Assignment 9

1. The Hamming distance between two sequence is given as the minimum number of symbol substitutions required to transform one sequence to other. Whereas the Euclidean distance is the root mean square distance. That is it is similar to the Pythagorean theorem.
   Using the given theory, the program for calculating the hamming and Euclidean distance between two sequences is as follows:

```
def euclidean(score1, score2):
    total=0
    for i in range(len(score1)):
        total+=(abs(score1[i]-score2[i]))**2
    total=total**0.5
    return total

def hamming(score1, score2):
    total=0
    for i in range(len(score1)):
        total+=abs(score1[i]-score2[i])
    return total

def scores(string):

AminoAcids=['A','C','D','E','F','G','H','I','K','L','M','N','P
','Q','R','S','T','V','W','Y']
    score=[0 for i in range(len(AminoAcids))]
    for i in range(len(string)):
        score[AminoAcids.index(string[i])]+=1
    for i in range(len(score)):
        score[i]=score[i]/len(string)
    return score

strings=['AMENLNMDLLYMAAAVMMGLAAIGAAIGIGILGGKFLEGAARQPDLIPLLRT
QFFIVMGLVDAIPMIAVGLGLYVMFAVA',

'AADVSAAVGATGQSGMTYRLGLSWDWDKSWWQTSTGRLTGYWDAGYTYWEGGDEGAGKHSL
SFAPVFVYEFAGDSIKPFIEAGIGVAAFSGTRVGDQNLGSSLNFEDRIGAGLKFANGQSVGV
RAIHYSNAGLKQPNDGIESYSLFYKIPI',

'MALLPAAPGAPARATPTRWPVGCFNRPWTKWSYDEALDGIKAAGYAWTGLLTASKPSLHHA
TATPEYLAALKQKSRHAA']
scorearr=[]
for i in strings:
    scorearr.append(scores(i))
euclideandist=[]
hammingdist=[]
for i in range(-1, len(scorearr)-1):
    print('euclidean score between {} and {}
are'.format(len(strings) if i==-1 else i+1, i+2),
euclidean(scorearr[i], scorearr[i+1]))
    euclideandist.append(euclidean(scorearr[i],
scorearr[i+1]))
for i in range(-1, len(scorearr)-1):
    print('hamming score between {} and {}
are'.format(len(strings) if i==-1 else i+1, i+2),
hamming(scorearr[i], scorearr[i+1]))
    hammingdist.append(hamming(scorearr[i], scorearr[i+1]))

print('The closest Euclidean distance is', min(euclideandist))
print('The closest Hamming distance is', min(hammingdist))
```

and the output is:

```
euclidean score between 3 and 1 are 0.2208681669138957
euclidean score between 1 and 2 are 0.20106216842153501
euclidean score between 2 and 3 are 0.20112952107271115
hamming score between 3 and 1 are 0.8433544303797469
hamming score between 1 and 2 are 0.665728476821192
hamming score between 2 and 3 are 0.726632576075111
The closest Euclidean distance is 0.20106216842153501
The closest Hamming distance is 0.665728476821192
```

2. We use UniProt to get the sequences with query name “beta barrel membrane proteins” and we get 703 reviewed proteins. We put the downloaded fasta file in CD-HIT and run the scan for the sequence Identity 40%, 50%, 75% and 90% and we get the sequences.

Percentage sequence Identity	Number of protein sequences (total=703)
40%	246
50%	305
75%	430
90%	510

(The Fasta files for these sequence identity are attached with the file names as “##percent” for the amount of percentage)

3. The PISCES server hasn’t come online since past week.

4. The PISCES server hasn’t come online since past week.

5. 50% cutoff sequence identity from UniProt and CD-HIT are (since PISCES hasn’t come online since past week, we are not including it.)

UniProt:	365
CD-HIT:	305