

LABORATORY REPORT #7

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

lab7_1.py

The screenshot shows the Spyder Python IDE interface. On the left, the code editor displays `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")
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("Warning: The sequence contains non-standard DNA bases. Proceeding, but results mi")
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
```

On the right, the Variable Explorer shows the following variables:

| Name | Type | Size | Value |
|-----------------|-------------|------|--|
| dna_sequence | str | 2990 | ATTAAGGTTTACCTCCCAAGGATACAAACCAACTTGATCTCTGTAGATCTGTCCTAA... |
| FASTA_FILE_NAME | str | 11 | covid.fasta |
| repeat_data | defaultdict | 4 | defaultdict object of collections module |

The IPython Console shows the output of the code execution:

```
In [4]:
```

| NC_026431.1 | NC_026432.1 | NC_026434.1 | Sequence | Count |
|-------------|-------------|-------------|----------|-------|
| TCG | ACT | ATA | A | 4 |

At the bottom, the status bar indicates: Conda: anaconda3 (Python 3.13.5) ✓ LSP: Python Line 132, Col 35 UTE-B CSLE RW Mem 50% 11°C Cloudy 3:24 PM 11/13/2025

The screenshot shows the Spyder Python IDE interface. On the left, the code editor displays two files: lab7_1.py and lab7_2.py. The code in lab7_2.py reads a multi-FASTA file and returns a dictionary of headers and sequences. On the right, the Variable Explorer shows three variables: dna_sequence (a string of length 29903), FASTA_FILE_NAME (a string 'covid.fasta'), and repeat_data (a defaultdict object). Below the Variable Explorer is the Console I/A window, which shows the command 'In [8]:'. At the bottom, the system tray shows the date and time as 11/13/2025.

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 |

