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

W.M. Ayesha Sanahari | 2017s16470 | S13722

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.

      Gene ID : 672
      RefSeq Accession ID : NG_005905.2

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

      RefSeq
      The Reference Sequence (RefSeq) collection provides a comprehensive, integrated, non-redundant, well-annotated set of sequences, including genomic DNA, transcripts, and proteins. RefSeq sequences form a foundation for medical, functional, and diversity studies. They provide a stable reference for genome annotation, gene identification and characterization, mutation and polymorphism analysis (especially RefSeqGene records), expression studies, and comparative analyses.

      RefSeq genomes are copies of selected assembled genomes available in GenBank. RefSeq transcript and protein records are generated by several processes including:

      Computation
          Eukaryotic Genome Annotation Pipeline
          Prokaryotic Genome Annotation Pipeline
      Manual curation
          Propagation from annotated genomes that are submitted to members of the International Nucleotide Sequence Database Collaboration (INSDC)

      The main features of the RefSeq collection include:

      - non-redundancy
      - explicitly linked nucleotide and protein sequences
      - updates to reflect current knowledge of sequence data and biology
      - data validation and format consistency
      - distinct accession series (all accessions include an underscore '_' character)
      - ongoing curation by NCBI staff and collaborators, with reviewed records indicated

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.

```
Input: DNA sequence in FASTA format
Output: sequence length

Open/ read sequence & store it in a variable
Remove blank lines
Remove the FASTA header
Define counter
For each letter/base in the sequence
    Increase the counter by 1
Return the counter/ length
```

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.

```
Input: DNA sequence in FASTA format
Output: No. of bases in each base type and total no. of bases

Open/read sequence & store it in a variable
Remove blank lines
Remove the FASTA header
Define counters for each type of base
For each letter/base in the sequence

        if A was found increase the no of A bases by 1
        else if T was found increase the no of T bases by 1
        else if G was found increase the no of G bases by 1
        else if C was found increase the no of C bases by 1
        else mention another letter found except A,T,G,C
Return the count of each base
Return the total no. of bases
```

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.