

BINF6410 Fall 2024

Assignment 2

Submit this assignment to the Courselink Dropbox that will be created. Please name the file with your username and the assignment number, for example my file would be named lewis_a2.py

Write a Python3 program that does the following:

1. Reads in a provided FASTA file containing a nucleotide sequence. **Please do not hardcode the name of the file.** The name should be presented to the program as a unix command line argument.
2. Reads in a provided text file containing a series of restriction enzymes. The file contains one enzyme per line, listing first the enzyme name, followed by a semicolon (;) and then the sequence recognized by the enzyme, with a ^ character indicating the cleavage site. *E.g: EcoRI;G^AATTC.* Like the FASTA file, the filename should be a command line argument.
3. Uses each of the restriction enzymes to find where it cuts the nucleotide sequence.
4. Prints out a report with the following information (See the end for the formatting of the report):
 - a. the names of the files used
 - b. the header information from the FASTA file
 - c. the length of the nucleotide sequence
 - d. for each enzyme:
 - i. the name of the enzyme and its sequence
 - ii. the number of cutting sites found (this could be zero)
 - iii. if cutting sites are found:
 1. how many cutting sites and how many fragments
 2. the sequences of all of the fragments that would result if the enzyme were applied. Display these in lines of 60 bases, with each line separated into groups of 10 bases. Display the sequence position the start of each line, and keep a running count between fragments
 - iv. if no cutting sites are found, print a clear understandable message informing the user of this.

Your program should work with any restriction enzyme and any properly formatted FASTA file. You can assume that the FASTA file will contain only one sequence. A small test-file will be provided to you along with a list of enzymes.

Finally, you may import code from a module to process the sequence from the input FASTA file. Otherwise, the project should be done with code from commands that you learned or will learn in lecture. Of course, you are welcome to research these commands and their use.

Your output should look something like this:

Restriction enzyme analysis of sequence from file testseq.fas.
Cutting with enzymes found in file enzymes.txt.

Sequence name: TESTSEQ
Sequence is 107 bases long.

There are 2 cutting sites for EcoRI, cutting at G^AATTC There are 3 fragments:

Length- 27
1 ATTATAAAAT TAAAATTATA TCCAATG
Length- 23
28 AATTCAATTA AATTAAATTA AAG
Length- 57
51 AATTCAATAA TATACCCCGG GGGGATCCAA TTAAAAGCTA AAAAAAAAAA AAAAAA

There are 1 cutting sites for BamHI, cutting at G^GATCC There are 2 fragments:

Length- 73
1 ATTATAAAAT TAAAATTATA TCCAATGAAT TCAATTAAAT TAAATTAAAG
AATTCAATAA
61 TATACCCCGG GGG
Length- 34
74 GATCCAATTA AAAGCTAAAA AAAAAAAAAA AAAA

There are no sites for PstI.

There are 1 cutting sites for AluI, cutting at AG^CT There are 2 fragments:

Length- 87
1 ATTATAAAAT TAAAATTATA TCCAATGAAT TCAATTAAAT TAAATTAAAG
AATTCAATAA
61 TATACCCCGG GGGGATCCAA TTAAAAG
Length- 20
88 CTAAAAAAAA AAAAAAAAAA

Your work will be evaluated on four axes:

1. **Scope:** The extent to which your code implements the features outlined in the specification (8 marks).
2. **Correctness:** The extent to which your code is consistent with the specification and is free of bugs (8 marks).
3. **Design:** The extent to which your code is well written (i.e. clear, efficient, elegant, logical) (4 marks).
4. **Style:** The extent to which your code is readable (i.e. commented, containing aptly named variables, etc) (4 marks).

```
cd "/Users/paniztayebi/Documents/ptayebi-Assignment2"
```

```
/usr/local/bin/python3 Assignment2.py nucleotide_seq.fasta restriction_enzymes.txt >
terminal_output.txt
```

```
Panizs-MacBook-Air:Python-6410 paniztayebi$ /usr/local/bin/python3 Assignment2.py
nucleotide_seq.fasta restriction_enzymes.txt
```

Restriction enzyme analysis of sequence from file nucleotide_seq.fasta.
Cutting with enzymes found in file restriction_enzymes.txt.

```
-----
Sequence name: TestSeq_1
Sequence is 6000 bases long.
```

```
-----
There are 3 cutting sites for EcoRI, cutting at G%AATTC
There are 4 fragments:
Length- 3747
Length- 584
Length- 695
Length- 974
```

```
-----
There are 0 cutting sites for BamHI, cutting at G%GATCC
There are no sites for BamHI.
```

There are 5 cutting sites for PstI, cutting at CTGCA%G

There are 6 fragments:

Length- 892

Length- 890

Length- 1345

Length- 2245

Length- 29

Length- 599

There are 40 cutting sites for AluI, cutting at AG%CT

There are 41 fragments:

Length- 234

Length- 32

Length- 687

Length- 307

Length- 7

Length- 275

Length- 273

Length- 67

Length- 29

Length- 88

Length- 27

Length- 137

Length- 68

Length- 51

Length- 99

Length- 368

Length- 25

Length- 295

Length- 19

Length- 29

Length- 6

Length- 158

Length- 82

Length- 34

Length- 42

Length- 11

Length- 221

Length- 5

Length- 45

Length- 54

Length- 256

Length- 51

Length- 118

Length- 76
Length- 139
Length- 8
Length- 430
Length- 101
Length- 296
Length- 716
Length- 34