Assignment

PF2

Assignment Background:

This short assignment contains three sections of questions that build upon and incorporate methods, techniques and concepts taught and demonstrated in the lesson materials. Each section will explore your understanding of strings, lists, tuples, and sets.

All questions aim to consider biological context. Timely submission will result in your assignment being returned within 14 days of the submission deadline. Marked Jupyter Notebook manuscripts will be returned with comments from your marker If you receive a pass or distinction, your marker will attach the model solutions for this assignment *via* the open pull request on your lesson's GitHub repository.

For the first two sections, **strings**, **lists**, and **tuples** you will be working with some unformatted sequencing data that you have received. Your task will be to re-format the data so that it will work with your analysis pipeline.

The data is given below:

```
seq1 = "SEQ001_ATCGATCGTAGCTAGCTAGCTA"
seq2 = "ATCG_CGTAGCTAG_CTAGCTA"

sequences = [
    "GCTAGTCGTAGCTCTAGCTAGCTAGCAAATAAGCTAGT",
    "TAGCTAGCTAGCTAGTCGTAGCTCAGCTAGCTA",
    "CTAGCAGCTATAGCTAGCTTCGTAGCAGCTAGCT",
    "ATCGATCGGCTATCGATCGATTCGTAGCTCGAT"
]
```

For the last section, **sets**, you will be working with a list of patient blood types for your clinical trial:

```
group_a_blood_types = ['A+', 'B+', 'O-', 'A+', 'AB+', 'O-', 'B+', 'A+',
'O+']
group_b_blood_types = ['O+', 'A-', 'B+', 'O+', 'A+', 'AB-', 'O-', 'A-']
```

Assignment Questions:

Strings

1. You have been given two sequences that should be one, `seq1` and `seq2`. Both sequences will need to cleaned and then joined together, write some Python code that achieves the following goals:
    - Remove the sequence ID from `seq1`, the characters before and including the _
    - `seq2` is missing some information, replace the underscores with `G`
    - Join the two sequences together, with `seq1` being first
2. Count how often `AG` appears in the string you created in question 1.

Lists

1. Part of your data analysis pipeline necessitates that the string be converted to a list:
    - Convert the sequence you generated above into a list
2. Demonstrate your understanding of list indexing:
    - Isolate and print the 5th item in the list
    - Isolate and print the 3rd to last item in the list
3. Count how many times "G" appears in the list you created in question one of this section.
4. Now that you have cleaned the sequence you want to add it to a list of sequences from the same experiment, write some Python code that achieves the following tasks:

- Add the cleaned/joined sequence from the **strings** section to the start of the list called `sequences`
- You only need the first 4 items in the list, create a slice that achieves this. Save this slice as a new variable
- You decide you want to remove the last item instead, achieve this through a list method. Save this as a new variable

Tuples

1. Convert the updated `sequences` (containing 5 items) into a Tuple:
2. Use indexing to access the 2nd sequence:
    - Save this as a new variable
    - Check if `TAGCTCT` appears in the 2nd sequence in the tuple:
        - Print the result using string formatting to highlight the answer (i.e. Does X appear in the sequence: ...)

Sets

1. You want to compare what blood types are unique or shared across group_a and group_b, the best way to do this is through sets:
    - Create sets called `unique_group_a` and `unique_group_b` from the given lists
    - Print the sets
2. Find all blood types present across both groups
3. Find blood types common to both groups
4. Find blood types that appear in only one group, not both (symmetric difference)

*HINT: See the Set operations sub-section of **Sets** for the various methods.*