Homework 2: due Monday, Mon. 13th, midnight

1.) Write a Python script that considers the following DNA sequence

```
ATTGGGGAGGAGGCGAGTTGAGCGGCGGCAGTTCGCCTGCGTGCGCTGCGCGG
CGTCGACATCTGATCCGCACCATGGAAATCCCCGCTCAATCTTTGGAGCAGGGAT
GCGGGGCGATCAAGATGGGGATGCGGGATGGGGGCGACGGTGTATTTCCGCCAG
AAGATTTCGCCGCGGGAGCTCGCGGTGCGTACGTGCATGTTCAAACGCACGGTG
CGCGCATGGCAGTGGCAGACTGATCAACGCAGCTGGAAGCATCCGAAGCGCGCG
GGCACGCGTGTCCTCGACGCGTGGCCTCACATGCTGTCGGGTCGGTTCAAGACC
GAAAGCCACCGACCGACGCGCGAGCAATGCGCTACGCGGATCGCGTTCGACACG
AGCCGCGCGCGAGGCAAGGCCGACGTATTCGATCTTCCAGAGGAAGCCTATTGG
CTCGAGTCGTAGTGCTCGATATGGTAGAGCAACATGAATCCCGGGCTAAGTACAA
GAAGTAACCCGGCAACGAGTGAGATTGCGACGAATAAACGCTTCACCATGATCGC
GCTCCTGAGTTGGTTGAGGTGAATTGGAAAGTCGATTCCTGGGGGATCATTCCCG
GCAAGGCGCGCAATCCCCGCATTGTTCTCAAGATCGCAACGCGATTCGTCAGGCC
GATCTTCATGGGGTGTCTCGCTGGTAGTGATTCCGTCGTGGCCCGCGCATGTGCA
TGACGGCATCCGGGGAG
```

and determines the answers to the following questions:

1.) What is the length of the sequence?
2.) What are the fractions of each triplet (i.e. combinations of 3 nucleotides) in the sequence? You should create a dictionary that allows you to count of the triplets and print out a table of the frequencies of triplets.
3.) Find the genetic complement of the above sequence (i.e. A is replaced by T, G is replaced by C etc.)

You can write either one script that carries out these steps or a separate script for each task. Also, indicate the answers to these question in a separate text file.

2.) In the file `human.fa` you will find 13 human protein sequences in fasta format. In particular, in such a file each sequence entry starts with '>' telling you information about the following sequence that is formatted by strings of 60 characters. Currently each entry is annotated by its gene name.

In the file `protein-coding_gene.txt` you will find more information about human genes and proteins. In particular, the file contains the information about gene names, their synonyms and gene location.

Write a python code named `parse.py` that:

1. allows you to parse the information of the human.fa.

2. You need to get the information of synonyms and location by parsing the corresponding information from protein-coding_gene.txt.
3. As output, I want you to write a separate fasta file. In particular, the information of each entry should be formatted by writing > gene name|synonyms (separated by commas)|gene location.
4. Sequence information needs to be exactly like in human.fa (meaning sequence needs to be in blocks of 60 characters).

You need to write a python code to carry out this task. To make it fun, every sequence and the auxiliary information needs to be represented by a class object.

3.) Using the same file `dna.fasta,` write a Python program called `orf.py` that parses these sequences through their gene identifiers (2nd column) and answers the following questions:

- What is the longest sequence and what is the shortest sequence?
- In molecular biology, a reading frame is a way of dividing the DNA sequence of nucleotides into a set of consecutive, non-overlapping triplets (or codons). For instance, the three possible forward reading frames for the sequence AGGTGACAC are:

  AGG TGA CAC (frame 1)
  A GGT GAC AC (frame 2)
  AG GTG ACA C (frame 3)

  An open reading frame (ORF) is the part of a reading frame that encodes a protein that starts with a start codon (ATG), and ends with a stop codon (TAA, TAG or TGA). For instance, ATGAAATAG is an ORF of length 9.
  For each sequence in `dna.fasta` find the longest ORF in each sequence and indicate in which frame you found it (Hint: if you find a start codon, the ORF ends at the next stop codon you encounter. However, ORFs can be overlapping. In other words, if the sequence looks like this AAATGAUCAACATGAAAUGAUAATAGCGUTAAAAA what is the ORF we are looking for?).