S13674
2017s16415
Hendalage DPB

BT-3172: Special Topics in Bioinformatics: Computing for Biologists
Practical 1: Introduction to computing for biologists

In this practical you will learn how to write algorithms to solve simple biological problems and implement them using Python. I recommend using PyCharm as your IDE for writing Python codes.

1) Calculating the length of a given DNA sequence.
   i) BRCA1 is an important tumor suppressor gene, which is crucial in DNA repair. Mutations in this gene are known to cause cancer, especially the breast cancer, in humans. As your first task, obtain the DNA sequence for the BRCA1 gene from the NCBI GenBank in FASTA format. Make sure you download the NCBI RefSeq gene sequence. Write the Gene ID of the obtained sequence. Write the RefSeq accession ID of the downloaded sequence.

      NG_005905

   ii) What is RefSeq and how RefSeq sequences are different from other GenBank nucleotide sequences?

   The Reference Sequence (**RefSeq**) database is an open access, annotated and curated collection of publicly available nucleotide sequences (DNA, RNA) and their protein products

   GenBank sequence records are owned by the original submitter and cannot be altered by a third party. **RefSeq** sequences are not part of the INSDC (the International Nucleotide Sequence Database Collaboration) but are derived from INSDC sequences to provide non-redundant curated data representing our current knowledge of known genes

   iii) Write the pseudocode for an algorithm to output the length of a given DNA sequence in FASTA format. In your first attempt, you have to use a for loop to count the length.

```
Import sequence and open it
Store it as a variable
Remove the header
Define a counter
Count all characters with the help of for loop
Get the output
```

   iv) Implement the above algorithm in Python. Save the code as "your_index_Q1_4code.py"
   Hint: use the open() function in Python to read the FASTA file content and save it in a variable.

v) Now, implement the same algorithm to count the length of the sequence, but this time use the len() fuction in Python. Save the code as “your_index_Q1_5code.py”

2) Calculating the nucleotide base counts of a given sequence

i) Write a pseudocode for an algorithm to calculate the nucleotide base counts of a given DNA sequence in FASTA format.

```
Import sequence file
Open the sequence in fasta format
Store as variable
Remove the fasta header
Define counters for each base
Make a for loop for identify different bases and count it
Print the output
```

ii) Implement the above algorithm in Python and save the code as ““your_index_Q2code.py”. Use the same BRCA1 gene you used in the previous exercise.