.append(amino_acid_percent)
    print(amino_acid,":",round(amino_acid_percent,1),"%")
```

```
S : 10.0 %
H : 7.5 %
W : 2.5 %
Y : 2.5 %
K : 2.5 %
T : 7.5 %
M : 2.5 %
I : 2.5 %
E : 5.0 %
C : 5.0 %
D : 5.0 %
V : 12.5 %
P : 10.0 %
G : 2.5 %
R : 12.5 %
Q : 5.0 %
A : 5.0 %
```

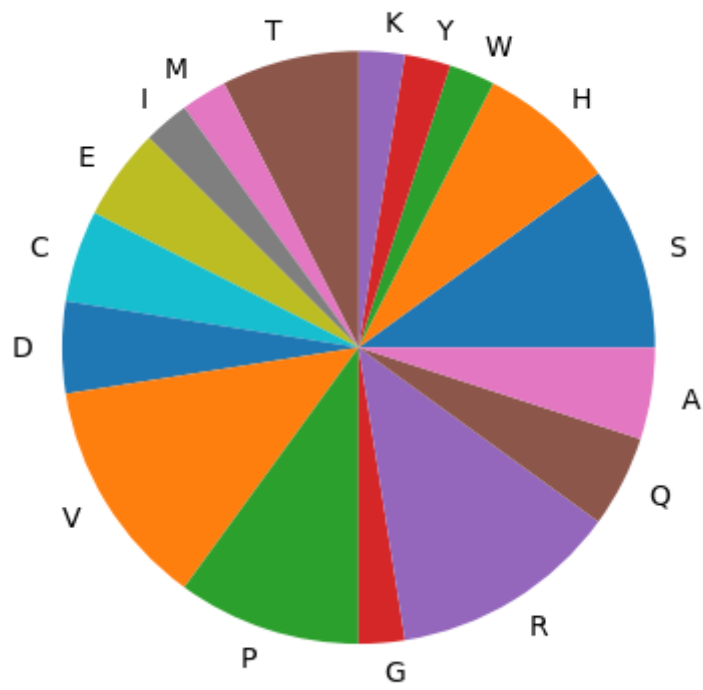
```
In [33]: #We graph the amino acids in a pie chart based on their frequency by percentage
#This subroutine takes the amino acid sequence breakdown by percentage and plots it
#INPUT:
#1)List of percentage for each amino acid in the sequence
#2)List of amino acid one letter codes present in the sequence
#OUTPUT
#Pie plot on screen showing frequency of amino acids in sequence by percentage

import matplotlib.pyplot as plt

plt.pie(percent, labels = amino_acids_list)
```

```
plt.title('Amino Acid Breakdown by Percentage')  
plt.show()
```

Amino Acid Breakdown by Percentage



In []: