1327380
Huiming Zhou

CIS 6060 — Assignment #1

Winter 2025

Due: *Fri., Jan. 31, 2025 @ 23:59*

Total Points: 50

Please submit your assignment solutions as **one zip** file (named as YOUR/UoG/ID_a1.zip, e.g. 1234567_a1.zip) to Dropbox under Assignment 1 before the due date.

Exercise 1 Transcription, Translation (10 pts)

The following is a partial DNA coding segment from *Homo sapiens* CD4 gene that encodes a membrane glycoprotein of T lymphocytes.

5'ACCGGGGAGTCCCTTTTAGGCACTTGCTTCTGG3'

1. Produce the corresponding RNA fragment.
2. Produce the partial primary protein sequence (one letter codes for amino acids) using reading frame 3. Note: Reading frame 3 means the transcription starts at the 3rd nucleotide base. (You can manually interpret the sequence, or use online tools you can find to do it.))

Submission

Submit your answers in a PDF file.

Choose any two from the following three exercises (each is 20 points) to complete.

Exercise 2 Contemporary Issues Summary (20 pts)

Students are more motivated to learn things that they believe are worth learning. In this exercise, you will look for connections between course material ("bioinformatics, or computational/mathematical biology") and recent events or developments that you find via online news sites, printed news sources, or broadcast media. You will write an summary on how bioinformatics apply to these current affairs (in plain English, **one page, single-spaced, font size 11** is good). Your summary should include an overview of the event(s), an introduction of the bioinformatics tools and how they've been applied, and the significance or new knowledge the bioinformatics have brought.

Tips: you can try to associate these areas with the selection of reflective journal review, or help with selecting the topic of your project.

Below is a story about the 2011 German outbreak, which will give you some directions to look for a contemporary issues, but please feel free to connect to any events (not necessarily to be very recent).

"The 2011 Germany E. coli O104:H4 outbreak is one of the first uses of genome sequencing to study the dynamics of a food-borne outbreak and provides further evidence that genomic tools can be used to investigate future outbreaks and provide greater insight into the emergence and spread of infectious diseases. As an example to show here, the bioinformatics applied is the genome sequencing tools (in your summary, you may need to introduce such sequencing tool with more information). It represented an early example of epidemiologists collaborating with computational biologists on four continents to stop an outbreak. They released bacterial DNA sequencing data from one of a patient, which elicited a burst of analyses carried out by computational biologists on four continents. For more information on the computational methods and data analysis, you can go to their project page on GitHub: `https://github.com/ehec-outbreak-crowdsourced/BGI-data-analysis/wiki`. "

Submission

Submit your answers in the same PDF file you had for exercise 1.

Exercise 3 (20 pts)

A FASTA file consists of one line of header, the so-called, "definition line" (which starts with the > symbol), followed by lines consisting only of sequence data (usually 60 characters per sequence except perhaps the last one). A multi-FASTA contains information regarding multiple sequences and it starts with the header line for one sequence followed by the sequence across multiple lines, followed by the header for the second sequence, followed by the second sequence, and so on. The assignment folder contains a file `multiprotein.fasta` containing four proteins.

You should write a python script which does the following: it should ask the user for the name of an input FASTA file. Let's say the filename they enter is called '`stuff.fasta`'. The script should then read the input file. It should then display to the screen information regarding each sequence. For each, it should display the header line, the first 10 characters followed by the the number of amino acids in each protein. This will serve as a useful tool to summarize the contents of a multi-FASTA file.

Submission

1. Submit the source code of this question as `asn1e3.py`. Make sure to include your name and student number as comments in the file. You can use Biopython libraries when available, or just write your own code completing the task.

2. Submit the documentation as `READMEasn1e3.txt`. It should include information on how to run your code, any dependency or libraries needed, etc..

3. Submit the screenshot of you running the code and the output in the PDF file you had before.

Guidelines

Sample output of the first two proteins in the given file `multiprotein.fasta`, so you can check your code.

```
Path to FASTA file:
```

(your code should ask for an input of the multi-fasta file, then followed by the output)

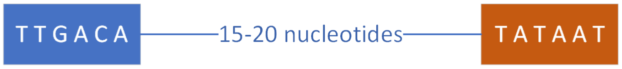
```
>1433G_HUMAN (P61981) 14-3-3 protein gamma (Protein kinase C inhibitor protein
1) (KCIP-1) [Homo sapiens]  VDREQLVQKA : 246
>ATP8_RAT (P11608) ATP synthase protein 8 (EC 3.6.3.14) (ATPase subunit 8)
(A6L) (Chargerin II) [Rattus norvegicus]  MPQLDTSTWF : 67
```

Exercise 4 (20 pts)

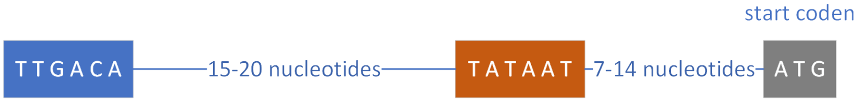
We want to find potential occurrences of promoter sequences in a genome (please read more information about promoters on wikipedia: http://en.wikipedia.org/wiki/Promoter_(genetics)).

In particular, we are searching for the common pattern that occurs, the sequence TTGACA upstream (before the beginning) at around -35 from the start of a gene and TATAAT at about -10. More precisely, we need to write a script which should ask the user for the name of a FASTA file (same function as you just did for Exercise 3), and then scan the file and output the following two sequences to the screen, if they exist:

1. the first sequence (q1) which starts with TTGACA and ends with TATAAT with anywhere between 15 and 20 nucleotides between TTGACA and TATAAT.



2. the second sequence (q2) which starts with TTGACA, followed by between 15 and 20 nucleotides, followed by TATAAT, followed by between 7 and 14 nucleotides, followed by the start codon, ATG. Everything but the start codon should be output to the screen for 2).



Notice that the two sequences above might not come from the same section. You can assume that (excluding the definition line), all lines but the last are 60 characters long (the last has less than or equal to 60).

You do not need to concern yourself with where the start of the gene is (the -35 and -10 numbers above). We are only concerned with the distance between the two sequences here.

Submission

1. Submit the source code of this question as `asn1e4.py`. Make sure to include your name and student number as comments in the file. You can use Biopython libraries when available, or just write your own code completing the task.

2. Submit the documentation as `READMEasn1e4.txt`. It should include information on how to run your code, any dependency or libraries needed, etc..

3. Submit the screenshot of you running the code and the output in the PDF file you had before.

Guidelines

Some example genomes are provided along with the assignment, which can be used to test your script.

Sample output of `BacGen.fasta` and `BacGenome.fasta` files are listed below for you to check your code.

BacGen.fasta

```
>gi|56160984|gb|CP000002.2| Bacillus licheniformis ATCC 14580, complete genome
q1(Potential promoter):  TTGACAGGCTTGTAGATACTCTATATAAT
q2 Sequence:  not found.
```

BacGenome.fasta

```
>gi|56160984|gb|CP000002.2| Bacillus licheniformis ATCC 14580, complete genome
q1(Potential promoter):  TTGACAGGCTTGTAGATACTCTATATAAT
q2 Sequence:  TTGACACTTAATTTTTTCTTTATGTATAATTAAACAAATG
```

Exercise 1 Transcription, Translation (10 pts)

The following is a partial DNA coding segment from *Homo sapiens* CD4 gene that encodes a membrane glycoprotein of T lymphocytes.

5'ACCGGGGAGTCCCTTTTAGGCACTTGCTTCTGG3'

1. Produce the corresponding RNA fragment.
2. Produce the partial primary protein sequence (one letter codes for amino acids) using reading frame 3. Note: Reading frame 3 means the transcription starts at the 3rd nucleotide base. (You can manually interpret the sequence, or use online tools you can find to do it.))

Submission

Submit your answers in a PDF file.

1.

5' ACC GGG GAG TCCCTT TTAGGC ACT TGC TTC TGG 3'

↓

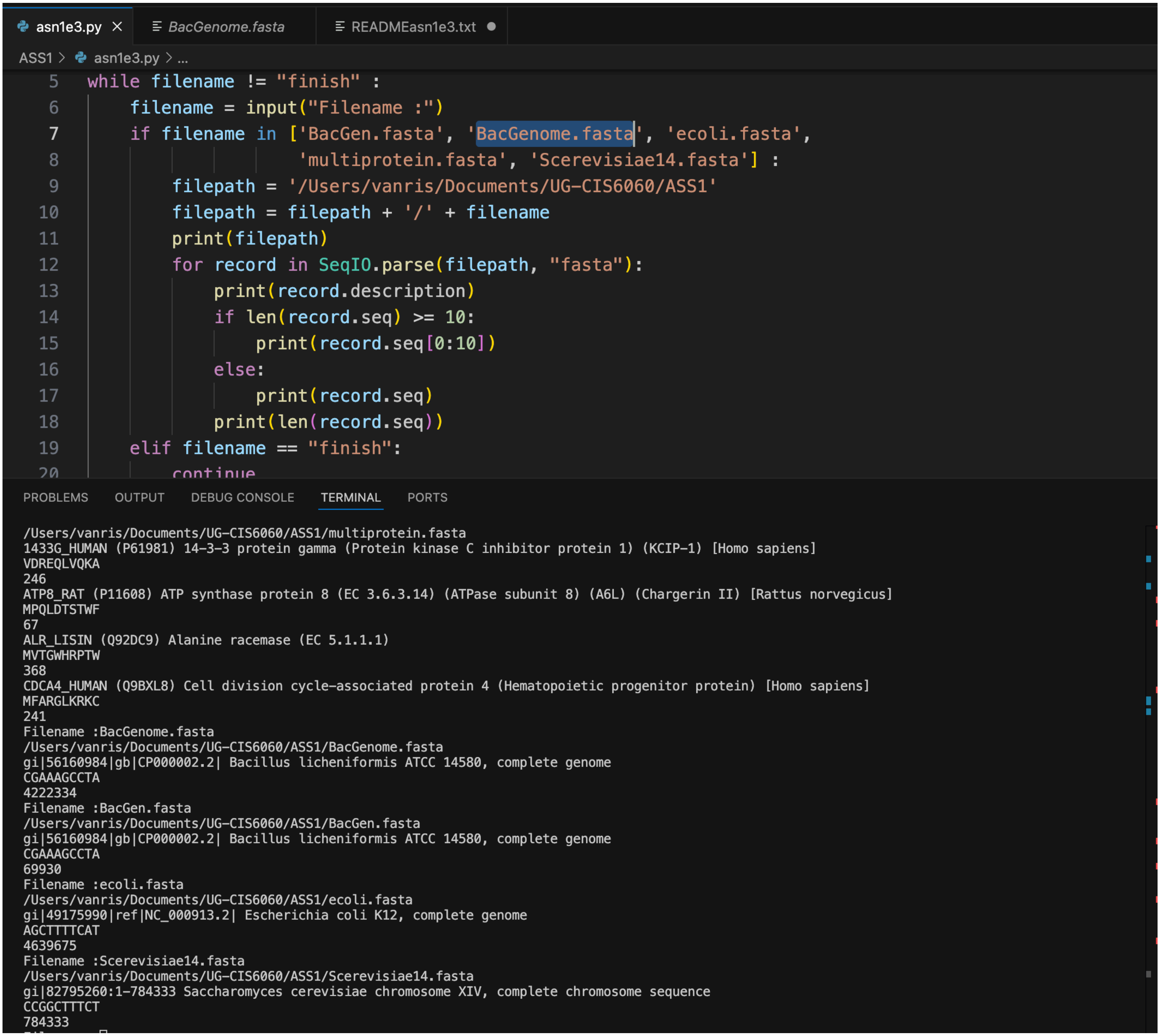
RNA

3' UGG CCC CUC AGG GAA AAU CG UGA ACG AAG ACC 5'

2.

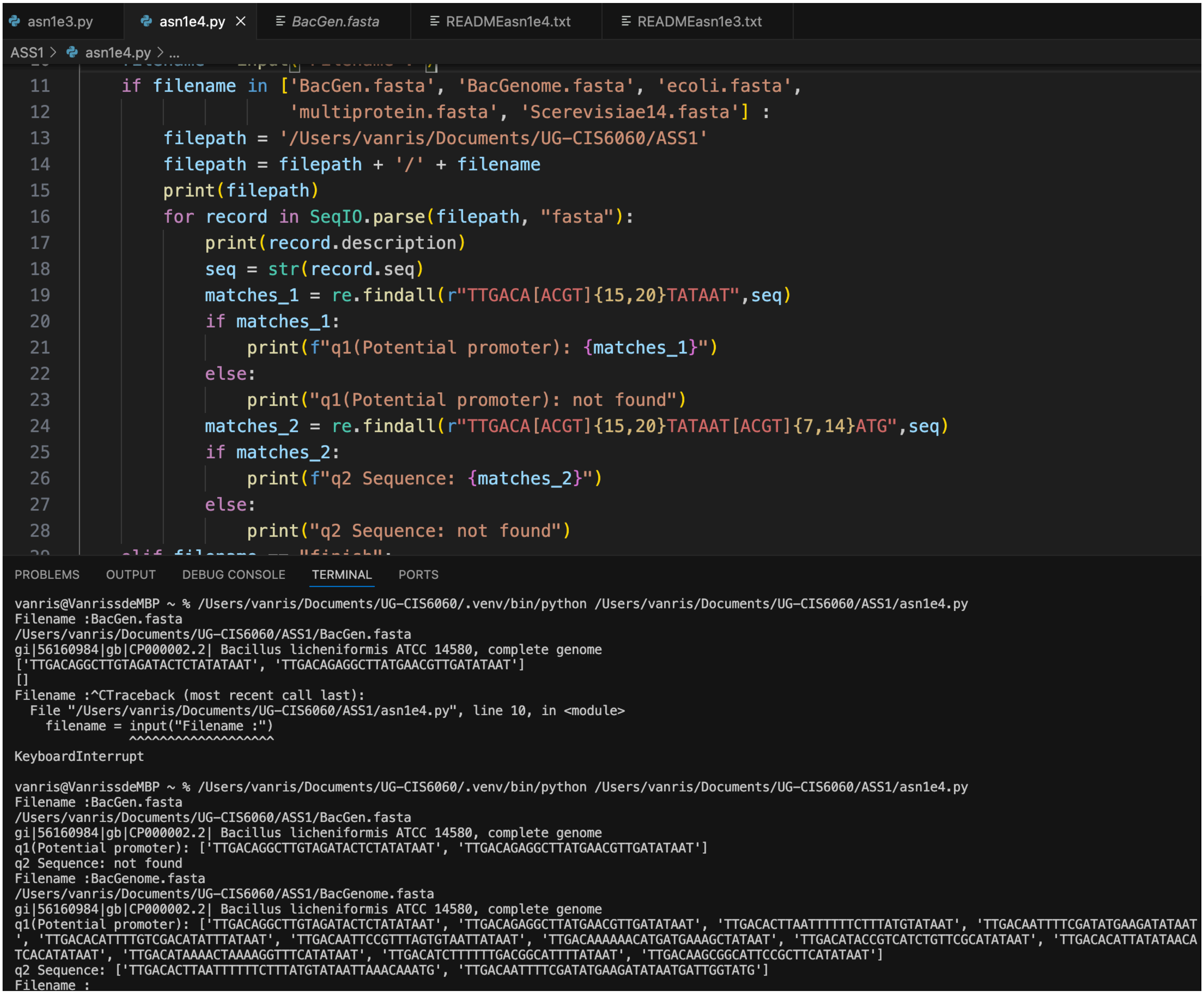
5' UG GCC CCU CAG GGA AAA UCC GUG AAC GAA GAC C 3'

A P Q G K S V N Z D



```python
while filename != "finish" :
    filename = input("Filename :")
    if filename in ['BacGen.fasta', 'BacGenome.fasta', 'ecoli.fasta',
                    'multiprotein.fasta', 'Scerevisiae14.fasta'] :
        filepath = '/Users/vanris/Documents/UG-CIS6060/ASS1'
        filepath = filepath + '/' + filename
        print(filepath)
        for record in SeqIO.parse(filepath, "fasta"):
            print(record.description)
            if len(record.seq) >= 10:
                print(record.seq[0:10])
            else:
                print(record.seq)
            print(len(record.seq))
    elif filename == "finish":
        continue
```

```
/Users/vanris/Documents/UG-CIS6060/ASS1/multiprotein.fasta
1433G_HUMAN (P61981) 14-3-3 protein gamma (Protein kinase C inhibitor protein 1) (KCIP-1) [Homo sapiens]
VDREQLVQKA
246
ATP8_RAT (P11608) ATP synthase protein 8 (EC 3.6.3.14) (ATPase subunit 8) (A6L) (Chargerin II) [Rattus norvegicus]
MPQLDTSTWF
67
ALR_LISIN (Q92DC9) Alanine racemase (EC 5.1.1.1)
MVTGWHRPTW
368
CDCA4_HUMAN (Q9BXL8) Cell division cycle-associated protein 4 (Hematopoietic progenitor protein) [Homo sapiens]
MFARGLKRKC
241
Filename :BacGenome.fasta
/Users/vanris/Documents/UG-CIS6060/ASS1/BacGenome.fasta
gi|56160984|gb|CP000002.2| Bacillus licheniformis ATCC 14580, complete genome
CGAAAGCCTA
4222334
Filename :BacGen.fasta
/Users/vanris/Documents/UG-CIS6060/ASS1/BacGen.fasta
gi|56160984|gb|CP000002.2| Bacillus licheniformis ATCC 14580, complete genome
CGAAAGCCTA
69930
Filename :ecoli.fasta
/Users/vanris/Documents/UG-CIS6060/ASS1/ecoli.fasta
gi|49175990|ref|NC_000913.2| Escherichia coli K12, complete genome
AGCTTTTCAT
4639675
Filename :Scerevisiae14.fasta
/Users/vanris/Documents/UG-CIS6060/ASS1/Scerevisiae14.fasta
gi|82795260:1-784333 Saccharomyces cerevisiae chromosome XIV, complete chromosome sequence
CCGGCTTTCT
784333
```



```python
    filename = input("Filename :")
    if filename in ['BacGen.fasta', 'BacGenome.fasta', 'ecoli.fasta',
                    'multiprotein.fasta', 'Scerevisiae14.fasta'] :
        filepath = '/Users/vanris/Documents/UG-CIS6060/ASS1'
        filepath = filepath + '/' + filename
        print(filepath)
        for record in SeqIO.parse(filepath, "fasta"):
            print(record.description)
            seq = str(record.seq)
            matches_1 = re.findall(r"TTGACA[ACGT]{15,20}TATAAT",seq)
            if matches_1:
                print(f"q1(Potential promoter): {matches_1}")
            else:
                print("q1(Potential promoter): not found")
            matches_2 = re.findall(r"TTGACA[ACGT]{15,20}TATAAT[ACGT]{7,14}ATG",seq)
            if matches_2:
                print(f"q2 Sequence: {matches_2}")
            else:
                print("q2 Sequence: not found")
    elif filename == "finish":
```

```
vanris@VanrissdeMBP ~ % /Users/vanris/Documents/UG-CIS6060/.venv/bin/python /Users/vanris/Documents/UG-CIS6060/ASS1/asn1e4.py
Filename :BacGen.fasta
/Users/vanris/Documents/UG-CIS6060/ASS1/BacGen.fasta
gi|56160984|gb|CP000002.2| Bacillus licheniformis ATCC 14580, complete genome
['TTGACAGGCTTGTAGATACTCTATATAAT', 'TTGACAGAGGCTTATGAACGTTGATATAAT']
[]
Filename :^CTraceback (most recent call last):
  File "/Users/vanris/Documents/UG-CIS6060/ASS1/asn1e4.py", line 10, in <module>
    filename = input("Filename :")
               ^^^^^^^^^^^^^^^^^^^

KeyboardInterrupt

vanris@VanrissdeMBP ~ % /Users/vanris/Documents/UG-CIS6060/.venv/bin/python /Users/vanris/Documents/UG-CIS6060/ASS1/asn1e4.py
Filename :BacGen.fasta
/Users/vanris/Documents/UG-CIS6060/ASS1/BacGen.fasta
gi|56160984|gb|CP000002.2| Bacillus licheniformis ATCC 14580, complete genome
q1(Potential promoter): ['TTGACAGGCTTGTAGATACTCTATATAAT', 'TTGACAGAGGCTTATGAACGTTGATATAAT']
q2 Sequence: not found
Filename :BacGenome.fasta
/Users/vanris/Documents/UG-CIS6060/ASS1/BacGenome.fasta
gi|56160984|gb|CP000002.2| Bacillus licheniformis ATCC 14580, complete genome
q1(Potential promoter): ['TTGACAGGCTTGTAGATACTCTATATAAT', 'TTGACAGAGGCTTATGAACGTTGATATAAT', 'TTGACACTTAATTTTTTCTTTATGTATAAT', 'TTGACAATTTTCGATATGAAGATATAAT
', 'TTGACACATTTTGTCGACATATTTATAAT', 'TTGACAATTCCGTTTAGTGTAATTATAAT', 'TTGACAAAAAAACATGATGAAAGCTATAAT', 'TTGACATACCGTCATCTGTTCGCATATAAT', 'TTGACACATTATATAACA
TCACATATAAT', 'TTGACATAAAACTAAAAGGTTTCATATAAT', 'TTGACATCTTTTTTGACGGCATTTTATAAT', 'TTGACAAGCGGCATTCCGCTTCATATAAT']
q2 Sequence: ['TTGACACTTAATTTTTTCTTTATGTATAATTAAACAAATG', 'TTGACAATTTTCGATATGAAGATATAATGATTGGTATG']
Filename :
```