INF281 Exercise 04

1. **Local alignment with DP**

   The DP algorithm can be used to identify optimal local alignments. Assume the scoring scheme as match: 1, mismatch: -1, and gap penalty: 1.

   (a) Complete the DP table to find the optimal local alignment.

	d		J	A	V	N	N
q			1	2	3	4	5
J	1						
A	2						
V	3						
A	4						
A	5						

   (b) Backtrack from $H_{9,6}$ and write down the local alignment.

	d		F	U	N	J	A	V	N	N	O	T
q			1	2	3	4	5	6	7	8	9	10
		0	0	0	0	0	0	0	0	0	0	0
F	1	0	1	0	0	0	0	0	0	0	0	0
U	2	0	0	2	1	0	0	0	0	0	0	0
N	3	0	0	1	3	2	1	0	1	1	0	0
T	4	0	0	0	2	2	1	0	0	0	0	1
O	5	0	0	0	1	1	1	0	0	0	1	0
N	6	0	0	0	1	0	0	0	1	1	0	0
J	7	0	0	0	0	2	1	0	0	0	0	0
A	8	0	0	0	0	1	3	2	1	0	0	0
V	9	0	0	0	0	0	2	**4**	3	2	1	0
A	10	0	0	0	0	0	1	3	3	2	1	0

2. **Dot matrix**

   A dot matrix is one of the simplest methods to identify local alignments.

   (a) Fill the table with dots.

	d	F	U	N	J	A	V	N	N	O	T
q		1	2	3	4	5	6	7	8	9	10
F	1										
U	2										
N	3										
T	4										
O	5										
N	6										
J	7										
A	8										
V	9										
A	10										

   (b) Identify all segment pairs with at least 3 contiguous dots along diagonals.

   (c) Identify all segment pairs with at least 3 contiguous dots along aniti-diagonals.

3. **N-grams**

   N-grams are n-letter words that can be used for database search methods. Create a table of 2-grams for q: ATGCAT.

   (a) List all 2-grams of q.

   (b) Fill the table with the 2-grams and the corresponding indices of q.

Index of q	2-gram of q

4. **Matching n-grams**

Calculate the scores of the segment pairs between q: CG and all 2-gram permutations of {A, C, G, T}.

Score matrix:

	A	T	G	C
A	2	-2	1	-2
T	-2	2	-2	1
G	1	-2	2	-2
C	-2	1	-2	2

(a) Fill the scores between CG and all its matching n-grams.

N-gram	Matching n-gram	Score
CG	AA	
CG	AC	
CG	AG	
CG	AT	
CG	CA	2 + 1 = 3
CG	CC	2 + (-2) = 0
CG	CG	2 + 2 = 4
CG	CT	2 + (-2) = 0
CG	GA	
CG	GC	
CG	GG	
CG	GT	
CG	TA	1 + 1 = 2
CG	TC	1 + (-2) = -1
CG	TG	1 + 2 = 3
CG	TT	1 + (-2) = -1

(b) Identify all matching n-grams when the threshold value T is 3.

5. **Lookup table for n-grams**

Create a 2-gram lookup table with indices and scores for the sequence q: ATGCAT.

Score matrix:

	A	T	G	C
A	2	1	-2	-2
T	1	2	-2	-2
G	-2	-2	2	-2
C	-2	-2	-2	2

T: 3

Pre-calculated scores of all segment pairs:

	AT	TG	GC	CA
AA	3	-1	-4	0
AC	0	-4	0	-4
AG	0	3	-4	-4
AT	4	-4	-4	-1
CA	-1	-4	-4	4
CC	-4	-4	0	0
CG	-4	0	-4	0
CT	0	-4	-4	3

	AT	TG	GC	CA
GA	-1	-4	0	0
GC	-4	-4	4	-4
GG	-4	0	0	-4
GT	0	-4	0	-1
TA	2	0	-4	0
TC	-4	0	0	-4
TG	-1	4	-4	-4
TT	3	0	-4	-1

(a) Fill the table.

N-gram of q	Indices of q	Matching n-grams	Scores of segment pair
AT			
TG			
GC			
CA			

(b) Create a lookup table for the matching n-grams with scores and indices.

Matching n-gram	Indices of q	Scores of segment pairs

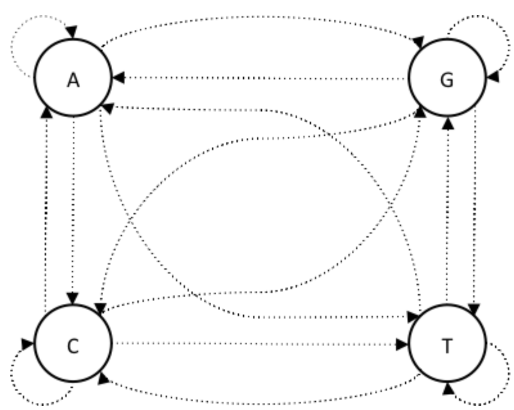
6. **Finite-state machine with 2-grams**

Use the 2-gram lookup table of q = ATGCAT to create a finite-state machine for all potential matching 2-grams.

Lookup table of 2-gram:

Matching 2-gram	**Indices of q**	**Scores of segment pairs**
AT	1, 5	4, 4
AA	1, 5	3, 3
TT	1, 5	3, 3
TG	2	4
AG	2	3
GC	3	4
CA	4	4
CT	4	3

(a) Add indices and scores to the corresponding edges.



(b) Use the finite-state machine to find the matching segment pairs and the scores.

1. d1: TCGGTAA

2. d2: ATAGC

7. **Finite-state machine with 3-grams**

   Add edges to connect nodes to create an overlap graph that can be used as a 3-gram finite-state machine.

   (a) List all 3-grams of AAACGGTA.

   (b) Use the finite-state machine to find the matching segment pairs and the scores.