Python exercise 4

ORF finder

V1. 28 October 2021

Copy the file U15422.fasta which is a text file in FASTA format and represents a DNA sequence. An experimental group has obtained data in this format from an organism and they wish to use it to identify proteins.

You will need to write a Python program which can read in a standard FASTA file as input, which is of the form:

```
>sequence identifier and details
DNA or Amino Acid sequence in one letter codes
```

For example:

```
>NM_174053.2 Bos taurus fibrillin 1 (FBN1), mRNA
CGCTCGGCATTATGCGGCGAGGGGGGCTGCTGGAGGTCGCCTTGGGATTTACCGTGCTCTT
AGCGTCCTACACGAGCCATGGGGCGGATACCAATTTGGAGGCTGGGAACGTGAAGGAAACC
AGAGCCAACCGGGCCAAGAGAAGAGGTGGCGGAGGACACAACGCGCTTAAAGGACCCAATG
TCTGTGGATCACGTTATAATGCTTACTGTTGCCCTGGATGGAAGACTTTACCTGGTGGAAA
```

Your programme should find all the possible open reading frames in the sequence (known as ORFs, beginning with Methionine (ATG[1]), and ending with a stop codon, i.e. TGA, TAA or TAG). You can assume that the longest ORF will be used, and you do not have to worry about any internal ones. You will have to analyse all six reading frames for the sequence, and predict ORFs in all of them.

You should output protein ORFs with the following FASTA format:

```
>orf_name
MAKSKDFPVKADFAAHVAIQSEFAHGV
>orf_name
MHILDECCAKSKDFPDFAAHVAYIQSEFAALLILPQWNSDAAHGV
```

`Orf_name` should be a unique and meaningful name for each ORF, which might include the original sequence identifier, the frame (forward or reverse) in which the ORF is found and the start position in original sequence. E.g. your ORF name could be something like:

```
NM_174053.2_F1_145
```

for an ORF starting at position 145 in the forward frame 1 in sequence NM_174053.2. You can choose what to include and how to present the orf name.

You should think carefully about the minimum size of an ORF. I suggest you add an option which controls this, e.g. you could use 50 amino acids as the minimum size.

Bonus tasks:

Error checking: spot and deal with improper formats, unknown amino acids, and any sequence gaps (there normally won't be any of course).

---

[1] Note: >90% of codons in bacterial genes start with ATG but around 10% don't. However, we'll assume they all do for the purposes of this exercise.