Project 1

Sequence alignment and dynamic programming

Due to 13/10/17

Problem

Input
- Two sequences of variable length, for example:
  - GGVTTF
  - MEAIAKY

- Gap penalty, for example g = -4

- Substitution matrix, for example BLOSUM 62

Steps
1. Compute scores for scoring matrix
2. Backtrack to identify all possible alignments

Output
k best alignments/ maximum l alignments (NW/SW)

Example

GGVTTF (m=6)
MGGETFA (n=7)
Gap = -4

1. Create a matrix S of dimension (m+1)x(n+1) with the first sequence as rows and the second sequences as columns

		M	G	G	E	T	F	A
G								
G								
V								
T								
T								
F								

Example

GGVTTF (m=6)
MGGETFA (n=7)
g = -4

1. Create a matrix S of dimension (m+1)x(n+1) with the first sequence as rows and the second sequences as columns
2. Fill the first row/column with multiples of g

		M	G	G	E	T	F	A
	0	-4	-8	-12	-16	-20	-24	-28
G	-4							
G	-8							
V	-12							
T	-16							
T	-20							
F	-24							

Example

GGVTTF (m=6)
MGGETFA (n=7)
g = -4

1. Create a matrix S of dimension (m+1)x(n+1) with the first sequence as rows and the second sequences as columns
2. Fill the first row/column with multiples of g
3. Fill other cells according to:

$$max\{S(i-1,j) + g, S(i,j-1) + g, S(i-1,j-1) + t(i,j)\}$$

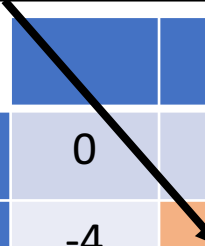|		M	G	G	E	T	F	A
	0	-4	-8	-12	-16	-20	-24	-28
G	-4							
G	-8							
V	-12							
T	-16							
T	-20							
F	-24							

Example

GGVTTF (m=6)
MGGETFA (n=7)
g = -4

1. Create a matrix S of dimension (m+1)x(n+1) with the first sequence as rows and the second sequences as columns
2. Fill the first row/column with multiples of g
3. Fill other cells according to:

$$max\{S(i-1,j) + g, S(i,j-1) + g, S(i-1,j-1) + t(i,j)\}$$

Max{-4+g,-4+g, 0+t('G','M')}

		M	G	G	E	T	F	A
	0	-4	-8	-12	-16	-20	-24	-28
G	-4							
G	-8							
V	-12							
T	-16							
T	-20							
F	-24							

Example

GGVTTF (m=6)
MGGETFA (n=7)
g = -4

1. Create a matrix S of dimension (m+1)x(n+1) with the first sequence as rows and the second sequences as columns
2. Fill the first row/column with multiples of g
3. Fill other cells according to:

$$max\{S(i-1,j)+g, S(i,j-1)+g, S(i-1,j-1)+t(i,j)\}$$

Max{-4+g,-4+g, 0+t('G','M')}   = Max{-8,-8,-3} = -3

BLOSUM 62

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	9																			
S	-1	4																		
T	-1	1	5																	
P	-3	-1	-1	7																
A	0	1	0	-1	4															
G	-3	0	-2	-2	0	6														
N	-3	1	0	-2	-2	0	6													
D	-3	0	-1	-1	-2	-1	1	6												
E	-4	0	-1	-1	-1	-2	0	2	5											
Q	-3	0	-1	-1	-1	-2	0	0	2	5										
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8									
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5								
K	-3	0	-1	-1	-1	-2	0	1	1	1	-1	2	5							
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5						
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4					
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4				
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	3	2	1	3	1	4			
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6		
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7	
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11

		M	G	G	E	T	F	A
	0	-4	-8	-12	-16	-20	-24	-28
G	-4	-3						
G	-8							
V	-12							
T	-16							
T	-20							
F	-24							

Example

GGVTTF (m=6)
MGGETFA (n=7)
g = -4

1. Create a matrix S of dimension (m+1)x(n+1) with the first sequence as rows and the second sequences as columns
2. Fill the first row/column with multiples of g
3. Fill other cells according to:

$$max\{S(i-1,j)+g, S(i,j-1)+g, S(i-1,j-1)+t(i,j)\}$$

Max{-8 +g,-3+g, -4+t('G','G')}   = Max{-12,-7,2} = 2

BLOSUM 62

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	9																			
S	-1	4																		
T	-1	1	5																	
P	-3	-1	-1	7																
A	0	1	0	-1	4															
G	-3	0	-2	-2	0	6														
N	-3	1	0	-2	-2	0	6													
D	-3	0	-1	-1	-2	-1	1	6												
E	-4	0	-1	-1	-1	-2	0	2	5											
Q	-3	0	-1	-1	-1	-2	0	0	2	5										
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8									
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5								
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5							
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5						
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4					
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4				
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	3	2	1	3	1	4			
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6		
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7	
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11

		M	G	G	E	T	F	A
	0	-4	-8	-12	-16	-20	-24	-28
G	-4	-3	2					
G	-8							
V	-12							
T	-16							
T	-20							
F	-24							

Example

GGVTTF (m=6)
MGGETFA (n=7)
g = -4

1. Create a matrix S of dimension (m+1)x(n+1) with the first sequence as rows and the second sequences as columns
2. Fill the first row/column with multiples of g
3. Fill other cells according to:

$$max\{S(i-1,j)+g, S(i,j-1)+g, S(i-1,j-1)+t(i,j)\}$$

BLOSUM 62

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	9																			
S	-1	4																		
T	-1	1	5																	
P	-3	-1	-1	7																
A	0	1	0	-1	4															
G	-3	0	-2	-2	0	6														
N	-3	1	0	-2	-2	0	6													
D	-3	0	-1	-1	-2	-1	1	6												
E	-4	0	-1	-1	-1	-2	0	2	5											
Q	-3	0	-1	-1	-1	-2	0	0	2	5										
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8									
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5								
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5							
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5						
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4					
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4				
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	3	2	1	3	1	4			
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6		
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7	
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11

		M	G	G	E	T	F	A
	0	-4	-8	-12	-16	-20	-24	-28
G	-4	-3	2	-2	-6	-10	-14	-18
G	-8	-7	3	8	4	0	-4	-8
V	-12	-7	-1	4	6	4	0	-4
T	-16	-11	-5	0	3	11	7	3
T	-20	-15	-9	-4	-1	8	9	7
F	-24	-19	-13	-8	-5	4	14	10

Example : global

1. Begin with the element of last row, last column

		M	G	G	E	T	F	A
	0	-4	-8	-12	-16	-20	-24	-28
G	-4	-3	2	-2	-6	-10	-14	-18
G	-8	-7	3	8	4	0	-4	-8
V	-12	-7	-1	4	6	4	0	-4
T	-16	-11	-5	0	3	11	7	3
T	-20	-15	-9	-4	-1	8	9	7
F	-24	-19	-13	-8	-5	4	14	10

Example : global

1. Begin with the element of last row, last column
2. Identify the previous step resulting in this value:
   - 14 + g ?
   - 7 + g ?
   - 9 + t('F','A') ?

		M	G	G	E	T	F	A
	0	-4	-8	-12	-16	-20	-24	-28
G	-4	-3	2	-2	-6	-10	-14	-18
G	-8	-7	3	8	4	0	-4	-8
V	-12	-7	-1	4	6	4	0	-4
T	-16	-11	-5	0	3	11	7	3
T	-20	-15	-9	-4	-1	8	9	7
F	-24	-19	-13	-8	-5	4	14	10

Example : global

1. Begin with the element of last row, last column
2. Identify the previous step resulting in this value:
   - 14 + g ?
   - 7 + g ?
   - 9 + t('F','A') ?

		M	G	G	E	T	F	A
	0	-4	-8	-12	-16	-20	-24	-28
G	-4	-3	2	-2	-6	-10	-14	-18
G	-8	-7	3	8	4	0	-4	-8
V	-12	-7	-1	4	6	4	0	-4
T	-16	-11	-5	0	3	11	7	3
T	-20	-15	-9	-4	-1	8	9	7
F	-24	-19	-13	-8	-5	4	14	10

Example : global

1. Begin with the element of last row, last column
2. Identify the previous step resulting in this value:
   - 14 + g ?
   - 7 + g ?
   - 9 + t('F','A') ?
3. Repeat step 2 until you reach (0,0)

		M	G	G	E	T	F	A
	0	-4	-8	-12	-16	-20	-24	-28
G	-4	-3	2	-2	-6	-10	-14	-18
G	-8	-7	3	8	4	0	-4	-8
V	-12	-7	-1	4	6	4	0	-4
T	-16	-11	-5	0	3	11	7	3
T	-20	-15	-9	-4	-1	8	9	7
F	-24	-19	-13	-8	-5	4	14	10

Example : global

1. Begin with the element of last row, last column
2. Identify the previous step resulting in this value:
   - 14 + g ?
   - 7 + g ?
   - 9 + t('F','A') ?
3. Repeat step 2 until you reach (0,0)
4. Find all possible alignments

M G G – E T F A
| | | | | | | |
- G G V T T F -

		M	G	G	E	T	F	A
	0	-4	-8	-12	-16	-20	-24	-28
G	-4	-3	2	-2	-6	-10	-14	-18
G	-8	-7	3	8	4	0	-4	-8
V	-12	-7	-1	4	6	4	0	-4
T	-16	-11	-5	0	3	11	7	3
T	-20	-15	-9	-4	-1	8	9	7
F	-24	-19	-13	-8	-5	4	14	10

Example : affine penalty gap

Different penalties for the initial gap I and the extension gap E

$max\{S(i-1,j)+g, S(i,j-1)+g, S(i-1,j-1)+t(i,j)\}$

AB-BBD vs AB-----BBD

- Matrix initializing must take this into account
- Need two additional matrices to keep gap information in memory, one for each sequence

We are not able to know if the gap was introduced before the values S(i-1,j) and S(I,j-1)

		M	G	G	E	T	F	A
G								
G								
V								
T								
T								
F								

Example : pénalité affine

Different penalties for the initial gap I and the extension gap E

AB-BBD vs AB-----BBD

- **First sequence:**
  - ▪ Previous value was not a gap
    - ➢ S(i-1,j) and I
  - ▪ Previous value was a gap
    - ➢ V(i-1,j) and E

$$V(i,j) = \max\{S(i-1,j) - I, V(i-1,j) - E\}$$

- **Second sequence:**
  - ▪ Previous value was not a gap
    - ➢ S(i,j-1) and I
  - ▪ Previous value was a gap
    - ➢ W(i,j-1) and E

$$W(i,j) = \max\{S(i,j-1) - I, W(i,j-1) - E\}$$

- **Scoring matrix S:**

$$S(i,j) = \max\{S(i-1,j-1) + t(i,j), V(i,j), W(i,j)\}$$

- Matrix initializing must take this into account
- Need two additional matrices to keep gap information in memory, one for each sequence

		M	G	G	E	T	F	A
	0	-4	-5	-6	-7	-8	-9	-10
G	-4							
G	-5							
V	-6							
T	-7							
T	-8							
F	-9							

Example with I = 4 and E = 1

Global vs local alignment

Global

- **The negatives values** are possible

- Backtracking begins at the **last row, last column**

Local

- Negatives values are replaced by **0** in the scoring matrix

- Backtracking begins at the **maximum value** in the scoring matrix

Advices for the project

- Like a scientific report (Intro, methods, results, discussion)
  - No copy-paste from the slides of the lessons
  - No print screen from the terminal
  - Illustrate your explanations with figures/plots

- Answer clearly to the questions

- No big block of code ! Structure your blocks smartly to explain your implementations

Schedule for the TP:
06/10/17
13/10/17 (last chance !)