

Projet 1

Alignement des séquences et programmation dynamique

À rendre pour le 25/10/19 12h (heure Bruxelles)

Énoncé du problème

Input

- Deux séquences de taille variable, par exemple :
 - GGVTTF
 - MEAIKY
- Pénalité de gap, par exemple $g = -4$
- Matrice de substitution, par exemple BLOSUM 62

Étapes

1. Calcule les scores de la matrice de scoring
2. Backtrace pour identifier tous les alignements possibles

Output

k meilleurs alignements/au maximum / alignements (NW/SW)

Exemple

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

1. Créer une matrice S de dimension (m+1)x(n+1) avec les lignes pour la 1ère séquence et les colonnes pour la seconde

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

Exemple

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

1. Créer une matrice S de dimension (m+1)x(n+1) avec les lignes pour la 1ère séquence et les colonnes pour la seconde
2. Remplir la première ligne/colonne avec les multiples de 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							

Exemple

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

1. Créer une matrice S de dimension (m+1)x(n+1) avec les lignes pour la 1ère séquence et les colonnes pour la seconde
2. Remplir la première ligne/colonne avec les multiples de g
3. Remplir les autres cellules selon :

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

		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							

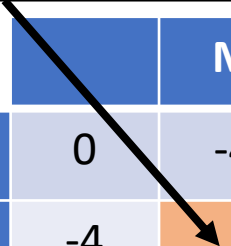
Exemple

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

1. Créer une matrice S de dimension (m+1)x(n+1) avec les lignes pour la 1ère séquence et les colonnes pour la seconde
2. Remplir la première ligne/colonne avec les multiples de g
3. Remplir les autres cellules selon :

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

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



		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							

Exemple

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

1. Créer une matrice S de dimension (m+1)x(n+1) avec les lignes pour la 1ère séquence et les colonnes pour la seconde
2. Remplir la première ligne/colonne avec les multiples de g
3. Remplir les autres cellules selon :

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

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

BLOSUM 62

BLOSUM 62

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	1	4						
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-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
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W

		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							

Exemple

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

1. Créer une matrice S de dimension (m+1)x(n+1) avec les lignes pour la 1ère séquence et les colonnes pour la seconde
2. Remplir la première ligne/colonne avec les multiples de g
3. Remplir les autres cellules selon :

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

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

BLOSUM 62

BLOSUM 62

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	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
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W

		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							

Exemple

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

1. Créer une matrice S de dimension (m+1)x(n+1) avec les lignes pour la 1ère séquence et les colonnes pour la seconde
2. Remplir la première ligne/colonne avec les multiples de g
3. Remplir les autres cellules selon :

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

BLOSUM 62

BLOSUM 62

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	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
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W

		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

Exemple : global

1. Commencer avec l'élément de la dernière ligne, dernière colonne

		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

Exemple : global

1. Commencer avec l'élément de la dernière ligne, dernière colonne
2. Identifier l'étape précédente qui a résulté en cette valeur:
 - $14 + g$?
 - $7 + g$?
 - $9 + t('F','A')$?

		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

Exemple : global

1. Commencer avec l'élément de la dernière ligne, dernière colonne
2. Identifier l'étape précédente qui a résulté en cette valeur:

- $14 + g ?$
- $7 + g ?$
- $9 + t('F','A') ?$

		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

Exemple : global

1. Commencer avec l'élément de la dernière ligne, dernière colonne
2. Identifier l'étape précédente qui a résulté en cette valeur:
 - $14 + g?$
 - $7 + g?$
 - $9 + t('F','A')$?
3. Répéter l'étape 2 jusqu'à arriver à (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

Exemple : global

1. Commencer avec l'élément de la dernière ligne, dernière colonne
2. Identifier l'étape précédente qui a résulté en cette valeur:
 - $14 + g ?$
 - $7 + g ?$
 - $9 + t('F','A') ?$
3. Répéter l'étape 2 jusqu'à arriver à (0,0)
4. Déterminer tous les alignements possibles

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

Exemple : pénalité affine

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

On ne sait pas savoir si le gap a été introduit avant les valeurs $S(i-1, j)$ et $S(i, j-1)$

Pénalités différentes pour le **gap initial I** et le **gap d'extension E**

AB-BBD vs AB-----BBD

- L'initialisation de la matrice score doit prendre ça en compte
- Besoin de 2 matrices supplémentaires pour garder l'information des gaps en mémoire, une pour chaque séquence

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

Exemple : pénalité affine

Pénalités différentes pour le **gap initial I** et le **gap d'extension E**

AB-BBD vs AB-----BBD

- **Gaps dans la sequence horizontale:**

- Valeur précédente n'était pas un gap
➤ $S(i-1, j)$ et I
- Valeur précédente était un gap
➤ $V(i-1, j)$ et E

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

- **Gaps dans la sequence verticale:**

- Valeur précédente n'était pas un gap
➤ $S(i, j-1)$ et I
- Valeur précédente était un gap
➤ $W(i, j-1)$ et E

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

- **Matrice de score S:**

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

- L'initialisation de la matrice score doit prendre ça en compte
- Besoin de 2 matrices supplémentaires pour garder l'information des gaps en mémoire, une pour chaque séquence

		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							

Exemple avec $I = 4$ et $E = 1$

Example $I = 4$ et $E = 1$

		M	G	G	E	T	F	A
	0	-4	-5	-6	-7	-8	-9	-10
G	-4	-3	2	1	-3	-4	-5	-6
G	-5	-7	3	8	4	3	2	1
V	-6	-4	-1	4	6	4	2	2
T	-7	-7	-2	3	3	11	7	6
T	-8	-8	-3	2	2	8	9	7
F	-9	-8	-4	1	0	6	14	10

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

V
gaps dans la
sequence
horizontale

		M	G	G	E	T	F	A
	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf
G	0	-8	-9	-10	-11	-12	-13	-14
G	0	-7	-2	-3	-7	-8	-9	-10
V	0	-8	-1	4	0	-1	-2	-3
T	0	-8	-2	3	2	0	-2	-2
T	0	-9	-3	2	1	7	3	2
F	0	-10	-4	1	0	6	5	3

W
gaps dans la
séquence
verticale

		M	G	G	E	T	F	A
	-Inf	0	0	0	0	0	0	0
G	-Inf	-8	-7	-2	-3	-4	-5	-6
G	-Inf	-9	-10	-1	4	3	2	1
V	-Inf	-10	-8	-5	0	2	1	0
T	-Inf	-11	-11	-6	-1	-1	7	6
T	-Inf	-12	-12	-7	-2	-2	4	5
F	-Inf	-13	-12	-8	-3	-4	2	10

Exemple avec $I = 12$ et $E = 2$

		M	G	G	E	T	F	A
	0	-12	-14	-16	-18	-20	-22	-24
G	-12	-3	-6	-8	-18	-20	-23	-22
G	-14	-15	3	0	-10	-13	-15	-17
V	-16	-13	-9	0	-2	-10	-14	-15
T	-18	-17	-11	-11	-1	3	-9	-11
T	-20	-19	-13	-13	-12	4	1	-9
F	-22	-20	-15	-16	-15	-8	10	-1

S

		M	G	G	E	T	F	A
	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf
G	0	-24	-26	-28	-30	-32	-34	-36
G	0	-15	-18	-20	-30	-32	-35	-34
V	0	-17	-9	-12	-22	-25	-27	-29
T	0	-19	-11	-12	-14	-22	-26	-27
T	0	-21	-13	-14	-13	-9	-21	-23
F	0	-23	-15	-16	-15	-8	-11	-21

V

gaps dans la sequence
horizontale

		M	G	G	E	T	F	A
	-Inf	0	0	0	0	0	0	0
G	-Inf	-24	-15	-17	-19	-21	-23	-25
G	-Inf	-26	-27	-9	-11	-13	-15	-17
V	-Inf	-28	-25	-21	-12	-14	-16	-18
T	-Inf	-30	-29	-23	-23	-13	-9	-11
T	-Inf	-32	-31	-25	-25	-24	-8	-10
F	-Inf	-34	-32	-27	-28	-27	-20	-2

W

gaps dans la séquence
verticale

'-GGVTTF', 'MGGETFA'

Alignement global vs local

Global

- Les **valeurs négatives** sont possibles dans la matrice score
- Le backtracking commence à la valeur de la **dernière ligne, dernière colonne**

Local

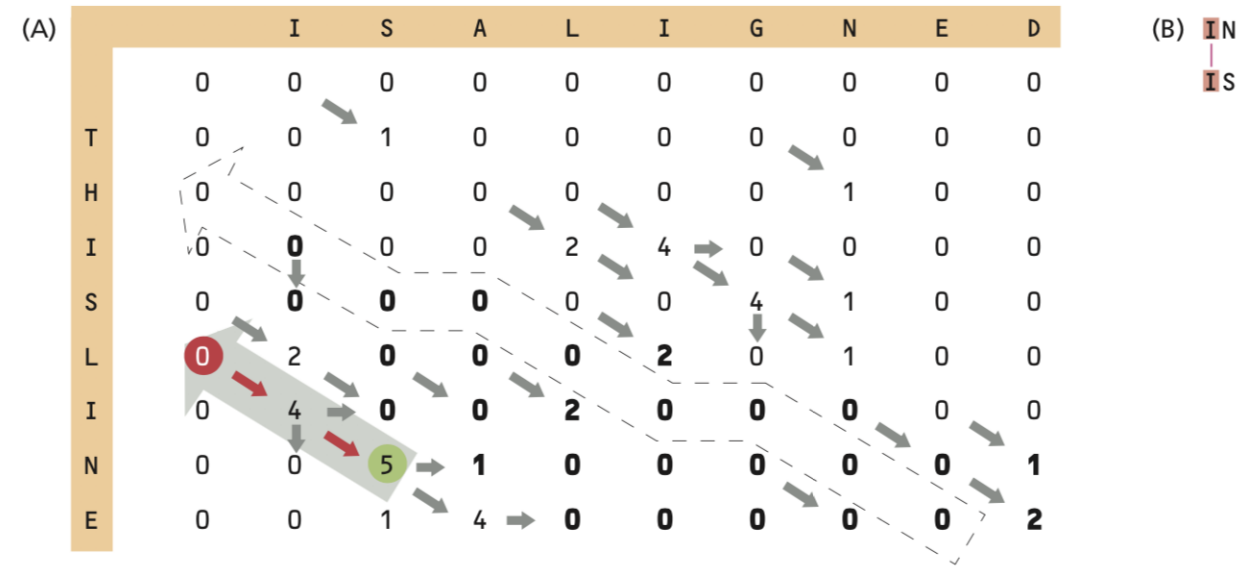
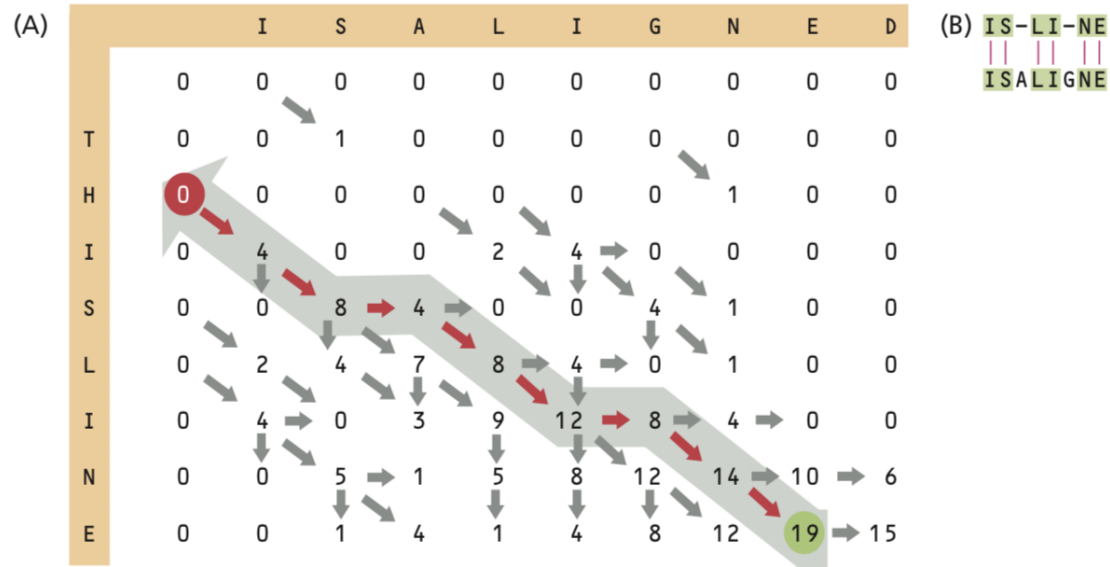
- Les valeurs négatives sont remplacées par **0** dans la matrice score, V et W
- Le backtracking commence à la **valeur maximale** de la matrice score

Plusieurs alignements sous-optimaux

- Mettre à 0 les positions de l'alignement optimal dans toutes les matrices (et les garder à 0)
- Recalculer les valeurs pour les 3 matrices
 - Hint n°1 : les éléments sont influencés uniquement par ce qui est en haut et à gauche d'eux
 - Hint n°2 : une fois qu'un élément ne change pas dans les trois matrices, cela signifie que tout ce qui se trouve à droite ne changera plus !
- Recommencer l'alignement local avec le plus haut score de la matrice

Exemple : restart

- Gap linéaire = -4



Exemple avec alignement local - 1

		I	S	A	L	I	G	N	E	D
	0	0	0	0	0	0	0	0	0	0
T	0	0	1	0	0	0	0	0	0	0
H	0	0	0	0	0	0	0	1	0	0
I	0	4	0	0	2	4	0	0	0	0
S	0	0	8	1	0	0	4	1	0	0
L	0	2	0	7	5	2	0	1	0	0
I	0	4	0	0	9	9	0	0	0	0
N	0	0	5	0	0	6	9	6	0	1
E	0	0	0	4	0	0	4	9	11	2

		I	S	A	L	I	G	N	E	D
	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf
T	0	0	0	0	0	0	0	0	0	0
H	0	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	0	0
S	0	0	0	0	0	0	0	0	0	0
L	0	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	0	0
N	0	0	0	0	0	0	0	0	0	0
E	0	0	0	0	0	0	0	0	0	0

		I	S	A	L	I	G	N	E	D
	-Inf	0	0	0	0	0	0	0	0	0
T	-Inf	0	0	0	0	0	0	0	0	0
H	-Inf	0	0	0	0	0	0	0	0	0
I	-Inf	0	0	0	0	0	0	0	0	0
S	-Inf	0	0	0	0	0	0	0	0	0
L	-Inf	0	0	0	0	0	0	0	0	0
I	-Inf	0	0	0	0	0	0	0	0	0
N	-Inf	0	0	0	0	0	0	0	0	0
E	-Inf	0	0	0	0	0	0	0	0	1

Score : 11.0

Identity percentage : 100.0%

Similarity rate : 100.0%

Percentage of gaps : 0.0%

NE - from 7 to 8

||

NE - from 7 to 8

Conseils pour la réalisation du projet

- Comme un rapport scientifique (Intro, Méthodes, Résultats, Discussion)
 - Pas de copier-coller des slides du cours
 - Pas de captures d'écran du terminal
 - Illustrez vos explications/discussions avec des graphiques/figures
 - Proof of concept !
- Répondez clairement aux questions posées
- Pas de gros blocs de code ! Structurez les blocs pour expliquer pertinemment vos implémentations



**KEEP
CALM
AND
BON
COURAGE**

Séances de TP :

Le 18/10/19

Le 25/10/19 (last chance !)