INF281 Exercise 04 solutions

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	
-		•	0	0	0	0	0	0	
	J	1	0	1	0	0	0	0	
	A	2	0	0	2		0	0	q:
	V	3	0	0	1	3	2	1	d:
	A	4	0	0	1	2	2	1	
	A	5	0	0	1	1	1	1	

(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
0	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

Solution:

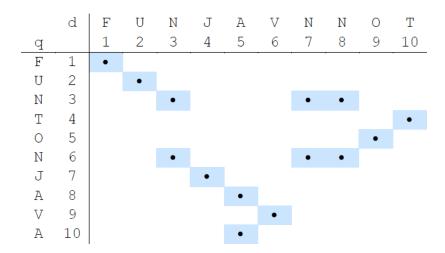
q: 6 NJAV 9

d: 3 NJAV 6

2. Dot matrix

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

(a) Fill the table with dots.



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

Solution:

q: 1 FUN 3 q: 6 NJAV 9 d: 1 FUN 3 d: 3 NJAV 6

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

Solution:

q: 4 TON 6 d: 8 NOT 10

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.

Solution:

AT, TG, GC, CA, AT

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

Index of q	2-gram of q
1	AT
2	TG
3	GC
4	CA
5	AT

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:

	А	Т	G	С
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	-2 + 1 = -1
CG	AC	-2 + (-2) = -4
CG	AG	-2 + 2 = 0
CG	AT	-2 + (-2) = -4
CG	CA	2 + 1 = 3
CG	CC	2 + (-2) = 0
CG	CG	2 + 2 = 4
CG	CT	2 + (-2) = 0
CG	GA	-2 + 1 = -1
CG	GC	-2 + (-2) = -4
CG	GG	-2 + 2 = 0
CG	GT	-2 + (-2) = -4
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.

Solution:

CA, CG, TG

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	С
A	2	1	-2	-2
Т	1	2	-2	-2
G	-2	-2	2	-2
С	-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

0	0
4	-4
0	-4
0	-1
-4	0
0	-4
-4	-4
-4	-1
	0 0 -4 0

(a) Fill the table.

N-gram of q	Indices of q	Matching n-grams	Scores of segment pair
AT	1, 5	AA, AT, TT	3, 4, 3
TG	2	AG, TG	3,4
GC	3	GC	4
CA	4	CA, CT	4, 3

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

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