

## LABORATORY REPORT #14

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

lab14\_1.py

The screenshot shows the Spyder Python IDE with the file lab14\_1.py open. The code defines functions for counting transitions and calculating probabilities for a sequence. The execution results are displayed in the Variable Explorer and the Console.

```
1 import math
2
3 def get_index(base):
4     if base == 'A': return 0
5     if base == 'C': return 1
6     if base == 'G': return 2
7     if base == 'T': return 3
8     return -1
9
10 def count_transitions(seq):
11     counts = [[0.0 for _ in range(4)] for _ in range(4)]
12     for i in range(len(seq) - 1):
13         idx_from = get_index(seq[i])
14         idx_to = get_index(seq[i+1])
15         if idx_from != -1 and idx_to != -1:
16             counts[idx_from][idx_to] += 1
17     return counts
18
19 def to_probabilities(counts):
20     probs = [[0.0 for _ in range(4)] for _ in range(4)]
21     for i in range(4):
22         row_sum = sum(counts[i])
23         if row_sum > 0:
24             for j in range(4):
25                 probs[i][j] = counts[i][j] / row_sum
26     return probs
27
28 s1 = "ATCGATTCGATATACACGTAT"
29 s2 = "CTCGACTAGTATGAAGTCACGGTTG"
30 s_test = "CAGGTTGGAACGTAA"
31 bases = ['A', 'C', 'G', 'T']
32
33 c1 = count_transitions(s1)
34 c2 = count_transitions(s2)
35
36 p1 = to_probabilities(c1)
37 p2 = to_probabilities(c2)
38
39 log_matrix = [[0.0 for _ in range(4)] for _ in range(4)]
40 epsilon = 1e-10
```

Variable Explorer:

Name	Type	Size	Value
bases	list	4	['A', 'C', 'G', 'T']
c1	list	4	[[0.0, 2.0, 0.0, 0.0], [2.0, 0.0, 3.0, 0.0], [2.0, 0.0, 0.0, 1.0], [3.0, ...]]
c2	list	4	[[1.0, 2.0, 2.0, 1.0], [1.0, 1.0, 2.0, 3.0], [2.0, 1.0, 0.0, 2.0], [2.0, ...]]
epsilon	float	1	1e-10
i	int	1	34
idx_from	int	1	0
idx_to	int	1	0
j	int	1	3
log_matrix	list	4	[[[-30.63411844815247, -0.4158374992788438, -31.634118448152467, 2.1699 ...]]
p1	list	4	[[0.0, 0.25, 0.0, 0.75], [0.4, 0.0, 0.6, 0.0], [0.6666666666666666, 0. ...]]
p2	list	4	[[0.16666666666666666, 0.3333333333333333, 0.3333333333333333, 0.1666 ...]]
p_minus	float	1	0.14285714285714285

Console:

```
In [21]: Xrunfile C:/Users/Antonio/Desktop/bioinfo/lab14/lab14_1.py --wdir
Log-likelihood Matrix:
      A      C      G      T
A -30.63 -0.42 -31.63  2.17
C  1.49 -30.41  1.07 -32.00
G  0.74 -30.90  0.00 -0.26
T  0.58  0.58 -31.41  0.00

Testing Sequence: CAGGTTGGAACGTAA
Total Log-Likelihood Score: -152.0126
Result: Sequence does NOT belong to a CpG island.
```

lab15\_2.py

The screenshot shows the Spyder Python IDE with the file lab15\_2.py open. The code defines functions for tokenizing text, training a model, and calculating log-likelihood. The execution results are displayed in the Variable Explorer and the Console.

```
1 Created on Thu Jun 22 14:43:22 2026
2
3 @author: Antonio
4 ---
5
6
7
8 import math
9 import re
10 from collections import defaultdict
11
12 def get_tokens(filename):
13     with open(filename, 'r', encoding='utf-8') as f:
14         text = f.read().lower()
15         text = re.sub(r'/\W+', '', text)
16         return text.split()
17
18 def train_model(tokens):
19     transitions = defaultdict(lambda: defaultdict(int))
20     counts = defaultdict(int)
21
22     for i in range(len(tokens) - 1):
23         prev, curr = tokens[i], tokens[i+1]
24         transitions[prev][curr] += count / counts[prev]
25         counts[prev] += 1
26
27     probs = defaultdict(dict)
28     for prev, followers in transitions.items():
29         for curr, count in followers.items():
30             probs[prev][curr] = count / counts[prev]
31     return probs
32
33 def get_log_likelihood(w1, w2, m_eminescu, m_stanescu):
34     p_e = m_eminescu.get(w1, {}).get(w2, 0)
35     p_s = m_stanescu.get(w1, {}).get(w2, 0)
36
37     if p_e == 0 and p_s == 0:
38         return 0
39     if p_e == 0:
40         return -10
41     if p_s == 0:
42         return 10
```

Variable Explorer:

Name	Type	Size	Value
attribution	str	7	NEUTRAL
i	int	1	118
j	int	1	3
model_eminescu	defaultdict	140	defaultdict object of collections module
model_stanescu	defaultdict	74	defaultdict object of collections module
score	int	1	0
segment	str	26	puneam deo marte dulce ma
tokens_eminescu	list	196	['fiind', 'balet', 'paduri', 'cutreieram', 'si', 'ma', 'culcam', 'ades ...]
tokens_stanescu	list	88	['primejdii', 'dulci', 'alcatuind', 'sub', 'gine', 'ma', 'te', 'investi ...]
tokens_suspicious	list	116	['fiind', 'balet', 'primejdii', 'cutreieram', 'prin', 'codrii', 'unde ...]
w1	str	5	dulce
w2	str	2	ma

Console:

```
Python 3.13.5 | packaged by Anaconda, Inc. | (main, Jun 12 2025, 16:37:03) [MSC v.1929 64 bit (AMD64)]
Type "copyright", "credits" or "license()" for more information.
Python 8.30.0 -- An enhanced Interactive Python. Type "?" for help.

In [1]: Xrunfile C:/Users/Antonio/Desktop/bioinfo/lab14/lab14_2.py --wdir
Index      Score      Attribution      Text Segment
-----
0          10.00      EMINESCU        fiind balet primejdii cutreieram prin
1           0.00      NEUTRAL         balet primejdii cutreieram prin codrii
2           0.00      NEUTRAL         primejdii cutreieram prin codrii unde
```

