Analyses de séquences

I_ Données stockées

A. Banques de données

On dispose de banques de données de séquences biologiques :

- Nucléique : EMBL *(UE)* et GenBank *(US)* qui contiennent approximativement 33 000 000 de séquences
- Protéique : SwissProt *(Suisse)*, UniProt *(UE)* et PIR qui contiennent approximativement 300 000 séquences
- Motifs peptidiques : ProSite qui possède 1800 motifs
- Structure : PDB qui contient 25 000 structures

Ces banques de données ont besoin d'énormément d'espace de stockage ainsi que de l'argent pour être entretenues.

à On observe parmi ces banques de données un problème de redondance dans les entrées enregistrées.

B. Lecture des informations

Les lignes de texte de la séquence sont notées selon un code qui permet de retrouver les informations données.

ID	Identificateur	AC	*Access codes*
DT	Dates d'accession	DE	Séquence correspondant à …
OS	Organisme	OC	Phylogénie
R_	Références bibliographiques	DR	*Databank references*
CC	Commentaires	FT	Features
XX	Séparateur		

C. Écriture des séquences-consensus

Une séquence consensus donne une petite séquence d'acides aminés pouvant varier, séparés par des tirets. L'écriture suit les règles suivantes :

[A,B,C]	Les acides aminés A, B et C sont possibles à cette position
{A,B,C}	Les acides aminés A, B et C ne sont pas possibles à cette position
x(#)	Il y a # acides aminés quelconques à cette position

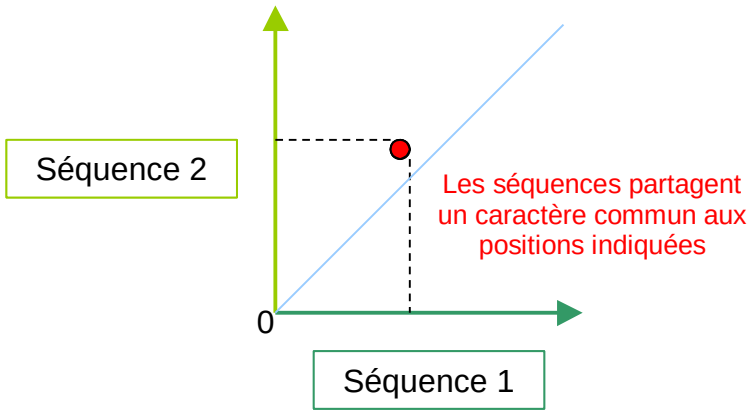
Ex : [R,K]-x(2)-{DE}-x(3)-Y

II_ Comparaison des données

A. Dot Plot

Un logiciel compare deux séquences et place un point à chaque fois que les séquences correspondent selon une matrice.

Ainsi on peut déceler des similitudes entre séquences ou des répétition lorsque l'on observe des droites sur le graphique.



Lorsque l'on utilise la matrice unitaire, cela signifie que l'on cherche l'exacte correspondance des séquences. Malheureusement, ce type de matrice appliquée à chaque caractère donnerait un graphique presque complètement noir.

Ex : matrice unitaire de l'ADN

	A	T	G	C
A	1	0	0	0
T	0	1	0	0
G	0	0	1	0
C	0	0	0	1

Pour éviter ce problème, le programme utilise deux paramètres :

- La fenêtre (*window*) : elle donne la longueur des chaînes de caractères à prendre en compte
- La rigueur (*stringency*) : c'est la valeur limite pour laquelle les chaînes comparées sont considérées semblables

Ex :

Fenêtre de 13 et rigueur de 6												
A	G	R	Y	D	E	A	L	M	Y	K	F	E
A	R	F	Y	I	K	A	Q	D	Y	K	P	N

à 5 acides aminés sont les mêmes, les chaînes sont différentes

B. Matrices

À l'intérieur du corps, des séquences polypeptidiques ou nucléotidiques différentes peuvent assurer les mêmes fonctions. Ainsi les acides aminés ou les nucléotides peuvent être considérés plus ou moins similaires s'ils ont les mêmes caractéristiques.

1) Matrices de protéines

- Matrice d'hydrophobicité :

On peut établir des matrices selon l'hydrophobicité des acides aminés. En effet, ce paramètre est important car il définit la position de l'acide aminé dans la structure protéique.

Méthode de détermination d'une matrice d'hydrophobicité

$$\text{int}\left[10 \times \left(1 - \frac{|i-j|}{|i-j|_{max}}\right)\right]$$

Matrice de différence : plus les acides aminés sont différents, plus la valeur est élevée

Rapportée au maximum

Matrice de similarité : plus les acides aminés sont similaires, plus la valeur est élevée

Rapportée à un nombre (ici 10)

l Matrices à taux de mutation :

On a établi des matrices appelées PAM250 et BLOSUM62 qui font la correspondance des acides aminés selon leur taux de mutation. On a comparé des séquences homologues et on a compté leur taux de mutation par acide aminé.

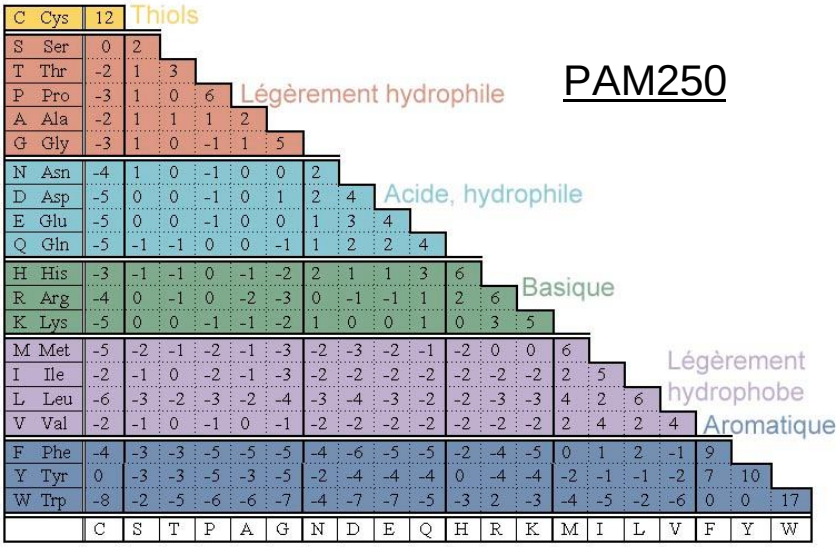
La matrice PAM (*Percent Accepted Mutation*) l'applique sur la séquence entière. Ainsi plus la valeur est élevée, plus l'acide aminé concerné est conservé.

La matrice BLOSUM ne compare que des morceaux de séquence qui ont un certain pourcentage d'identité. Ainsi il existe plusieurs matrices BLOSUM selon le pourcentage d'identité choisi (62 est la plus fréquente).

Remarque : Symboles utilisés

Symbole	Signification
B	Asx : Asp ou Asn
Z	Glx : Glu ou Gln
X	Acide aminé quelconque

PAM250



	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C Cys	12																			
S Ser	0	2																		
T Thr	-2	1	3																	
P Pro	-3	1	0	6																
A Ala	-2	1	1	1	2															
G Gly	-3	1	0	-1	1	5														
N Asn	-4	1	0	-1	0	0	2													
D Asp	-5	0	0	-1	0	1	2	4												
E Glu	-5	0	0	-1	0	0	1	3	4											
Q Gln	-5	-1	-1	0	0	-1	1	2	2	4										
H His	-3	-1	-1	0	-1	-2	2	1	1	3	6									
R Arg	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6								
K Lys	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5							
M Met	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6						
I Ile	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5					
L Leu	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6				
V Val	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4			
F Phe	-4	-3	-3	-5	-5	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9		
Y Tyr	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10	
W Trp	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17

2) Matrices d'acides nucléiques

Il existe aussi des matrices pour comparer les acides nucléiques. La matrice de similarité couramment employée est la matrice +5/-4 car son maximum est à 5 et son minimum à -4. Elle utilise aussi des nouveaux symboles.

Symbole	Signification
M	A ou C
R	A ou G
W	A ou T
S	C ou G
Y	C ou T
K	G ou T
V	Pas T
H	Pas G
D	Pas C
B	Pas A
X/N	N'importe quelle base

C. Algorithmes

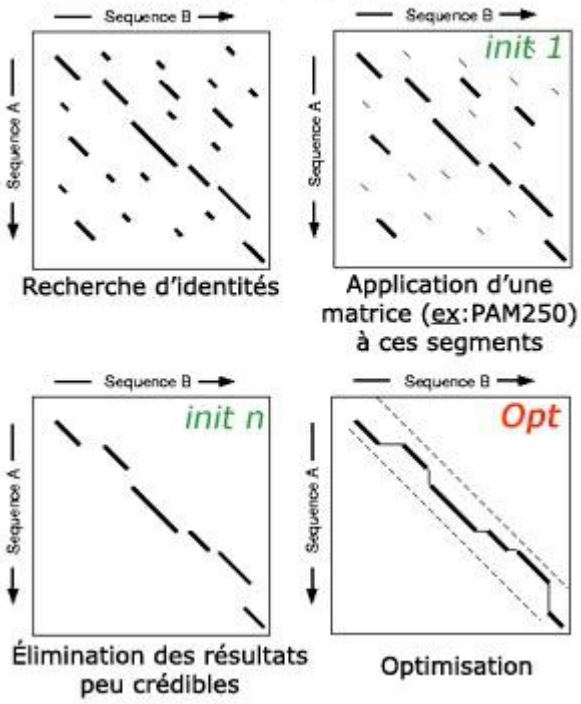
Des algorithmes peuvent appliquer une matrice entre une séquence donnée et toutes celles d'une base de données, pour ensuite citer quelles sont les séquences les plus similaires.

l FASTA :

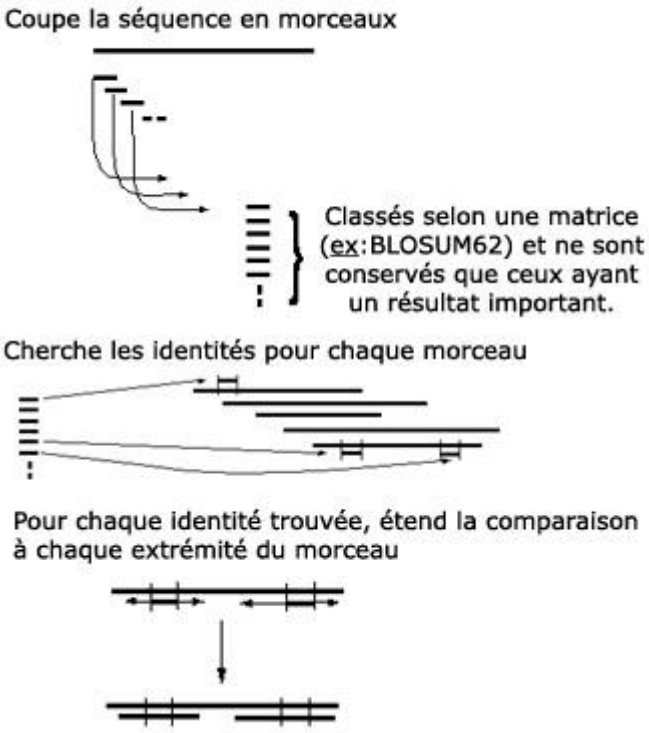
Cet algorithme a surtout été conçu pour les acides nucléiques. On l'utilise souvent pour comparer des séquences complètes (phylogénie).

Les séquences comparées donnent une valeur de *Opt* ; plus cette valeur est élevée, plus les séquences sont similaires.

FASTA



BLAST



- BLAST :

Cet algorithme ne compare que des morceaux de séquence. On obtient alors la probabilité que l'alignement soit du au hasard. Plus cette probabilité est faible, plus les séquences sont similaires.

Attention, il convient de vérifier comment les séquences sont alignées pour vérifier que ces résultats ne sont pas simplement des "hallucinations informatiques".

- BLITZ :

Encore un autre algorithme, qui donne comme pour BLAST la probabilité que l'alignement soit du au hasard (*prediction number*).

Il est important de vérifier que les résultats se croisent entre eux, pour une même banque de données. Cela confirmera l'hypothèse que la séquence comparée correspond bien à une protéine déjà connue.

Attention, les résultats obtenus par ces méthodes ne sont *que des hypothèses de travail*. Elles doivent faire l'objet d'un jugement rigoureux.

D. Alignements *pairwise*

On peut aussi utiliser une méthode d'alignement appelée *pairwise*, qui permet d'aligner des séquences de manière optimale en créant des espaces appelés *gap* au sein des séquences.

Ces *gap* dépendent de deux paramètres : la pénalité de création d'un gap pcg et la pénalité d'élongation d'un gap peg. Il est important de vérifier, d'après les connaissances que l'on a de l'origine de la séquence, que ces alignements sont corrects, et qu'ils ne sont pas des hallucinations informatiques.