 Due: 7^{th} October

CMPSC 300
Bioinformatics
Fall 2019

Lab 5:
Working With Genbank Files

Homo sapiens cystic fibrosis transmembrane conductance regulator (CFTR) gene, exons 23, 24a, and 24

GenBank: M96936.1

FASTA Graphics

Go to:

```
LOCUS       HUMCFTMCR              1458 bp    DNA     linear   PRI 26-JUL-2016
DEFINITION  Homo sapiens cystic fibrosis transmembrane conductance regulator
            (CFTR) gene, exons 23, 24a, and 24.
ACCESSION   M96936
VERSION     M96936.1
KEYWORDS    alternative splicing; cystic fibrosis; cystic fibrosis
            transmembrane conductance regulator; transmembrane conductance
            regulator.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 1458)
  AUTHORS   Yoshimura,K., Chu,C.S. and Crystal,R.G.
  TITLE     Alternative splicing of intron 23 of the human cystic fibrosis
            transmembrane conductance regulator gene resulting in a novel exon
            and transcript coding for a shortened intracytoplasmic C terminus
  JOURNAL   J. Biol. Chem. 268 (1), 686-690 (1993)
   PUBMED   7678008
FEATURES             Location/Qualifiers
     source          1..1458
                     /organism="Homo sapiens"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:9606"
```

Figure 1: A Genbank record comes from PubMed and contains contains diverse information for each known DNA sequence.

GitHub starter link

https://classroom.github.com/a/aDNdN-Ed

To use this link, please follow the steps below.

- Click on the link and accept the assignment.
- Once the importing task has completed, click on the created assignment link which will take you to your newly created GitHub repository for this lab.
- Clone this repository (bearing your name) and work on the practical locally.
- As you are working on your practical, you are to commit and push regularly. You can use the following commands to add a single file, you must be in the directory where the file is located (or add the path to the file in the command):
  - `git add -A`
  - `git commit -m ‘‘Your notes about commit here’’`

- `git push`

Alternatively, you can use the following commands to add multiple files from your repository:

- `git commit <`*nameOfFile*`> -m ‘‘Your notes about commit here’’`
- `git push`

Be sure to read the `README.md` file in the GitHub Classroom repository for instructions on how to complete your first assignment.

Objectives

- To apply information from Genbank record files for analysis.
- To generate Python3 code to manipulate genetic sequences in a variety of ways and make basic comparisons between sequences.
- To modify your existing code (if you choose) from lab 3 to be able to load Genbank files to display information about transcription and translation, and to conduct a comparative analysis.

```
ORIGIN
        1 acagtaattc tctgtgaaca caggatagaa gcaatgctgg aatgccaaca atttttggtg
       61 agtctttata actttactta agatctcatt gcccttgtaa ttcttgataa caatctcaca
      121 tgtgatagtt cctgcaaatt gcaacaatgt acaagttctt ttcaaaaata tgtatcatac
      181 agccatccag ctttactcaa aatagctgca caagtttttc actttgatct gagccatgtg
      241 gtgaggttga aatatagtaa atctaaaatg gcagcatatt actaagttat gtttataaat
      301 aggatatata tacttttga gccctttatt tggggaccaa gtcatacaaa atactctact
      361 gtttaagatt ttaaaaaagg tccctgtgat tctttcaata actaaatgtc ccatggatgt
      421 ggtctgggac aggcctagtt gtcttacagt ctgatttatg gtattaatga caaagttgag
      481 aggcacattt catttttcta gccatgattt gggttcaggt agtacctttc tcaaccacct
```

Figure 2: Each Genbank record has the DNA sequence data that can be placed into a `FASTA` file.

Developing Your Own Tool

A Single tool for DNA mutation detection that loads data from Genbank files

Comparisons between genetic material are extremely common in Bioinformatics. As discussed in class, the presence of alternative SNPs between samples of DNA, may help researchers to conclude whether a disorder can be expected from the organisms. Furthermore, mutations between code samples may also also serve to help researchers determine and potentially explain reasons for genetic abnormalities which may appear in protein formation, structure and, therefore, function.

In this part of the lab you will develop a Python3 program that can manipulate sequences, transcribe and translate them and find mutations or general differences in sequences. Your working code that you are are to submit will be a modified version of the Python3 code (file: `src/mutDetect_i.py`) that you submitted for lab3's deliverable.

Required

Please follow the below steps to create your Python3 program.

1. Go to PubMed and download a single Genbank file of your choosing for a DNA record.
2. Make a copy of the downloaded file and rename the copy's filename to include the word, `mutation`. You are to manually edit this mutation file to include at least three (3) different types of mutations to the DNA (*note: we discussed all these three mutations types in class*).
3. Create a new Python3 program called: `src/mutDetect_ii.py`. You are welcome to use your solution from lab3's deliverable to begin this program however you will need to add the below functions and integrate their functionalities into you code.
4. Functions to write;
   (a) **A function to load Genbank files**: The program must implement a function to load two Genbank files *from steps 1 and 2, discussed above* and to isolate the DNA sequence material in each file using the BioPython library. You are free to chose how these files are loaded – by command line inputs or by simply asking the user to enter filenames using an `input()` function.
   (b) **A function to save FASTA files**: Your program must have a function to save a `FASTA` file from each Genbank that is processed.
5. Once the DNA from each file has been obtained by parsing, your program is to transcribe and translate both DNA sequences to produce protein sequences (the function of the tool from the deliverable of lab3).
6. After translation, your tool is to complete a comparison of the sequences, and to print out any differences found between both pairs of DNA, RNA and Protein.

Sample Output

Your program should have an output that resembles the following;

```
bioinformatics4Life$ ./mutDetect_ii.py <AnyKeyToStart>

 Welcome to mutDetect_ii!
 A program to compare DNA, make protein and compare protein sequences.

__Getting a Genbank sequence__
 Enter the GenBank file name :wild.gb
 Genbank record: M96936.1

__Writing a FASTA file__
 Genbank record: M96936.1

__Getting a Genbank sequence__
 Enter the GenBank file name :mut.gb
```

```
 Genbank record: M96936.1

__Writing a FASTA file__
 Genbank record: M96936.1
 + Length of first sequence  : 1458
 + Length of second sequence : 1458

__Comparing sequences__
 + Seq char not the same at pos:  1
 First seq char   :  C
 Second  seq char :  T
 + Seq char not the same at pos:  11
 First seq char   :  C
 Second  seq char :  G
 + Sequences are same length:  True

  __Translation__
 + Original DNA       : ACAGTAATTCTCTGTGAACACAGGATAGAAGCAATGCTG ... , length is : 1458
 + PROTEIN from RNA   : TVILCEHRIEAMLECQQFLVSLYNFT*DLIALVILDNNLTCDSSCKLQQCTSSFQK
 + protein1 sequence  : TVILCEHRIEAMLECQQFLVSLYNFT*DLIALVILDNNLTCDSSCKLQQCTSS ...

 __Translation__
 + Original DNA       : ATAGTAATTCTGTGTGAACACAGGATAGAAGCAATGCTGGAATGCCA, length is : 1458
 + PROTEIN from RNA   : IVILCEHRIEAMLECQQFLVSLYNFT*DLIALVILDNNLTCDSSCKLQQCTSSFQK
 + protein2 sequence  : IVILCEHRIEAMLECQQFLVSLYNFT*DLIALVILDNNLTCDSSCKLQQCTSS ...

 __Comparing sequences__
 + Seq char not the same at pos:  0
 First seq char   :  T
 Second  seq char :  I
```

Ethical Consideration: Questions in Blue

Please read the National Public Radio article, *U.S. Justice Department Charges 35 People In Fraudulent Genetic Testing Scheme* (link `https://www.npr.org/sections/health-shots/2019/09/27/765230011/u-s-justice-department-charges-35-people-in-fraudulent-genetic-testing-scheme`). You are to respond to the below *Questions in Blue*, shown below, which are also found in the Markdown file, `ethical/reflections.md`, in your repository.

1. What is the business incentive for the company (or the article) that prepared cancer-risk assessments for patients?
2. Explain what practices were employed by the company?
3. Which population did the company mentioned in the article target? In your opinion, why was this particular population chosen by the company?
4. How were medical histories collected by doctors of the company? Comment on the ethical or unethical implications of the company's practices.

5. According to the article, how was medical data collected for the cancer-risk assessment? In your opinion, what types of problems do you foresee happening as a result of the way that data was handled?

6. Describe the technique used by the company in their feedback to patients. Comment on the ethical or unethical practices of their feedback method.

Required Deliverables

- Your completed activity should be saved as `src/mutDetect_ii.py` in the repository that you will push to GitHub.
- Please include your Genbank files (the one from the your download and the one that you edited to add the three (3) types of mutations.
- Ethical reflections: Please respond to the reflection questions in the Markdown file, `ethics/reflections.md` concerning the NPR article, *U.S. Justice Department Charges 35 People In Fraudulent Genetic Testing Scheme*.

Please see the instructor if you have questions about the assignment submission.