Python Programming - Assessed Exercise No.1

Issued: Friday 22 October 2021
Due: Monday 1 November 2021 – 10am

Guidance on using Jupyter Notebook to write your script is at the end of the instructions.

Background

When DNA sequences are produced on a sequencer a measure of quality is required. This quality score, initially developed by the ***Phred*** program, is called the Q value and is assigned to each base as it is predicted.

Phred quality scores Q are defined as a property which is logarithmically related to the base-calling error probabilities P.

$$Q = -10 \log_{10} P$$

or

$$P = 10^{\frac{-Q}{10}}$$

For example, if ***Phred*** assigns a quality score of 30 to a base, the chances that this base is called incorrectly are 1 in 1000.

A sample file of sequence data, including Q scores, is available on Blackboard. The file is called **seq_sample.fastq**.

This file is a sample set of Illumina sequencing reads in fastq format. The reads are paired end reads, which means that each DNA fragment is sequenced from both ends. The fragment itself is of known length, possibly 2Kb, and each read is 101bp. The format for each entry is below and the lines have been truncated for clarity:

@sample_43/1
CTCGTTTAACGCAGACTCATCTAAACATAACCCTCTGAAAGAATACAA...
+
_bbecceegceegihiihchffehghhghhhibdgffdgff_egfhhhifihhiii`ffghiigbddggdedeeab...
@sample_43/2
CAAGGACCCTATTGTTAAATGCTCCTGTAAGCCATATGCAGGAATTTG....
+
bbbeeeeeggggebeghiiiihdhihaghihiihhihhghiagfhhhchhfhhhifhihgfhiiiiiifiiiiifgggf....

The first line begins with a "@" followed by the sequence identifier and an optional description. The "/1" signifies this is the first read from the pair. The second line is the sequence itself and the third is just a "+" that may optionally be followed by the sequence

identifier again. The fourth line is the quality scores for the read sequence. This is then repeated again for the second read from the read. Note that each set of paired end sequences covers eight lines. The quality scores use letters to overcome the problem of representing multi digit numbers and the actual scores are calculated from the ascii values of the characters used. For Illumina data the score is the ascii value minus 64, which means that "A" with an ascii value of 65 would represent a score of 1.

Tasks

Write a Python script in Jupyter Notebook that reads the **seq_sample.fastq** file and filters out any read pairs where at least one of the sequences has an average Q score below 30.

Your script should output 2 files. The first will contain the sequences where both reads have an average score above and including 30 and the second those sequences that have at least 1 read with an average score below 30.

Important: The output files are required to have the following names:

above30.fastq
below30.fastq

The output files should be in the same fastq format as the original.

Hint

You can get the ascii value for any character in python by using the ***ord*** command:

```
ch = "A"
val = ord(ch)
print val
```

This will print: 65

IMPORTANT:

Your script should not require or use any command line arguments.
It should open all files within the same directory as the script is run. Do not use relative paths to any files.
Do not change the name of the fastq input file.
The code must be commented.
Your code must run with Python3 and you cannot use any 3rd party libraries apart from those included with Anaconda Python/Jupyter Notebook.
The output files must be named as detailed above.
Marks will be deducted if any of the above are not followed.

What to Hand In

You must submit 1 Jupyter Notebook file to Dropbox on Blackboard.

You need to name your Jupyter Notebook file as:

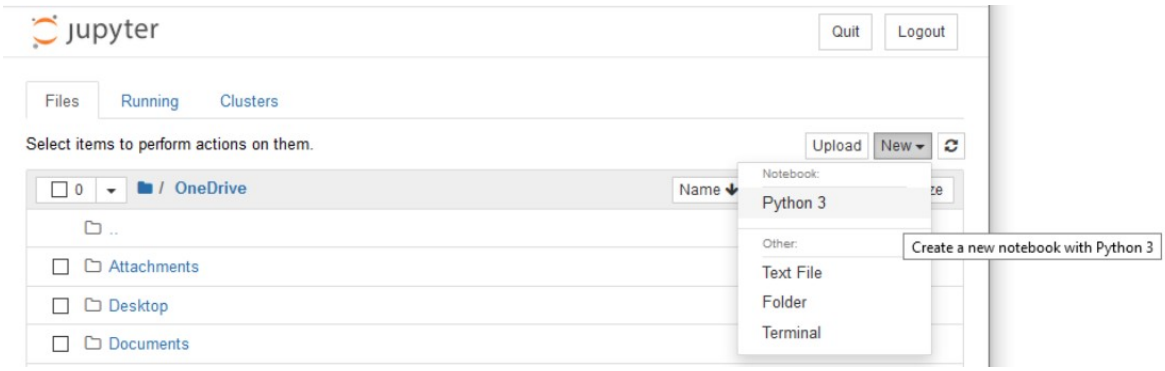
Surname_forename_degree_Python1.ipynb

The "degree" will be ABB or Bionf

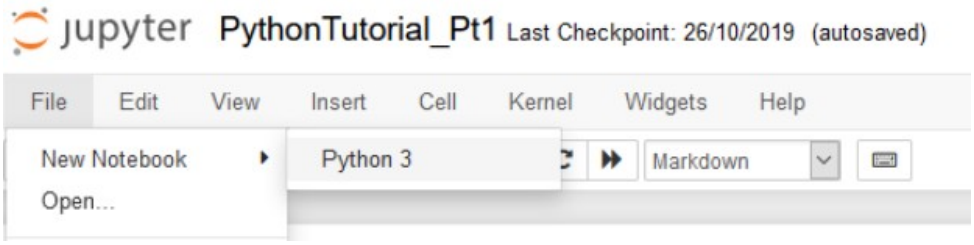
You do not need to submit your output files.

Using Jupyter Notebook to Write your Python Script

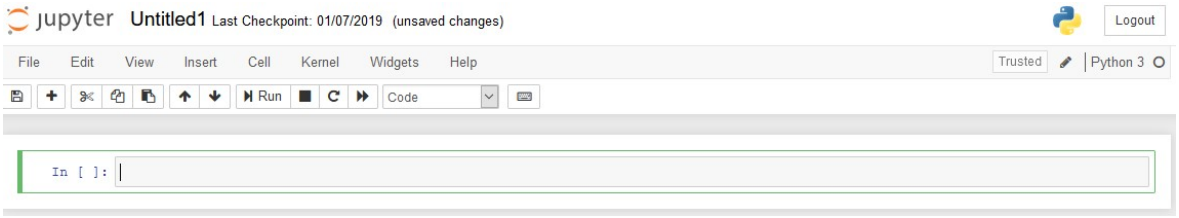
A Python script can be created from the Jupyter Notebook start page. Click "**New**" and select "**Python3**":



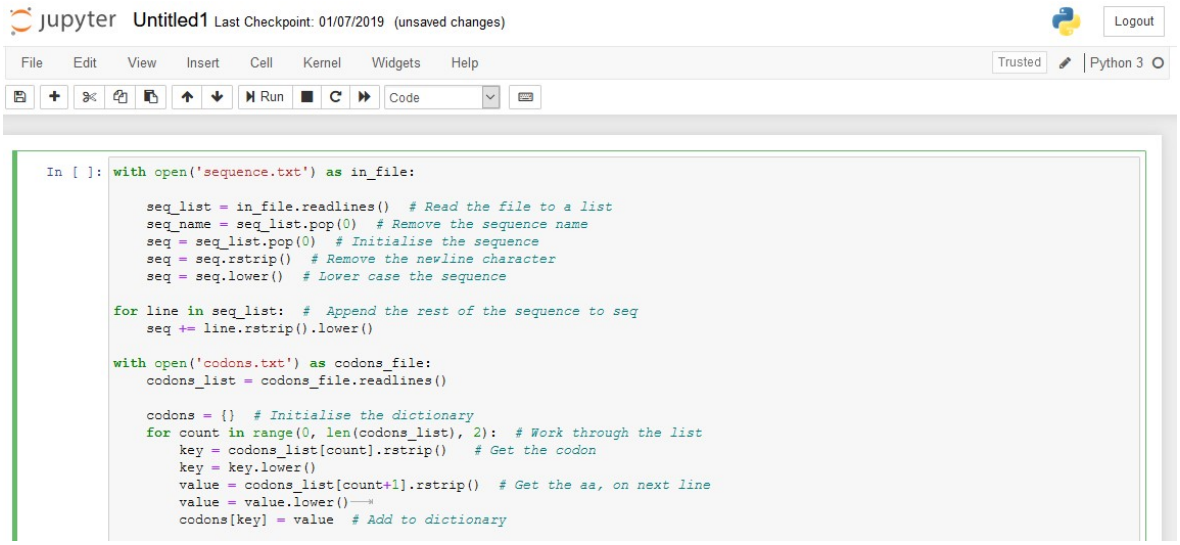
You can also create one from an existing Notebook by clicking "**File**" and selecting "**New Notebook**" and then "**Python 3**":



The new notebook will have a single cell into which you can write and test your code:



Write your code within this cell, which will expand automatically:



Save the Jupyter Notebook and upload to Dropbox on Blackboard.

Ensure you rename the file as required.