NCBI has DNA codes

1. In the top, from database scroll down menu, select Nucleotide
2. Enter the code of the DNA you're interested in
3. Click on FASTA
4. Copy the DNA code and save it to a txt document
5. download amino acid sequence to check your work at the end. In NCBI page, find CDS in the left side. Inside details at the bottom copy what comes after translation.
6. save it to a txt file.

In [8]:

```
pwd
```

Out[8]:

```
'C:\\Users\\maria\\Documents\\python for github'
```

In [13]:

```
inputfile = 'DNA3.txt'
```

In [24]:

```
f= open(inputfile, 'r')
seq = f.read()
```

In [26]:

```
seq
```

Out[26]:

```
'GGTCAGAAAAAGCCCTCTCCATGTCTACTCACGATACATCCCTGAAAACCACTGAGGAAGTGGCTTTTCA\nGATCATCTTGCTTTGCCAGTTTGGGGT
TGGGACTTTTGCCAATGTATTTCTCTTTGTCTATAATTTCTCT\nCCAATCTCGACTGGTTCTAAACAGAGGCCCAGACAAGTGATTTTAAGACACATGG
CTGTGGCCAATGCCT\nTAACTCTCTTCCTCACTATATTTCCAAACAACATGATGACTTTTGCTCCAATTATTCCTCAAACTGACCT\nCAAATGTAAAT
TAGAATTCTTCACTCGCCTCGTGGCAAGAAGCACAAACTTGTGTTCAACTTGTGTTCTG\nAGTATCCATCAGTTTGTCACACTTGTTCCTGTTAATTCA
GGTAAAGGAATACTCAGAGCAAGTGTCACAA\nACATGGCAAGTTATTCTTGTTACAGTTGTTGGTTCTTCAGTGTCTTAAATAACATCTACATTCCAAT
TAA\nGGTCACTGGTCCACAGTTAACAGACAATAACAATAACTCTAAAAGCAAGTTGTTCTGTTCCACTTCTGAT\nTTCAGTGTAGGCATTGTCTTCTT
GAGGTTTGCCCATGATGCCACATTCATGAGCATCATGGTCTGGACCA\nGTGTCTCCATGGTACTTCTCCTCCATAGACATTGTCAGAGAATGCAGTACA
TATTCACTCTCAATCAGGA\nCCCCAGGGGCCAAGCAGAGACCACAGCAACCCATACTATCCTGATGCTGGTAGTCACATTTGTTGGCTTT\nTATCTTC
TAAGTCTTATTTGTATCATCTTTTACACCTATTTTATATATTCTCATCATTCCCTGAGGCATT\nGCAATGACATTTTGGTTTCGGGTTTCCCTACAATT
TCTCCTTTACTGTTGACCTTCAGAGACCCTAAGGG\nTCCTTGTTCTGTGTTCTTCAACTGTTGAAAGCCAGAGTCACTAAAAATGCCAAACACAGAAGA
CAGCTTT\nGCTAATACCATTAAATACTTTATTCCATAAATATGTTTTTAAAAGCTTGTATGAACAAGGTATGGTGCTC\nACTGCTATACTTATAAAAG
AGTAAGGTTATAATCACTTGTTGATATGAAAAGATTTCTGGTTGGAATCTG\nATTGAAACAGTGAGTTATTCACCACCCTCCATTCTCT'
```

In [27]:

```
#to remove new lines from seq, we can use replace but we have to reassign it to seq if we want to save the result
s
seq=seq.replace('\n', '')
```

In [30]:

```
seq = seq.replace('\r', '')
```

In [37]:

```
table = {
'ATA':'I', 'ATC':'I', 'ATT':'I', 'ATG':'M',
'ACA':'T', 'ACC':'T', 'ACG':'T', 'ACT':'T',
'AAC':'N', 'AAT':'N', 'AAA':'K', 'AAG':'K',
'AGC':'S', 'AGT':'S', 'AGA':'R', 'AGG':'R',
'CTA':'L', 'CTC':'L', 'CTG':'L', 'CTT':'L',
'CCA':'P', 'CCC':'P', 'CCG':'P', 'CCT':'P',
'CAC':'H', 'CAT':'H', 'CAA':'Q', 'CAG':'Q',
'CGA':'R', 'CGC':'R', 'CGG':'R', 'CGT':'R',
'GTA':'V', 'GTC':'V', 'GTG':'V', 'GTT':'V',
'GCA':'A', 'GCC':'A', 'GCG':'A', 'GCT':'A',
'GAC':'D', 'GAT':'D', 'GAA':'E', 'GAG':'E',
'GGA':'G', 'GGC':'G', 'GGG':'G', 'GGT':'G',
'TCA':'S', 'TCC':'S', 'TCG':'S', 'TCT':'S',
'TTC':'F', 'TTT':'F', 'TTA':'L', 'TTG':'L',
'TAC':'Y', 'TAT':'Y', 'TAA':'_', 'TAG':'_',
'TGC':'C', 'TGT':'C', 'TGA':'_', 'TGG':'W',
}
```

In [40]:

```
table['CCT']
```

Out[40]:

```
'P'
```

In [44]:

```
len(seq)%3
```

Out[44]:

```
2
```

In [59]:

```
def translate(seq):
    """Translates a string containing a nucleotide sequence into a string containing the corresponding sequence o
f amino acids.
    Neuceotides are translated into triplets using a table dictionary; each amino acid is encoded with a string o
f length 1."""
    protein = ''
    if len(seq) % 3 ==0: # check if sequence is divisible by 3

        for i in range(0, len(seq), 3):     # loop over the sequence
            codon = seq[i: i+3]   # extract the codon
            protein += table[codon]
    return protein
```

In [50]:

```
translate('ATA')
```

Out[50]:

```
'I'
```

In [52]:

```
def read_seq(filename):
    """Reads txt file """
    with open(filename, 'r') as f:
        seq = f.read()
    seq = seq.replace('\n', '')
    seq = seq.replace('\r', '')
    return seq
```

In [55]:

```
protien_ref = read_seq('protien.txt')
```

In [56]:

```
DNA = read_seq('DNA3.txt')
```

In [68]:

```
protien_ref
```

Out[68]:

```
'MSTHDTSLKTTEEVAFQIILLCQFGVGTFANVFLFVYNFSPISTGSKQRPRQVILRHMAVANALTLFLTIFPNNMMTFAPIIPQTDLKCKLEFFTRLVA
RSTNLCSTCVLSIHQFVTLVPVNSGKGILRASVTNMASYSCYSCWFFSVLNNIYIPIKVTGPQLTDNNNNSKSKLFCSTSDFSVGIVFLRFAHDATFMSI
MVWTSVSMVLLLHRHCQRMQYIFTLNQDPRGQAETTATHTILMLVVTFVGFYLLSLICIIFYTYFIYSHHSLRHCNDILVSGFPTISPLLLTFRDPKGPC
SVFFN'
```

In [73]:

```
print(translate(DNA[20:932]))
```

```
MSTHDTSLKTTEEVAFQIILLCQFGVGTFANVFLFVYNFSPISTGSKQRPRQVILRHMAVANALTLFLTIFPNNMMTFAPIIPQTDLKCKLEFFTRLVAR
STNLCSTCVLSIHQFVTLVPVNSGKGILRASVTNMASYSCYSCWFFSVLNNIYIPIKVTGPQLTDNNNNSKSKLFCSTSDFSVGIVFLRFAHDATFMSIM
VWTSVSMVLLLHRHCQRMQYIFTLNQDPRGQAETTATHTILMLVVTFVGFYLLSLICIIFYTYFIYSHHSLRHCNDILVSGFPTISPLLLTFRDPKGPCS
VFFN
```

In [74]:

```
protein_translated = translate(DNA[20:932])
```

In [75]:

```
protien_ref == protein_translated
```

Out[75]:

```
True
```

DNA code in the website contains stop codon while in the translated code it does not.

transcription of this protien starts at 21 and ends at 938. Therefore, we need to slice the DNA sequence to start from 20 and end before the stop codon

In [ ]:

```

```