

## 5 Exercises – Database search

### 1. 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
<b>1</b>	<b>AT</b>
<b>2</b>	<b>TG</b>
<b>3</b>	<b>GC</b>
<b>4</b>	<b>CA</b>
<b>5</b>	<b>AT</b>

## 2. 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	$-2 + 1 = -1$
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

### 3. 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		AT	TG	GC	CA
AA	3	-1	-4	0	GA	-1	-4	0	0
AC	0	-4	0	-4	GC	-4	-4	4	-4
AG	0	3	-4	-4	GG	-4	0	0	-4
AT	4	-4	-4	-1	GT	0	-4	0	-1
CA	-1	-4	-4	4	TA	2	0	-4	0
CC	-4	-4	0	0	TC	-4	0	0	-4
CG	-4	0	-4	0	TG	-1	4	-4	-4
CT	0	-4	-4	3	TT	3	0	-4	-1

(a) Fill the table.

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

(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	<b>1, 5</b>	<b>3</b>
AT	<b>1, 5</b>	<b>4</b>
TT	<b>1, 5</b>	<b>3</b>
AG	<b>2</b>	<b>3</b>
TG	<b>2</b>	<b>4</b>
GC	<b>3</b>	<b>4</b>
CA	<b>4</b>	<b>4</b>
CT	<b>4</b>	<b>3</b>

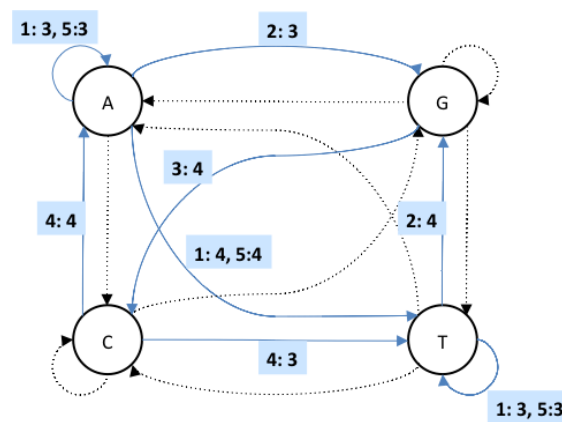
#### 4. Finite-state machine with 2-grams

Use the 2-gram lookup table of  $q = \text{ATGCAT}$  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

**Solution:**

q: 1 AT 2	Score: 3	q: 5 AT 6	Score: 3
d: 6 AA 7		d: 6 AA 7	

2. d2: ATAGC

**Solution:**

q: 1 AT 2	Score: 4	q: 5 AT 6	Score: 4
d: 1 AT 2		d: 1 AT 2	
q: 2 TG 3	Score: 3	q: 3 GC 4	Score: 4
d: 3 AG 4		d: 4 GC 5	

## 5. 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.

**Solution:**

AAA, AAC, ACG, CGG, GGT, GTA

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

