Basic Concepts of Biology

1.1 (E) Genetic Code
In this exercise we will have a quick look at the genetic code (DNA) and how to translate it into protein sequences.

1.2 (H) Statistics
Note: In this exercise 'frequency' means the 'relative fraction'.
For example, the frequency of 'A' in 'ATAGGC' is $\frac{2}{6}$ or 0.333333, not 2.

Your task is to complete the Genome class in the genome.py file. The __init__ method should accept only a single parameter: the DNA sequence of a genome. The sequence should either be stored in a class variable for later use or processed during initialization.

(a) Complete the get_gC_content() method. It returns the GC content of the genome sequence, i.e. the combined fraction of G and C. Round the final value to 6 digits.

(b) Complete the get_codon_dist() method. Calculate the expected codon distribution based on the frequencies of the four different nucleotides in the genome sequence. Round each codon frequency to 6 digits. The return value of the method is a tree-like structure of nested dictionaries. Each node in the tree represents one of the four different nucleotides and the leaves contain the corresponding codon frequency. The path along the tree forms the different codons. For example, the codon frequency for 'AAA' can be accessed using dict_tree['A']['A']['A'].

(c) Complete the get_amino_acid_dist() method. It returns a dictionary containing the expected amino acid distribution of the genome sequence: the keys represent the 20 standard amino acids, the values the corresponding amino acid frequency. Calculate the amino acid distribution based on the previously calculated codon frequencies. Once again, round the final values to 6 digits.
1.3 (H) Open Reading Frames (ORFs) (10)

Complete the program called orffinder.py which does the following:

(a) The function get_orfs takes two parameters: a DNA sequence (string) and a minimum number of amino acids (see below).

(b) If the input sequence is not a valid DNA sequence it raises a TypeError exception.

(c) Find all ORFs that encode for proteins with the provided minimum number of amino acids, i.e. the resulting polypeptide has at least that many amino acids. Search in all six reading frames (primary and reverse-complementary strand).

(d) The return value is a list of ORFs represented as 4-tuples containing (in this order) the position of the first DNA residue, the position of the last DNA residue (including stop codon), the translated amino acid sequence as a single string, and a flag which is True if the ORF is parsed from the reversed strand (otherwise False). Positions start at index 0 (first nucleotide in the primary strand) and ORFs in the reverse-complementary strand must be indexed according to their position on the primary strand (i.e. start > stop, except if they are circular).

Example: (4269, 4141, 'MSEGISLEHIVAFLTKNITVSRYLTVLYRIRQKRKLENE', True)

(e) If multiple ORFs overlap and end with the same stop codon return only the longest ORF.

(f) Consider circular DNA, e.g. plasmids.

(g) The DNA index always has to be non-negative and less than the genome length.