2 Exercises – Global pairwise alignment

1. **Pairwise alignments**

   Pairwise alignments are two aligned sequences of DNA, RNA, or protein. DNA and RNA sequences consist of four different nucleotides, whereas protein sequences consist of 20 different amino acids. The “-” sign is used to represent a blank or a gap, which indicates an insertion or a deletion from one sequence to the other.

   Use the simple scoring scheme and calcuate the score of the following alignments.

   Scoring scheme:
   $R_{ab} = 1$ for a = b
   $R_{ab} = 0$ for a $\neq$ b
   g = 1

   (a) Alignment 1

   ```
   q: ATGCT
   d: CA--T
   ```

   Solution: -1

   (b) Alignment 2

   ```
   q: CAGCT
   d: C-A-T
   ```

   Solution: 0

2. **Brute force approach**

   A brute force approach can be used to find the optimal alignment. Use the sequences *q* and *d* below and answer the questions.

   Sequences:

   `q: CG, d: AC`

   Scoring scheme:
   $R_{ab} = 1$ for a = b
   $R_{ab} = 0$ for a $\neq$ b
   g = 1

   (a) Identify all possible alignments.

   Solution:

   ```
   l = 4  CG--    C-G-    C--G    -CG-    -C-G    --CG
          --AC    -A-C    -AC-    A--C    A-C-    AC--

   l = 3  CG-   CG-   C-G   C-G   -CG   -CG
          A-C   -AC   AC-   -AC   A-C   AC-

   l = 2  CG
          AC
   ```

(b) Identify the optimal alignment with its score.

Solution:

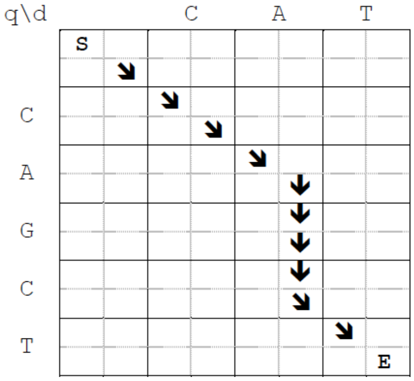
```
CG
AC    Score: 0
```

3. Table representation

An alignment can be represented as a table with arrows. Vertical and horizontal arrows indicate gaps, while diagonal arrows indicate matches and mismatches.

Identify the alignment that corresponds to the arrows in the following tables.

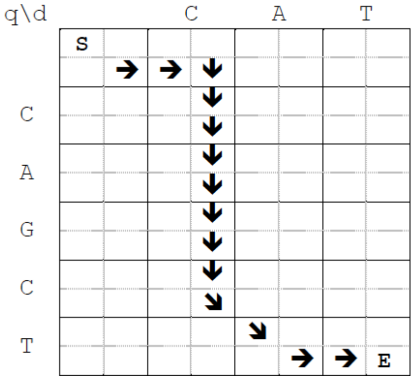
(a) Table 1



Solution:

```
q: CAGCT
d: CA--T
```

(b) Table 2



Solution:

```
q: -CAGCT-
d: C----AT
```

4. **DP table cell update rules**

Dynamic programming (DP) is an algorithm that uses table cells to memorize the sub-solutions of the target solution. DP requires three candidate scores and selects the maximum score among them when updating a cell.

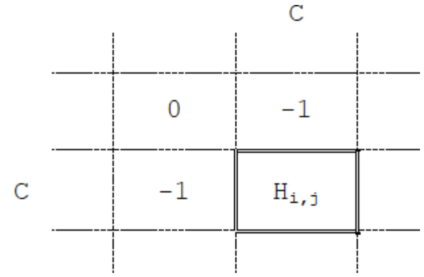
$$\begin{aligned} H_{i,j}^{(0)} &= H_{i-1,j} - g && \textit{(vertical)} \\ H_{i,j}^{(1)} &= H_{i,j-1} - g && \textit{(horizontal)} \\ H_{i,j}^{(2)} &= H_{i-1,j-1} + R_{a,b} && \textit{(diagonal)} \end{aligned}$$

Use the simple scoring scheme below to calcualte $H_{i,j}$ in Table A and B.

Scoring scheme:

$R_{ab} = 1$ for a = b
$R_{ab} = 0$ for a $\neq$ b
g = 1

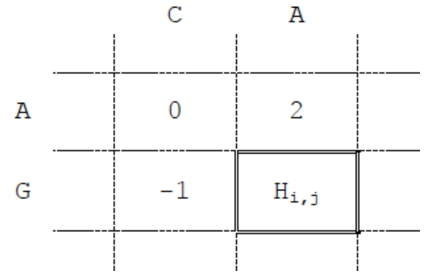
(a) Table A



		C
	0	-1
C	-1	$H_{i,j}$

Solution: 1

(b) Table B



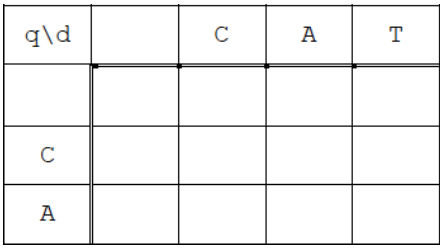
	C	A
A	0	2
G	-1	$H_{i,j}$

Solution: 1

5. **DP initialization**

   Initialization is the first step of the DP procedures.

   (a) Initialize the DP table with gap penalty 3.



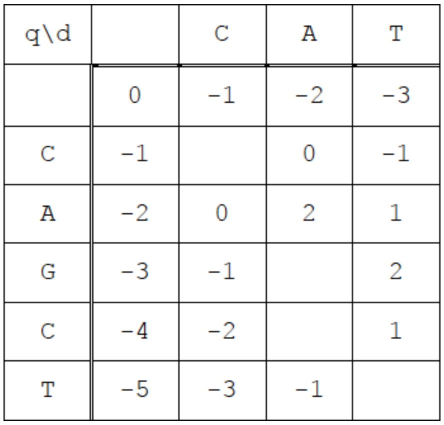
q\d		C	A	T
C				
A				

6. **DP global alignment**

   The score of optimal global alignment is found in the cell of the bottom-right corner after updating all cells.

   Scoring scheme:
   $R_{ab} = 1$ for a = b
   $R_{ab} = 0$ for a $\neq$ b
   g = 1

   (a) Use the simple scorning scheme and fill the empty cells with appropriate scores.



q\d		C	A	T
	0	-1	-2	-3
C	-1		0	-1
A	-2	0	2	1
G	-3	-1		2
C	-4	-2		1
T	-5	-3	-1	

   (b) What is the optimal score of the alignemnt?

   Solution: 1

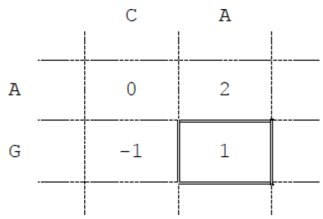
7. **DP backtrack**

Backtracking is a process to find the alignment with the optical score. It requires recalculations of the three candidate scores.

Scoring scheme:
R_{ab} = 1 for a = b
R_{ab} = 0 for a $\neq$ b
g = 1

(a) Which type of candidate score – vertical, horizontal, or diagonal – is used to update the cell with a double border? Assume that the simple scoring scheme has been used.

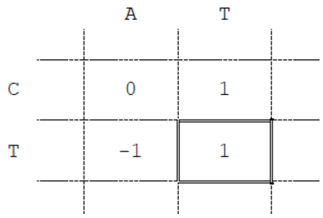
- Table 1



	C	A
A	0	2
G	-1	1

Solution: Vertical

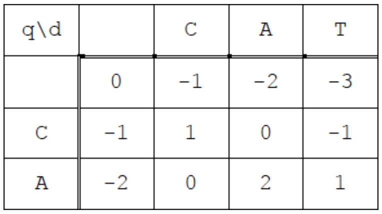
- Table 2



	A	T
C	0	1
T	-1	1

Solution: Diagonal

(b) Use backtracking to find the optimal global alignment.



q\d		C	A	T
	0	-1	-2	-3
C	-1	1	0	-1
A	-2	0	2	1

Solution:

```
q: CA-
d: CAT
```