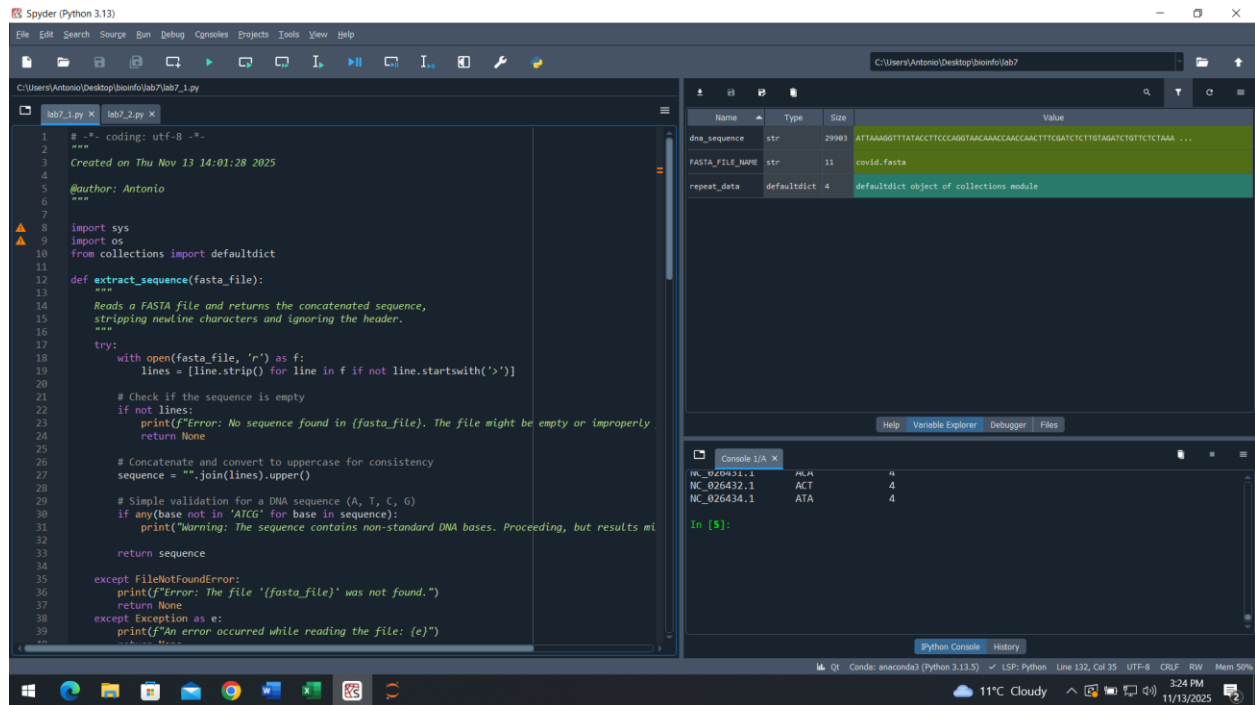


LABORATORY REPORT #7

GEORGESCU Mihai-Antonio, 1242EA
Bioinformatics, 4th year 1st semester, 2025-2026

lab7_1.py



The screenshot displays the Spyder Python IDE interface. The main editor window shows a Python script named `lab7_1.py` with the following content:

```
1 # -*- coding: utf-8 -*-
2 """
3 Created on Thu Nov 13 14:01:28 2025
4
5 @author: Antonio
6 """
7
8 import sys
9 import os
10 from collections import defaultdict
11
12 def extract_sequence(fasta_file):
13     """
14     Reads a FASTA file and returns the concatenated sequence,
15     stripping newline characters and ignoring the header.
16     """
17     try:
18         with open(fasta_file, 'r') as f:
19             lines = [line.strip() for line in f if not line.startswith('>')]
20
21         # Check if the sequence is empty
22         if not lines:
23             print(f"Error: No sequence found in {fasta_file}. The file might be empty or improperly formatted.")
24             return None
25
26         # Concatenate and convert to uppercase for consistency
27         sequence = "".join(lines).upper()
28
29         # Simple validation for a DNA sequence (A, T, C, G)
30         if any(base not in 'ATCG' for base in sequence):
31             print(f"Warning: The sequence contains non-standard DNA bases. Proceeding, but results may be affected.")
32
33         return sequence
34
35 except FileNotFoundError:
36     print(f"Error: The file '{fasta_file}' was not found.")
37     return None
38 except Exception as e:
39     print(f"An error occurred while reading the file: {e}")
40
```

The right-hand pane shows the Variable Explorer and the Python Console. The Variable Explorer displays the following variables:

Name	Type	Size	Value
dna_sequence	str	29903	ATTAAAGGTTTATACCTTCCAGGTACAAACCAACCACTTCGATCTCTGTAGATCTGTCTCTETAAA ...
FASTA_FILE_NAME	str	11	covid.fasta
repeat_data	defaultdict	4	defaultdict object of collections module

The Python Console shows the following output:

```
In [5]:
```

The status bar at the bottom indicates the current file is `lab7_1.py`, the interpreter is `Python 3.11.5`, and the current line is 132, column 35.

The screenshot shows the Spyder Python IDE interface. The main editor displays a Python script named 'lab7_2.py' with the following content:

```
1 # -*- coding: utf-8 -*-
2 """
3 Created on Thu Nov 13 15:23:18 2025
4
5 @author: Antonio
6 """
7
8 import pandas as pd
9 import matplotlib.pyplot as plt
10 from collections import defaultdict
11 import textwrap
12 import os
13 import sys
14
15 # --- 1. FASTA File Handling ---
16
17 def read_multi_fasta(fasta_file):
18     """Reads a multi-FASTA file and returns a dictionary of (header: sequence)."""
19     sequences = {}
20     current_header = None
21     current_sequence = []
22
23     try:
24         with open(fasta_file, 'r') as f:
25             for line in f:
26                 line = line.strip()
27                 if not line:
28                     continue
29
30                 if line.startswith('>'):
31                     if current_header and current_sequence:
32                         sequences[current_header] = "".join(current_sequence).upper()
33
34                     # Use only the first word (often the Accession ID or short name) as the ID
35                     current_header = line[1:].split()[0]
36                     current_sequence = []
37                 else:
38                     current_sequence.append(line)
39
40     except Exception as e:
41         print(f"Error reading FASTA file: {e}")
42
43 if __name__ == '__main__':
44     fasta_file = 'covid.fasta'
45     sequences = read_multi_fasta(fasta_file)
46     print(sequences)
```

The Variable Explorer on the right shows the following variables:

Name	Type	Size	Value
dna_sequence	str	29903	ATTAAAGGTTTATACCTTCCTCCAGGTAACAACCAACCACTTCGATCTCTGTAGATCTGTCTCTTAA ...
FASTA_FILE_NAME	str	11	covid.fasta
repeat_data	defaultdict	4	defaultdict object of collections module

The Console at the bottom shows the output of the script:

```
In [8]:
```

DataFrame of Top Repeats per Genome:

Genome ID	Top Motif	Total Copies
NC_026438.1	AGA	15
NC_026435.1	AAG	14
NC_026422.1	AGA	14
NC_026423.1	GAA	12
NC_026433.1	AAA	8
NC_026436.1	ATG	8
NC_026437.1	AGA	6
NC_026431.1	ACA	4
NC_026432.1	ACT	4
NC_026434.1	ATA	4

