IFT 3295 – Bio-Informatique

Devoir 1

Nom Prénom : Boukharouba Boumediene

Matricule : 20032279

Nom Prénom: Kassis Charbel

Matricule : 0976458

Partie 1: ASSEMBLAGE DE SÉQUENCES

1) Quelle est la différence entre un tel alignement et l'alignement global ?

Concernant l'alignement global, on a deux séquences ayant à peu près la même longueur et le but est de trouver la similarité maximale entre les deux séquences et cela en trouvant le meilleur alignement possible. L'alignement proposé a deux séquences de longueurs différentes (ou pas) et le but est de trouver le meilleur alignement entre une sous séquence de la première séquence et une sous séquence de la deuxième séquence. Du coup, on peut se retrouver avec un alignement suffixe/préfixe ou préfixe/suffixe

2) Qu'elles doivent être les valeurs de la première ligne (V(0,j) ∀j) ? Et celle de la première colonne (V(i,0) ∀i) de la table de programmation dynamique V? Justifiez votre réponse.

Les valeurs de la première ligne (V(0,j) ∀j) Et celle de la première colonne (V(i,0) ∀i) de la table de programmation dynamique V seront initialisées à 0. En évitant de pénaliser la ligne ou la colonne on généralise le cas suffixe/préfixe, préfixe/suffixe car si on traite cas par cas on aura que

- Dans le cas d'un suffixe /préfixe on aura que les valeurs de la première ligne seront initialisées à 0 celle des colonnes seront pénalisantes (-8, -16 ...)
- Dans le cas de préfixe /suffixe on aura que les valeurs de la première colonne seront initialisées à 0 et celle des lignes seront pénalisantes (-8,-16, ...)

3)Quelles sont les équations de récurrence à utiliser pour remplir la table de programmation dynamique?

Pour tout i,j V(i,0) = V(0,j) = 0

V(i,j) = Max [$V(i-1,j-1)+ P(v_i,w_j)$, $V(i,j-1)+ P(-,w_j)$, $V(i-1,j)+P(v_i,-)$]

Commentaire explicatif :

$V(i-1,j-1)+ P(v_i,w_j)$ *: Flèche diagonale où* $P(v_i,w_j)$ *== match si char[i] = char[j] sinon c'est un mismatch*

$V(i,j-1)+ P(-,w_j)$, $V(i-1,j)+P(v_i,-)$ *: lorsqu'on a un indel.*

4) Comment peut-on retrouver l'alignement avec le meilleur chevauchement à partir de la table de programmation dynamique? Vous devez décrire la procédure entière pour retrouver l'alignement.

Après avoir rempli le tableau de programmation dynamique avec les règles de récurrences présentés ci-dessus, on cherche la valeur maximale de la dernière ligne et de la dernière colonne. Puis retrace le chemin en suivant les flèches jusqu'à une case qui n'a pas de flèches

La séquence alignée est engendrée à partir de cette dernière case jusqu'à la fin de la table c'est-à-dire jusqu'à la coordonnée de la valeur maximale de la dernière ligne/colonne

5)

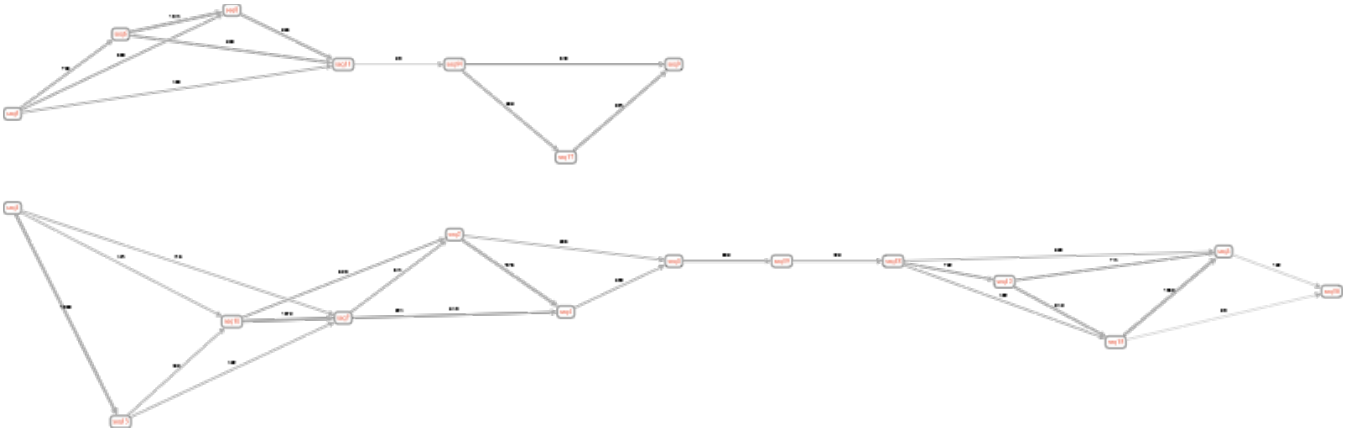
Le code de cette question se trouve dans le fichier **FonctionChevauchement.py.**

Partie2 : ASSEMBLAGE DE FRAGEMENT:

Question1 : le code de cette question se trouve dans le fichier Question1.py: Ce code va générer un EdgeList dans le fichier edgelist.txt

Question2:

A) **L'effet à ce seuil sur le graphe résultant :** Vue que sans ce seuil, chacun des reads (n reads) est relié à prés n-1 autres reads, alors on aura que les reads forward seront relié aux reads reverse. Cela dit, en imposant se seuil on aura de partie distincte, une qui représente le brin forward et une autre les reverse.



B) Les reads reverse se trouvent dans le graphe où il n'y a pas le reads 2 :

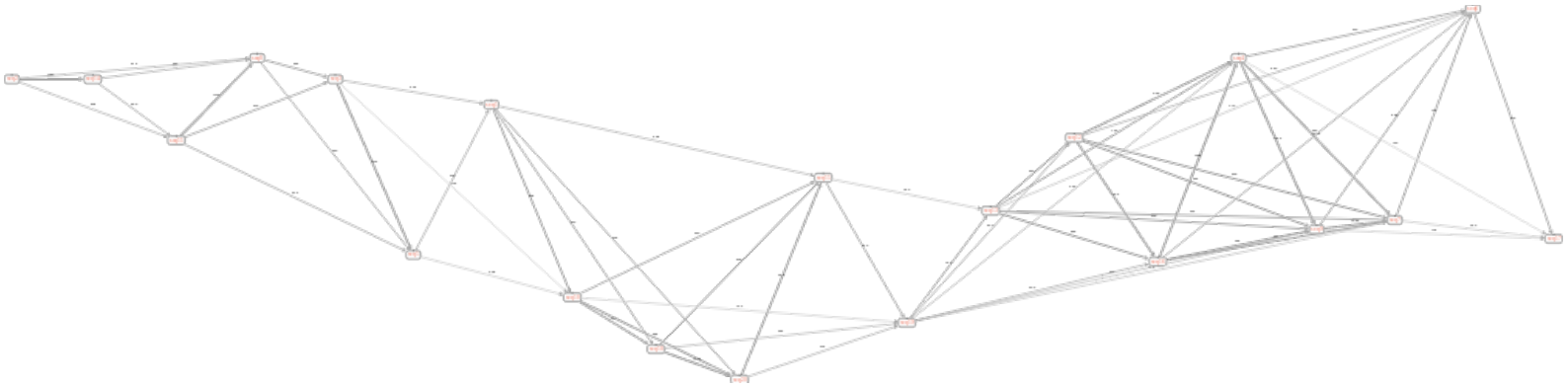
Reads_6, Reads_7, Reads_9, Reads_12, Reads_15, Reads_10, Reads_18

QUESTION3:

A)

Le code cette question se trouve dans le fichier Question3.py : Ce code génère d'abord un nouveau fichier FASTQ (reverse_reads.fq) en replaçant les reads inverses par leur inverse complément, puis un fichier EdgeList(edgelist_reverse.txt) à partir de ce ceci, puis après avoir appliqué l'algorithme de

réduction transitive, on génère un dernier fichier EdgeList (edgelist_reverse_reduite.txt) qui contiendra la solution.



On remarque qu'on a un seul graphe où tous les reads_ sont interconnectés, il n'y a pas de deuxième graphe.

B) -

Reads_2 ,Reads_14 ,Reads_11 ,Reads_8 ,Reads_3 ,Reads_5 ,Reads_1 ,Reads_10 ,Reads_18 ,Reads_20 ,Reads_15 ,Reads_19 ,Reads_13 ,Reads_12 ,Reads_16 ,Reads_4 ,Reads_9 ,Reads_7 ,Reads_6 ,Reads_17

Partie3 : Recherche d'introns et de Blast

Cadre de lecture 1 :

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

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

L L C S V G - R F T T S V L G R G K E K T - K E T P S T K S K F L L Y G C K M S

R - N L K S - F L Y - R K F L - G L L V W C V - L R Y I R T L C - V E V G L Q N

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

K C G Y S K - R C S L I V K Y S L D F D D V A Q K K N T L I T L F L K N F V S L

L E K F M L H M W L A V S - H - S Y - F F Y I L P E M S N R C C F P L L F I F N

I A F F F F - F A T R S P R F S A M L R Q W F F V - V V Q Q C C A S L Q E E R P

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

L S L E E F K W I G S K S E I Y L V - P G E I - I W S Q H M I F L R V I L L S N

I A K Y S L K I P L I K I I Y L R K V F R I V P K N - E Y I S G I K G - I I C I

- V L I Q Y S - N F S I L L K S T F C H - N Y S K G R M H L F N I L S - F F C R

L F I - K K A T L M I Q T A S - I L I L C C L T E S L I I N S I I L I N L P R -

C N Y I W F C K V Y S S N L L F W C Q F F N K V L I M G K

Cadre de lecture 2 :

G Q S L P Y H M F P D R A K A L F - G L G E H K Y Y E K Y - - K L L T G F E K -

L Y Y L T C C F C R G - P L F L F L H V Y F T Y G - **M C Q R C G L K L I V I I C**

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

G K I - N L N S F T K E N F C R D C - C G V Y S - D T L E P S V E - K W D Y R I

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

N V D T Q N R D V L - L - N T H W I L T M - H R K K I H - L H C F - K I L C R C

- K S L C Y T C G L L F H S T E V T D F F T Y Y Q K C R T G A V F L C F S F L T

L P F F F F S L L Q D H H G F Q P C S D S G S L C R L F N S V V P A Y R R K G Q

T H R R I I I W H S P T Q - - D - - L - M S L S S L K S L K K S - - L P K - L I

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

L L S I V - K Y L - S K L F T - E K Y S E - F L K I K S I F L V - K D K - S V Y

E Y - S N I L K T S V F Y L K V L F V I K I I A K V E C T C L I Y S H D S F A D

C S F R R K Q H - - F K Q L P E F - F C V V S Q K A L S - I P - F - L I Y Q D N

V I T F G F V R Y T A V I S Y F G V S F S I K F - L W A

Cadre de lecture 3:

D K V C L T I C F L I V R K P F S R D L E N T S I M K S T D E S Y - Q V S K N N

F T I - R V A S A E D D R Y S C F C M Y I S R M V K C A S V V V - N - - - - Y A

S L F S W L E I Y Y I R P W K R K R K N I K R N A - Y K V Q I L T L W M - N V Q

V K F E I L I P L L K K I S V G I A S V V C I V K I H - N P L L S R S G I T E L

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

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

R K V Y V T H V A C C F I A L K L L I F L H I T R N V E Q V L F S S A F H F - H

C L F F F L V C Y K I T T V F S H A Q T V V L C V G C S T V L C Q P T G G K A R

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

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

C - V - S E N T S N Q N Y L L E K S I Q N S S - K L R V Y F W Y K R I N N L Y M

S I N P I F L K L Q Y F T - K Y F L S L K L - Q R - N A L V - Y T L M I L L Q I

V H L E E S N T N D S N S F L N F N F V L S H R K P Y H K F H N S N - F T K I M

- L H L V L - G I Q Q - S P I L V S V F Q - S F D Y G Q

>geneX
MCQRCGLKLIVIICFFVQLARDLLHPSLEEEKKKHKKKRLVQSPNSYFMDVKCPGCYK
ITTVFSHAQTVVLCVGCSTVLCQPTGGKARLTEGCSFRRKQH

A) Dans quel cadre de lecture se trouve le codon start de la séquence protéique :

Afin de répondre à cette question nous allons inspecter les trois cadres et voir celui qui possède le codon start (Additionnant à cela une partie de la séquence protéique correspondante)

- **Réponse : Cadre de lecture numéro 2.**

B) Décrivez un algorithme de programmation dynamique qui vous permet de retrouver les différents fragments de la protéine X au sein de la séquence nucléotidique.Vous ne devez considérer que les plus longs fragments qui ne se chevauchent pas. Comme indice, vous pouvez supposer qu'il ne devrait pas avoir de mismatch entre les fragments de la séquence protéique et la traduction des régions nucléotidiques correspondants :

L'idée est d'utiliser un tableau de programmation dynamique qui prendra en paramètre la séquence (séquence.fasta) traduit en séquence protéique et le gêne X et appliquer les règle suivante :

```
If  ( Vi =Vj ):
        D(i-1, j-1) +1
Else:
```

0

Ensuite après avoir rempli le tableau, on Parcourt le tableau et à chaque cellule on vérifie si

FragmentDeproteine: [] //question b

//Parcourir le tableau de programmation dynamique

For i in range(len(GeneX)):

For j range(len(Sequence))

****

Question b:

On parcourt le tableau jusqu'à ce qu'il y a un match, si True alors on parcours le tableau en diagonale et tant qu'il n y a pas de 0 c'est qu'il n y a pas de missMatch et on continue à écrire la séquence dans FragmentDeproteine

If (tableau[i][j]>0):

While (tableau[i+1][j+1] >0):

FragmentDeproteine.append(**tableau[i])**

Print(FragmentDeProteine) //print la séquence (reponse b)

****

Question2:

	2eme Cadre	**3eme Cadre**	**2eme Cadre**	
GeneX	**[1:55]**	**[56:92]**	**[93:100]**	
Sequencefasta Protéique	**[66:121]**	**[287:324]**	**[480:488]**	
Sequence ARN	**[199:364]**	**[864:974]**	**[1441:1465]**	

Passant de la sequence.fasta protéique à la séquence ARN on va multiplier les intervalles par 3 et additionner 1 pour le cadron2 et multiplier par 3 et addition 2 pour le cadron3.

<u>L'ARNm Mature (après épissage) : (il faut changer les U en T puis coller les intervalles de la séquence ARN</u>

**AUGUGCCAGCGUUGUGGUUUAAAACUAAUAGUAAUAAUAUGCUUCUUUGUUCAG
UUGGCUAGAGAUUUACUACAUCCGUCCUUGGAAGAGGAAAAGAAAAAACAUAAA
AAGAAACGCCUAGUACAAAGUCCAAAUUCUUACUUUAUGGAUGUAAAAUGUCCA
GGUGCUACAAGAUCACCACGGUUUUCAGCCAUGCUCAGACAGUGGUUCUUUGUG
UAGGUUGUUCAACAGUGUUGUGCCAGCCUACAGGAGGAAAGGCCAGACUCACAG
AAGGAUUGUUCAUUUAGAAGAAAGCAACAC**

2- en vous servant de l'outil Blastp et/ou uniport, identifiez le nom de la protéine X ainsi que sa fonction.

Nom de la protéine :

Ribosomal protein S27-like

RPS27L

La protéine ribosomale S27-like est un régulateur physiologique de p53 qui supprime l'instabilité génomique et la tumorigénèse.