Data Science in Bioinformatics

Day 2

For all tasks please use text files for your answers and briefly comment your scripts within the code. Everybody participating should also finish the worksheet from Day 1

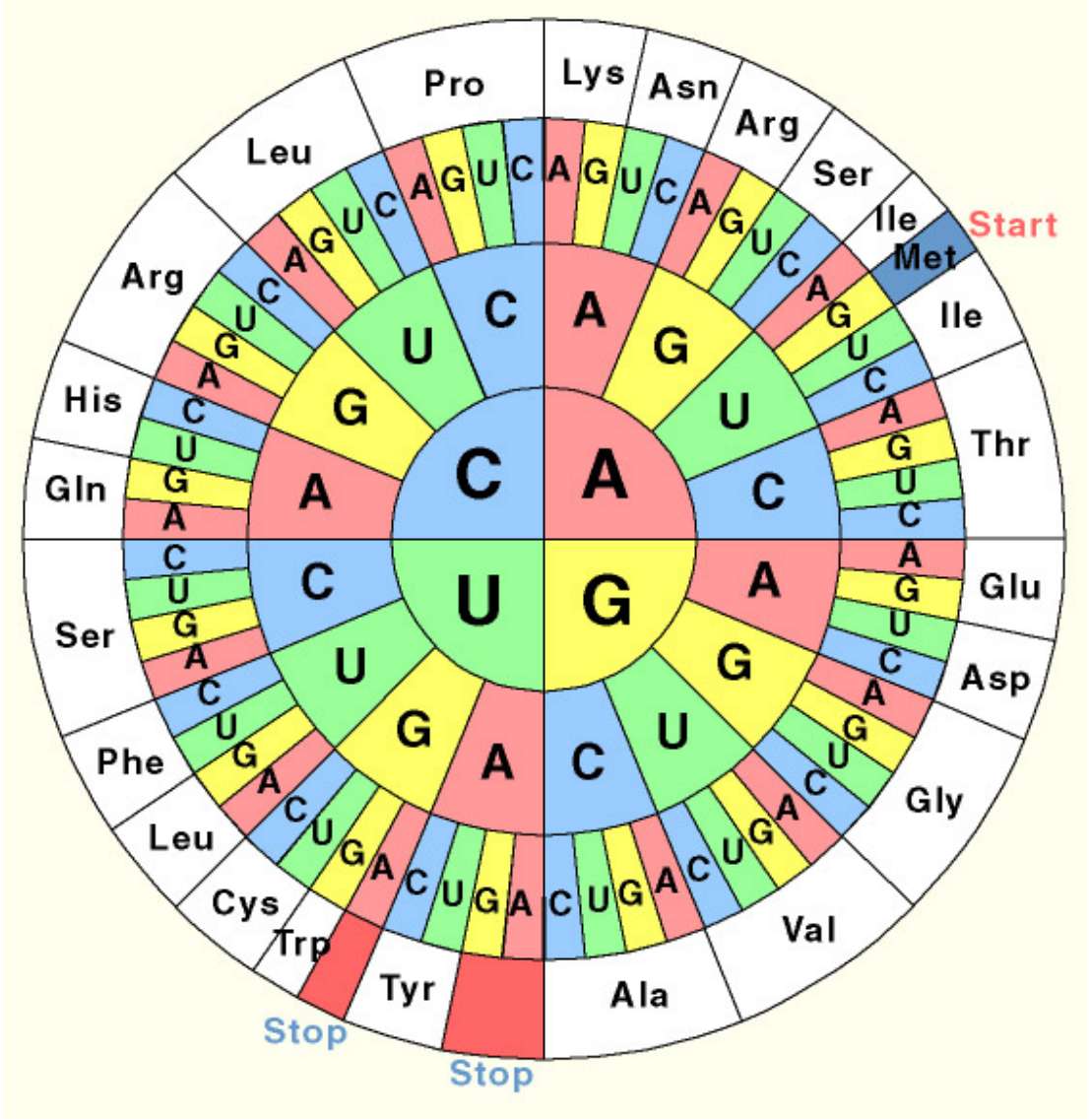
1) **Restriction enzymes**
   A restriction enzyme is capable of cleaving a DNA strand into fragments via palindromic recognition sites. What is special about the fragments after cleaving with specific restriction enzymes? Please find five restriction enzymes yourself and store them with identifier, sequence (and maybe necessary additional information 😉) in a capable format within a text file. Then write a script that is capable of reading a random DNA sequence from the terminal command line and check for your found restriction enzymes. The script should output the input sequence, all hits for restriction enzymes and the correctly cleaved fragments, all in a formatted manner.

2) **Protein biosynthesis**
   Protein biosynthesis is a biological core process, that produces all necessary enzymes and proteins a cell needs to live. The genetic information is stored in DNA strands, encoded in base triplets within open reading frames (ORFs). An ORF always starts with a START codon and ends with a STOP codon in the same reading frame. An ORF can contain a multitude of smaller ORFs, which commence with START codons downstream of the first START, lying within the same reading frame and forming a shorter ORF with the same STOP.

   For the expression of the genetic information, a reverse complement messenger RNA (mRNA, whats different compared to DNA?) copy of the gene is made, that is further processed (transcription). The mRNA is translated at Ribosomes into an amino acid chain. Please write a script that is capable of:

   a. Reading a sequence sample file in FASTA format
   b. Identify all ORFs, count them and output the number
   c. Exclude ORFs starting within a longer ORF, using the same STOP codon; mind the reading frame
   d. Output ORFs > 40 amino acids only
   e. Read in the provided codon.tsv file into a default python file structure, then translate and output the reverse complement mRNA sequence onto the terminal and the translated protein sequence (Header format: >(START,STOP)_protein) for the > 40 AA ORFs, switchable between one and three letter code into a file.

   Look at the attached figure. What do you recognize about the amino acids and their encoding?


Pro
Lys
Asn
Arg
Ser
Ile
Start
Met
Ile
Thr
Glu
Asp
Gly
Val
Ala
Tyr
Stop
Trp
Cys
Leu
Phe
Ser
Gln
His
Arg
Leu
C
A
G
U