Lab 1 - Python fundamentals

Purpose of this lab is to use fundamentals of python to transform raw data to representable format. More specifically, you will perform a step in decoding DNA data. This can be a common topic for a data analyst or data scientist in a medtech company. You do not need prior knowledge about DNA for this lab.

> 💡 **About DNA data:** DNA (or deoxyribonucleic acid) is the building block of life. They can be decoded into four letters: A, T, C and G, which represents adenine, thymine, cytosine and guanine. International scientists have been striving to analyze DNA data to improve our understanding of genetic diseases. This research is highly data-intensive as a single human cell can be decoded into 6 billions of DNA letters. Check out The Human Genome Project if you are interested to know more.

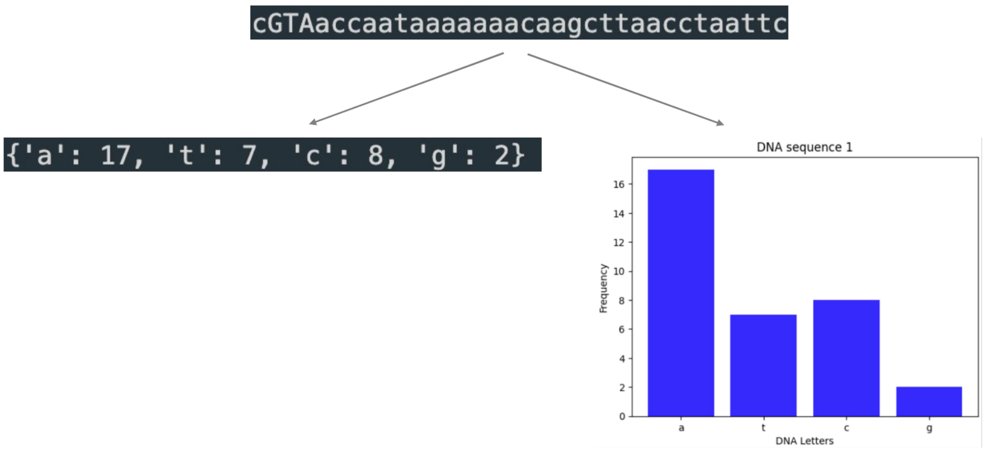
Tasks

Task 1

You are given several simplified sequences of DNA codes*. Your task is to count the number of different DNA letters in each sequence. In reality, this is an essential step for scientists to understand the genetic information from the human cell**. Following the instructions below:

- read the text file called *dna_raw.txt*
  - each sequene is composed of two lines of data: the first line beginning with > sign is the sequence ID, while the following line is the actual sequence
  - the actual sequence is not case-sensitive, which means that lower and upper case letters are treated the same
- for each sequence, create a dictionary to count the number of each DNA letter in that sequence
- for each sequence, graph the frequency of DNA letters for each sequence

Below is an example of a dictionary and graph from a sequence:



```
cGTAaccaataaaaaaacaagcttaacctaattc
```

```
{'a': 17, 't': 7, 'c': 8, 'g': 2}
```

Task 2 (bonus)

- there can be raw DNA data files with different number of sequences. Based on your solution code above, create a function that is able to take in new data files in similar format as *dna_raw.txt* with different number of sequences, and produce the same results above

- you have received also another DNA codes that is more complicated. The file is called *dna_raw_complicated.txt*. In this file, each DNA sequence can be composed of multiple lines of data in the text file. In this case, are you able to solve the task in the same way as before? If not, update your code to solve the same task with the new data

Grading

If you have taken ideas of codes from someone or found them online, it is **important** that you state the source and understand how the codes work. Write a comment next to these codes.

Criteria for G:

- solved task 1 correctly
- added relevant comments on the codes
- the codes are well-structured

Criteria for VG:

- solved both tasks 1 and 2 correctly

Submission

- create a public github repo when you start working with this lab
- the repo should show that you have made commits into it
- when you are done, send the repo URL to Learnpoint by the deadline

---

Note:

* The data are modified from the sample data here: https://github.com/mahmoudparsian/data-algorithms-with-spark/tree/master

** This task is much simplied compared to data processing of actual DNA data to fit the purpose of our course