Travaux Pratiques : Alignement local et blast

1 Alignement local : Smith et Waterman

1.1 Principe

Avec l'alignement des séquences locales, vous n'êtes pas contraint d'aligner l'ensemble des deux séquences comme vous l'avez réalisé avec l'alignement global; vous pouvez simplement utiliser des parties de chacune des séquences pour obtenir un score maximum. Ainsi si on utilise deux séquences S1 et S2 et la même grille de notation que vous avez vu la semaine dernière, vous obtenez l'alignement local optimal S1" et S2" suivant:

```
S1 = GCCCTAGCG
S2 = GCGCAATG

S1    = GCCCTAGCG
S1'' =       GCG
S2'' =       GCG
S2    =       GCGCAATG
```

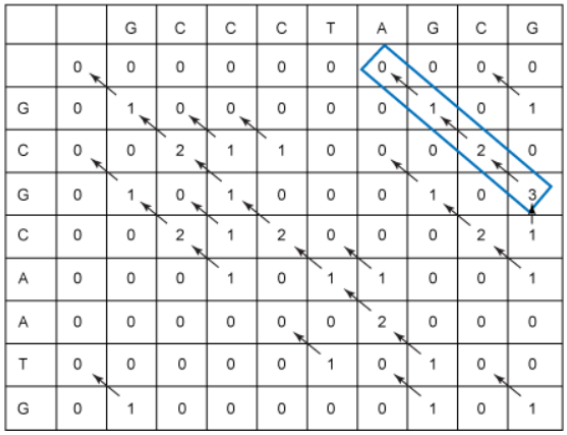
S1 = GCCCTAGCG
S2 = GCGCAATG

Dans l'algorithme Smith-Waterman, votre alignement local n'a pas besoin de se terminer à la fin de l'une ou de l'autre séquence, donc vous n'avez pas besoin de commencer votre traceback dans le coin inférieur droit; on le démarre dans la cellule avec le score le plus élevé.

L'algorithme Smith-Waterman diffère de l'algorithme Needleman-Wunsch en trois points :

- Dans l'étape d'initialisation, la première ligne et la première colonne sont toutes remplies de zéros.
- Deuxièmement, lorsque vous remplissez le tableau, si un score devient négatif, vous mettez 0 à la place, et vous ajoutez le pointeur en arrière seulement pour les cellules qui ont des scores positifs.
- Enfin, dans le traçage (traceback), vous commencez par la cellule qui a le score le plus élevé et retournez à la case précédente jusqu' à ce que vous atteigniez une cellule avec un score de 0. Sinon, le traçage fonctionne exactement comme dans l'algorithme Needleman-Wunsch.

Voici un exemple de table résumant les différents points (match : 1, mismatch : -1, déletion : -2):



		G	C	C	C	T	A	G	C	G
	0	0	0	0	0	0	0	0	0	0
G	0	1	0	0	0	0	0	1	0	1
C	0	0	2	1	1	0	0	0	2	0
G	0	1	0	1	0	0	0	1	0	3
C	0	0	2	1	2	0	0	0	2	1
A	0	0	0	1	0	1	1	0	0	1
A	0	0	0	0	0	0	2	0	0	0
T	0	0	0	0	0	1	0	1	0	0
G	0	1	0	0	0	0	0	1	0	1

Exercice 1 : Réutilisez et inspirez-vous des fonctions codées la semaine dernière pour :

- Construire la matrice.
- L'initialiser.
- Coder une nouvelle fonction qui va remplir les cellules.
- Coder une nouvelle fonction traceback.

2 Limitations

2.1 Mettons notre code à l'épreuve

Vous avez pu remarquer que pour un alignement quelqu'il soit, on crée une matrice de taille (taille de séquence 1*taille de séquence 2) est créée.

Alignez localement les séquences contenues dans les fichiers *gene1.txt* et *gene2.txt*.

Question : Qu'observez vous ?

2.2 Les alignements protéiques

L'alignement des séquences protéiques est plus complexe que les séquences nucléiques. Il existe 22 acides aminés chez l'Homme, on ne travaille donc plus sur un dictionnaire à 4 lettres mais à 22. Certains de ses acides aminés ont des propriétés physico-chimiques proches (voir *figure 1*).

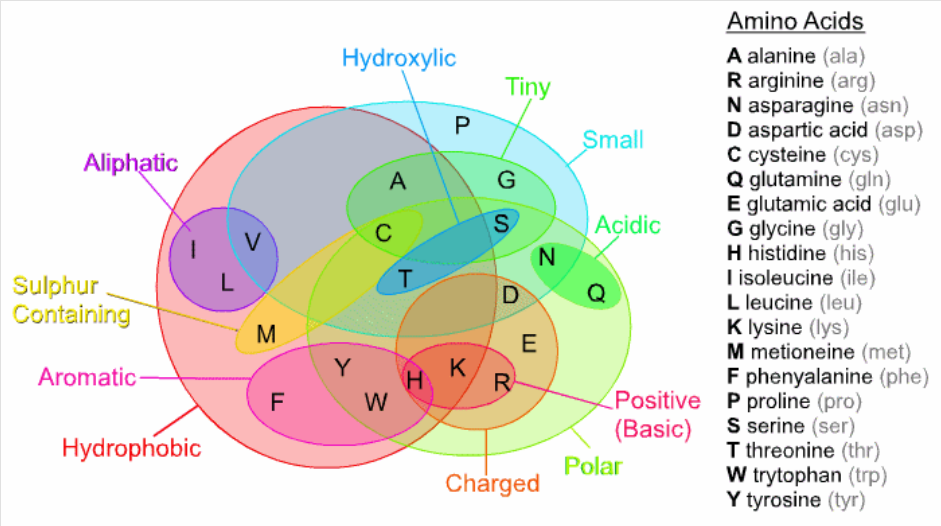
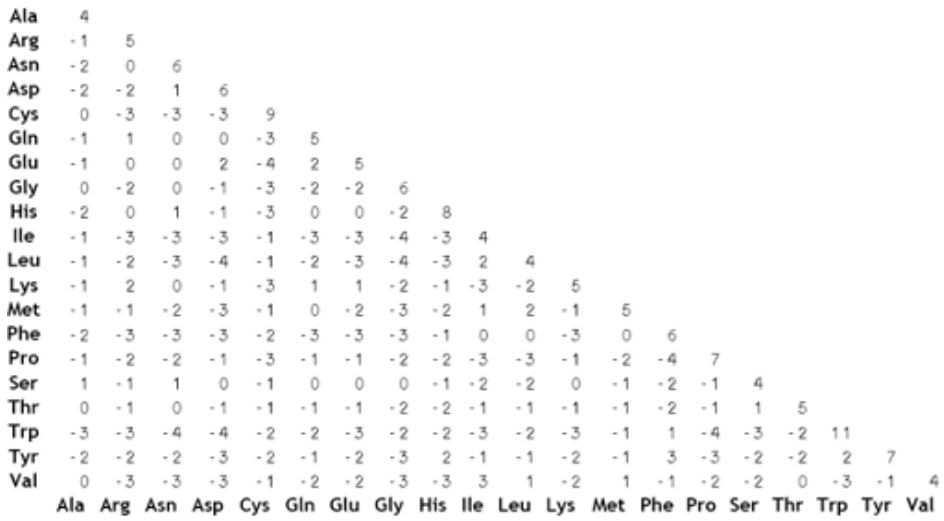


Figure 1: Diagramme de Venn des propriété phisico-chimiques

Or la substitution d'un acide aminé par un autre ayant les mêmes propriétés a moins de chance d'induire un grand changement sur la fonction de la protéine. C'est pour cela qu'il a été développé des matrices de substitutions spécifiques pour tenir compte des caractéristiques des différents acides aminés. Les plus connues étant les matrices BLOSSUM (BLOcks SUbstitution Matrix) et PAM (Point Accepted Mutation).



	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Figure 2: Exemple de matrice de substitution BLOSSUM62

3 Blast : basic local alignment search tool

3.1 Concept

Blast est un outil pour aligner localement des séquences mais il utilise une heuristique différente et plus optimisée. Cette heuristique permet notamment d'aligner une séquence face à une base de donnée qui contient une grande quantité de séquence.

Blast fonctionne en plusieurs étapes :

1 **Découpage de la séquence en k-mer :**

Le terme k-mer se réfère à tous les "sous mots"(substring) possible de longueur k contenus dans une chaîne de caractères. Par exemple, les 3-mers de la séquence ATCGATG sont ATC, TCG, CGA, GAT, ATG.

2 **Énumération de tous les mots correspondants possibles :**

Tout les k-mers de taille 3 possibles (AAA,ATA,ACT...) sont alignés avec chaque k-mer de notre séquence d'intérêt.

Les 3-mers qui ont des scores supérieurs à un seuil fixé sont gardés dans la suite des étapes.

3 **Organisation des k-mers ayant obtenu le plus de points dans un arbre de recherche efficace**

4 **Recherche dans la base de donnée les séquences possédant des matchs exacts avec nos k-mers retenus**

5 **Extension du matchs exacts dans la séquence de la base de données :**

BLAST va alors essayer de voir si cette région homologue s'étend au-delà du k-mer de départ. Il va alors essayer d'étendre en amont et en aval du k-mer pour voir si le score d'homologie augmente avec cette tentative d'extension.

Si les deux séquences présentent effectivement une homologie locale autour du k-mer de départ, l'extension va conduire à une augmentation effective du score, car de nouveaux nucléotides vont se trouver alignés. Si au contraire la tentative d'extension ne permet pas d'augmenter le score, parce que l'homologie ne continue pas, BLAST s'arrête. Si le score final après extension est supérieur à un seuil donné, l'alignement est conservé pour l'analyse finale.

6 **Analyse du score et évaluation de la pertinence :**

La recherche exhaustive avec BLAST retourne en général plusieurs dizaines d'alignements avec la séquence d'intérêt. Cependant, on ne peut rejeter l'hypothèse que ces résultats soient du au hasard (les bases de données contiennent énormément de séquences). BLAST évalue ses alignement en analysant la distribution des scores d'alignement entre la séquence d'intérêt et la banque. Il ajuste cette distribution à une fonction de densité théorique, ce qui lui permet de calculer la probabilité et l'espérance mathématique de trouver un alignement donnant un score donné dans la banque, uniquement du fait du hasard. Les paramètres de cette fonction de densité varient en fonction des compositions en nucléotides ou acides aminés de la séquence et de la banque analysée.

3.2 Les différents Blast

Différents outils blast existent et sont optimisés pour certains types de données ou pour réaliser des alignemnents particuliers.

- **blastn** : alignement de nucléotides, séquence nucléotidique contre une base de données de séquences nucléotidiques.
- **blastp** : alignement de protéines, séquence de protéine contre une base de données de séquences de protéines.
- **blastx** : alignement de séquence nucléotidique traduite en séquence de protéine contre une base de données de séquences de protéines.
- **tblastn** : alignement de séquence de protéine contre une base de données de séquences nucléotidiques traduites en séquences de protéines
- **tblastx** : alignement de séquence nucléotidique traduite en séquence de protéine contre une base de données de séquences nucléotidiques traduites en séquences de protéines.

3.3 Utilisation de Blast en ligne

Nous allons maintenant utiliser Blast en ligne et voir les possibilités qu'il y a derrière en terme d'anlyse.

- Allez sur https://blast.ncbi.nlm.nih.gov/Blast.cgi, et choisissez blastn
- Chargez le fichier sequence-unknown.txt dans "Enter Query Sequence".
- Dans "Program selection", choisissez blastn.
- Activez l'option "Show results in a new window".
- Lancez le BLAST.
- Quel résultat obtenez vous ? A quoi semble correspondre la séquence ?
- A quoi correspond le max score,total score, la query cover, l'e-value et l'identité ? (Aidez-vous de l'aide en ligne de blast)
- Retournez sur la page indiquée précédemment et cliquez sur blastx.
- Lancez une requête avec la *sequence-unknown.txt*.
- Quelles informations supplémentaires avez vous accès ?
- Combien de domaines d'activités sont référencés ? Dans quelle région ?

Bonus : Il est possible de modifier plusieurs paramètres dans les différents outils de BLAST (dans Algorithm parameters). Modifiez-les et observez s'il y a des différences.