

22W-COM SCI-CM121-LEC-1 / 22W-COM SCI-CM221-LEC-1 / 22W-CHEM-CM160A-LEC-1 / 22W-CHEM-CM260A-LEC-1 / 22W-BIOINFO-M221-LEC-1 / 22W-HUM GEN-M260A-LEC-1

Homework 1

WILLIAM CULVER RANDALL

TOTAL POINTS

25 / 25

QUESTION 1

1 Problem 1 10 / 10

- ✓ - **0 pts** Correct
- **10 pts** did not solve / no code

QUESTION 2

Problem 2 5 pts

2.1 2a 2 / 2

- ✓ - **0 pts** Correct
- **2 pts** Not Correct

2.2 2b 3 / 3

- ✓ - **0 pts** Correct
- **1 pts** 4 * P(CG) with no/wrong explanation
- **1 pts** Partially correct with good explanation
- **2 pts** Partially correct with no/bad explanation
- **3 pts** Not correct

QUESTION 3

Problem 3 10 pts

3.1 3a 1 / 1

- ✓ - **0 pts** Correct

3.2 3b 3 / 3

- ✓ - **0 pts** Correct
- **2 pts** intermediate graphs not shown/work not shown
- **1 pts** small error that propagates to next problems

3.3 3c 2 / 2

- ✓ - **0 pts** Correct
- **1 pts** no path
- **0 pts** correct conditional on previous error

3.4 3d 2 / 2

- ✓ - **0 pts** Correct
- **1 pts** Click here to replace this description.
- **2 pts** Click here to replace this description.

3.5 3e 2 / 2

- ✓ - **0 pts** Correct
- **1 pts** Click here to replace this description.
- **2 pts** Click here to replace this description.

homework1

problem 1

k-Mer Composition

```
#!/usr/bin/env python3
from itertools import product
import re

f = open("rosalind_kmer.txt", "r")
next(f)
s = f.read().replace("\n", "")

def permutations_with_repeats(alphabet, repeatNumber):
    alph = list(alphabet)
    out = []
    for c in product(alph, repeat = repeatNumber):
        out.append("".join(c))
    return sorted(out)

bases = 'ACTG'
perms = permutations_with_repeats(bases, len(bases))

for perm in perms:
    print(len(re.findall(f'(?={perm})', s)), end=' ')
```

problem 2

Question

Please be sure to show all corresponding work.

Assume the dictionary {A, C, G, T} with the following distribution:

- $P(A) = 0.1$
- $P(G) = 0.2$
- $P(C) = 0.2$
- $P(T) = 0.5$

part a

1 Problem 1 10 / 10

✓ - 0 pts Correct

- 10 pts did not solve / no code

homework1

problem 1

k-Mer Composition

```
#!/usr/bin/env python3
from itertools import product
import re

f = open("rosalind_kmer.txt", "r")
next(f)
s = f.read().replace("\n", "")

def permutations_with_repeats(alphabet, repeatNumber):
    alph = list(alphabet)
    out = []
    for c in product(alph, repeat = repeatNumber):
        out.append("".join(c))
    return sorted(out)

bases = 'ACTG'
perms = permutations_with_repeats(bases, len(bases))

for perm in perms:
    print(len(re.findall(f'(?={perm})', s)), end=' ')
```

problem 2

Question

Please be sure to show all corresponding work.

Assume the dictionary {A, C, G, T} with the following distribution:

- $P(A) = 0.1$
- $P(G) = 0.2$
- $P(C) = 0.2$
- $P(T) = 0.5$

part a

What is the expected frequency of the sequence CG in a sequence of length 3 (i.e., probability)? Hint: how many ways can you put CG in a sequence of length 3 and what is its probability? (2 points)

answer a

$$P(C) * P(G) * 1 + 1 * P(C) * P(G) \\ 0.2 * 0.2 * 1 + 1 * 0.2 * 0.2 = 0.08$$

part b

What is the expected frequency of the sequence CG in a sequence of length 5? (3 points)

answer b

There are 4 ways that **CG** can appear in a sequence of length 5 once.

*CG * * **
**CG * **
** * CG **
** * *CG*

There are 3 ways that **CG** can appear in a sequence of length 5 twice.

*CGCG **
*CG * CG*
**CGCG*

$$P(X = 1) = \\ P(CG * * * | *CG * * | * *CG * | * * *CG) - \\ 2 * P(CGCG * | CG * CG | *CGCG)$$

$$P(X = 1) = 4 * (0.2 * 0.2 * 1 * 1 * 1) - 2 * 3 * (0.2 * 0.2 * 0.2 * 0.2 * 1)(1) \\ = 0.1504 \quad (2)$$

$$P(X = 2) = P(CGCG * | CG * CG | *CGCG) \quad (3)$$

$$= 3 * (0.2 * 0.2 * 0.2 * 0.2 * 1) \quad (4)$$

$$= 0.0048 \quad (5)$$

2.12a 2 / 2

✓ - 0 pts Correct

- 2 pts Not Correct

What is the expected frequency of the sequence CG in a sequence of length 3 (i.e., probability)? Hint: how many ways can you put CG in a sequence of length 3 and what is its probability? (2 points)

answer a

$$P(C) * P(G) * 1 + 1 * P(C) * P(G) \\ 0.2 * 0.2 * 1 + 1 * 0.2 * 0.2 = 0.08$$

part b

What is the expected frequency of the sequence CG in a sequence of length 5? (3 points)

answer b

There are 4 ways that **CG** can appear in a sequence of length 5 once.

*CG * * **
**CG * **
** * CG **
** * *CG*

There are 3 ways that **CG** can appear in a sequence of length 5 twice.

*CGCG **
*CG * CG*
**CGCG*

$$P(X = 1) = \\ P(CG * * * | *CG * * | * *CG * | * * *CG) - \\ 2 * P(CGCG * | CG * CG | *CGCG)$$

$$P(X = 1) = 4 * (0.2 * 0.2 * 1 * 1 * 1) - 2 * 3 * (0.2 * 0.2 * 0.2 * 0.2 * 1)(1) \\ = 0.1504 \quad (2)$$

$$P(X = 2) = P(CGCG * | CG * CG | *CGCG) \quad (3)$$

$$= 3 * (0.2 * 0.2 * 0.2 * 0.2 * 1) \quad (4)$$

$$= 0.0048 \quad (5)$$

$$E(X) = \sum_{x \in X} x * P(x) \quad (6)$$

$$= 1 * P(X = 1) + 2 * P(X = 2) \quad (7)$$

$$= 1 * 0.1504 + 2 * 0.0048 \quad (8)$$

$$= 0.16 \quad (9)$$

problem 3

Question

Assume the following DNA sequence:

ATGATCGAGATC

part a

Draw the simple de Bruijn graph with $k = 3$, aka, `DeBruijn3(ATGATCGAGATC)`, that does not collapse any nodes. (1 point)

answer a



part b

Draw all intermediate collapsed graphs that collapse on common nodes. There are three such cases so be sure to provide three such graphs. (3 points)

answer b

2.2 2b 3 / 3

✓ - **0 pts** Correct

- **1 pts** 4 * P(CG) with no/wrong explanation
- **1 pts** Partially correct with good explanation
- **2 pts** Partially correct with no/bad explanation
- **3 pts** Not correct

$$E(X) = \sum_{x \in X} x * P(x) \quad (6)$$

$$= 1 * P(X = 1) + 2 * P(X = 2) \quad (7)$$

$$= 1 * 0.1504 + 2 * 0.0048 \quad (8)$$

$$= 0.16 \quad (9)$$

problem 3

Question

Assume the following DNA sequence:

ATGATCGAGATC

part a

Draw the simple de Bruijn graph with $k = 3$, aka, `DeBruijn3(ATGATCGAGATC)`, that does not collapse any nodes. (1 point)

answer a



part b

Draw all intermediate collapsed graphs that collapse on common nodes. There are three such cases so be sure to provide three such graphs. (3 points)

answer b

3.13a 1 / 1

✓ - 0 pts Correct

$$E(X) = \sum_{x \in X} x * P(x) \quad (6)$$

$$= 1 * P(X = 1) + 2 * P(X = 2) \quad (7)$$

$$= 1 * 0.1504 + 2 * 0.0048 \quad (8)$$

$$= 0.16 \quad (9)$$

problem 3

Question

Assume the following DNA sequence:

ATGATCGAGATC

part a

Draw the simple de Bruijn graph with $k = 3$, aka, `DeBruijn3(ATGATCGAGATC)`, that does not collapse any nodes. (1 point)

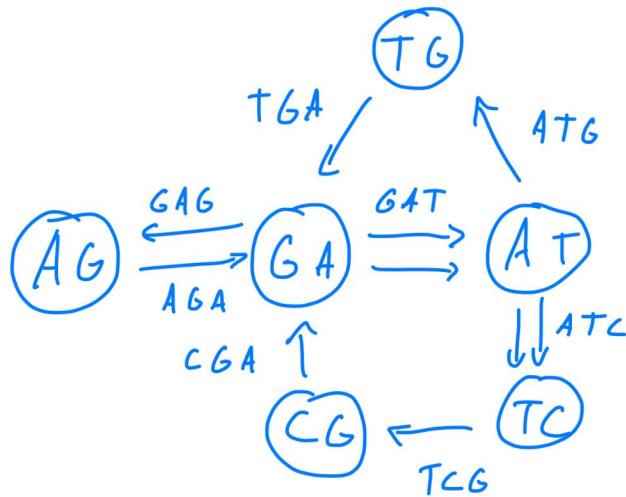
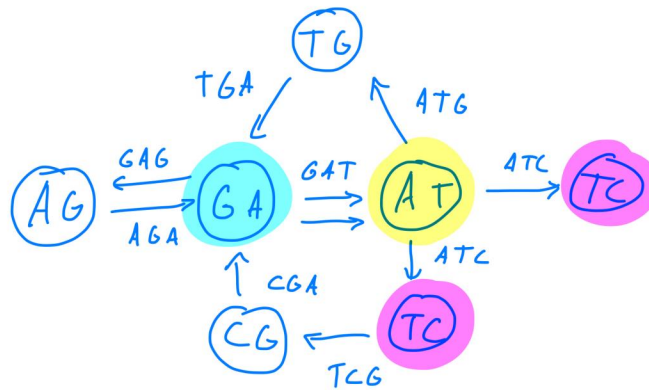
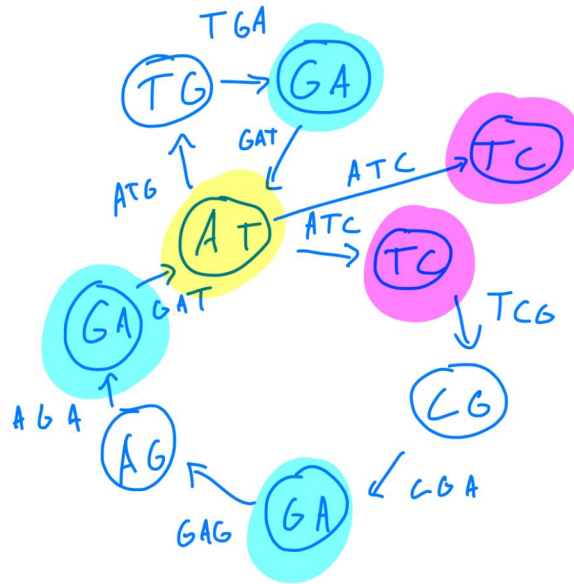
answer a



part b

Draw all intermediate collapsed graphs that collapse on common nodes. There are three such cases so be sure to provide three such graphs. (3 points)

answer b



3.2 3b 3 / 3

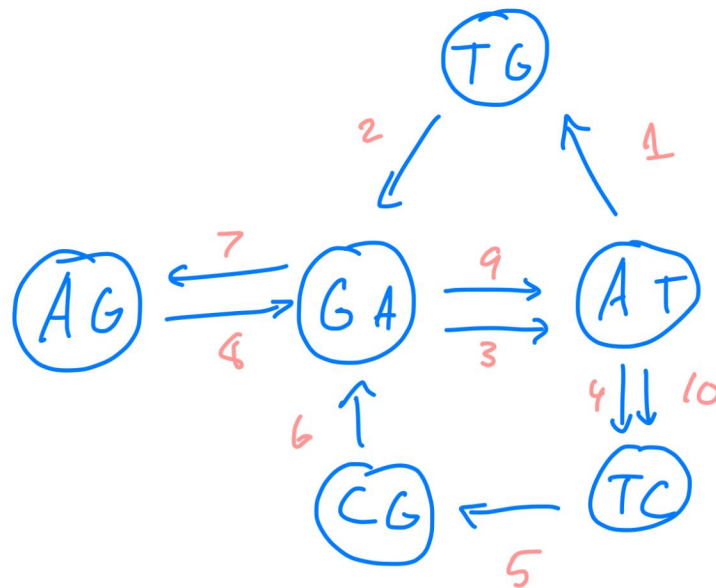
✓ - 0 pts Correct

- 2 pts intermediate graphs not shown/work not shown
- 1 pts small error that propagates to next problems

part c

Redraw your final graph from (b) and find the Eulerian path that corresponds to the original sequence but do not label the edges with their corresponding k-mer. Instead, label the edges on the Eulerian path edges with unique increasing integers starting with 1 (e.g. 1, 2, ...). (2 points)

answer c



part d

Find an Eulerian path that starts with a different k-mer. You can simply write down the edge labels here and make sure to write down the corresponding sequence. (2 points)

answer d

ATCGAGATGATC

part e

Draw the final Bruijn graph with $k = 5$, `DeBruijn5(ATGATCGAGATC)`, that collapses any duplicated nodes. (1 point)

answer e

3.3 3c 2 / 2

✓ - 0 pts Correct

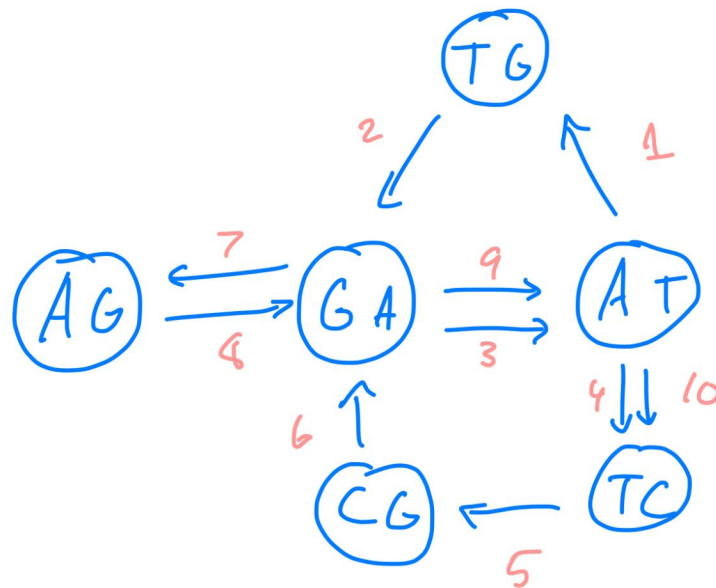
- 1 pts no path

- 0 pts correct conditional on previous error

part c

Redraw your final graph from (b) and find the Eulerian path that corresponds to the original sequence but do not label the edges with their corresponding k-mer. Instead, label the edges on the Eulerian path edges with unique increasing integers starting with 1 (e.g. 1, 2, ...). (2 points)

answer c



part d

Find an Eulerian path that starts with a different k-mer. You can simply write down the edge labels here and make sure to write down the corresponding sequence. (2 points)

answer d

ATCGAGATGATC

part e

Draw the final Bruijn graph with $k = 5$, `DeBruijn5(ATGATCGAGATC)`, that collapses any duplicated nodes. (1 point)

answer e

3.4 3d 2 / 2

✓ - 0 pts Correct

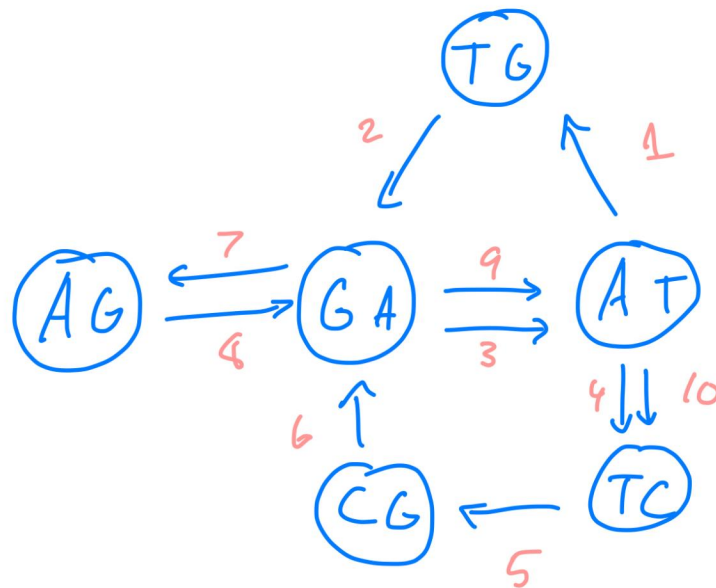
- 1 pts Click here to replace this description.

- 2 pts Click here to replace this description.

part c

Redraw your final graph from (b) and find the Eulerian path that corresponds to the original sequence but do not label the edges with their corresponding k-mer. Instead, label the edges on the Eulerian path edges with unique increasing integers starting with 1 (e.g. 1, 2, ...). (2 points)

answer c



part d

Find an Eulerian path that starts with a different k-mer. You can simply write down the edge labels here and make sure to write down the corresponding sequence. (2 points)

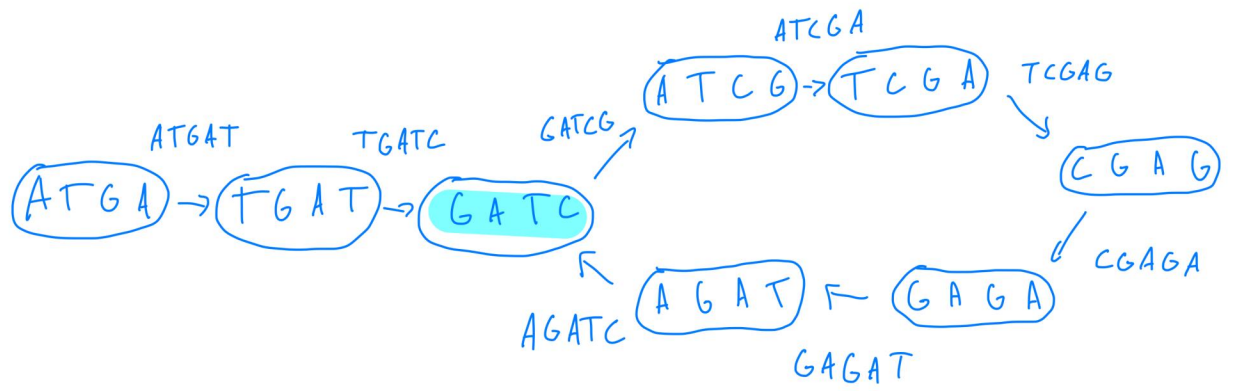
answer d

ATCGAGATGATC

part e

Draw the final Bruijn graph with $k = 5$, `DeBruijn5(ATGATCGAGATC)`, that collapses any duplicated nodes. (1 point)

answer e



3.5 3e 2 / 2

✓ - 0 pts Correct

- 1 pts Click here to replace this description.

- 2 pts Click here to replace this description.