Faculté des Sciences de la Nature et de la Vie
Département BMC, Université de Jijel

Cours de Bioinformatique

Licence Biochimie/Sciences pharmacologiques

Dr. BOULHISSA Ilham

Année universitaire: 2023-2024

Chapitre 3 : Alignement des séquences

Alignement des séquences biologiques

Mutation: modification rare (accidentelle ou provoquée) de l'information génétique (ADN, ARN ou protéine)

Les Mutations: Participent à l'évolution de l'espèce car elles constituent un élément clé de la biodiversité

Mutation héréditaire: Séquence mutée transmise à la génération suivante

À l'échelle moléculaire, les mutations peuvent être des:

1- Substitutions (changement ponctuelle d'un nucléotide par un autre)

2- Insertion (ajout d'un ou plusieurs nucléotides)

3- Délétion (délétion d'un ou plusieurs nucléotides)

• Les mutations provoquent des différences plus au moins importantes sur les structures des séquences biologiques (ADN, ARN ou protéines).

• Cause directe de la biodiversité et la divergence entre les espèces

• Exemple: Acétylcholinestérase

Acétylcholinestérase humaine (543 aa par chaine) et celle des souris (548 aa par chaine) exercent la même fonction bien qu'ils possèdent des structures différentes.

Comparaison des 10 premiers acides aminés de l'AChE issues des deux espèces:

AChE humaine EGREDAELLV
AChE des souris: EGREDPQLLV

Ces changements sont-ils acceptables ? Peuvent-ils influencer l'activité enzymatique ?

Comparaison des séquences: Alignement

Alignement: Permet de déterminer le degré de ressemblance entre les séquences afin d'indiquer que:

- La structure des séquences alignées est identique (notion d'identité).
- Les fonctions biologiques des séquences sont proches (notion de similarité)
- L'origine des séquences alignées est commune ou éloignée (notion d'homologie)

Si la fonction ou l'origine d'une séquence donnée sont inconnus, l'alignement de celle-ci avec des séquences issues des banques de données biologiques permet de tirer un maximum d'informations quant à la structure, la fonction et l'origine de la séquence inconnue.

Notions de base

En Bioinformatique, plusieurs termes sont employés pour décrire la notion de « ressemblance » entre deux séquences biologiques :

Notion d'identité: Ressemblance parfaite entre deux séquences (les deux séquences sont identique à 100%)

Notion de similarité: elle se mesure en % d'identité. Deux protéines sont similaires si elles ont un grand pourcentage d'identité e en assurant la même fonction biologique.

Notion d'homologie: elle a une connotation évolutive; 2 séquences sont homologues si elles ont un ancêtre commun même si elles n'assurent pas la même fonction biologique.

L'alignement des séquences nécessite la mise en œuvre de procédures de calculs afin de définir Les segments identiques entre deux séquences.

Notion de score: Le score élémentaire (noté "s") est une entité numérique que l'on attribue à chaque couple de nucléotides (aa) des deux séquences à comparer. Dans l'exemple ci-dessous, il prend la valeur de 1 lorsque les deux nucléotides des deux séquences sont identiques, et la valeur de 0 dans le cas contraire.

Exemple: Soit les deux séquences

Séquence1 A G C T A C C T G T
Séquence2 A A G T A G C T T T
Score élémentaire (s) **1+0+0+1+1+0+1+1+0+1 =6**

Score global (S): Total des scores élémentaires

$$S = \sum_{i=1}^{n} si$$

La somme des scores élémentaires est égale à six (score global).
Il y a six points identiques entre les deux séquences .
Le pourcentage d'identité soit 60% entre les deux séquences ([6/10] x100).
Le score global a donc permis de quantifier la ressemblance (l'identité) entre les deux séquences.

Alignement des séquences nucléiques

Les matrices nucléiques

Il existe peu de matrices pour les acides nucléiques car il n'y a que 4 lettres pour leur alphabet.

1- Matrice d'identité

Dans cette matrice, on attribue la valeur de 1 lorsque les deux nucléotides sont identiques et zéro s'ils ne le sont pas. Cette matrice rend compte de l'identité des bases pour chacune des positions de la comparaison, on parle ainsi de bon ou de mauvais appariement ou bien de bonne ou mauvaise association

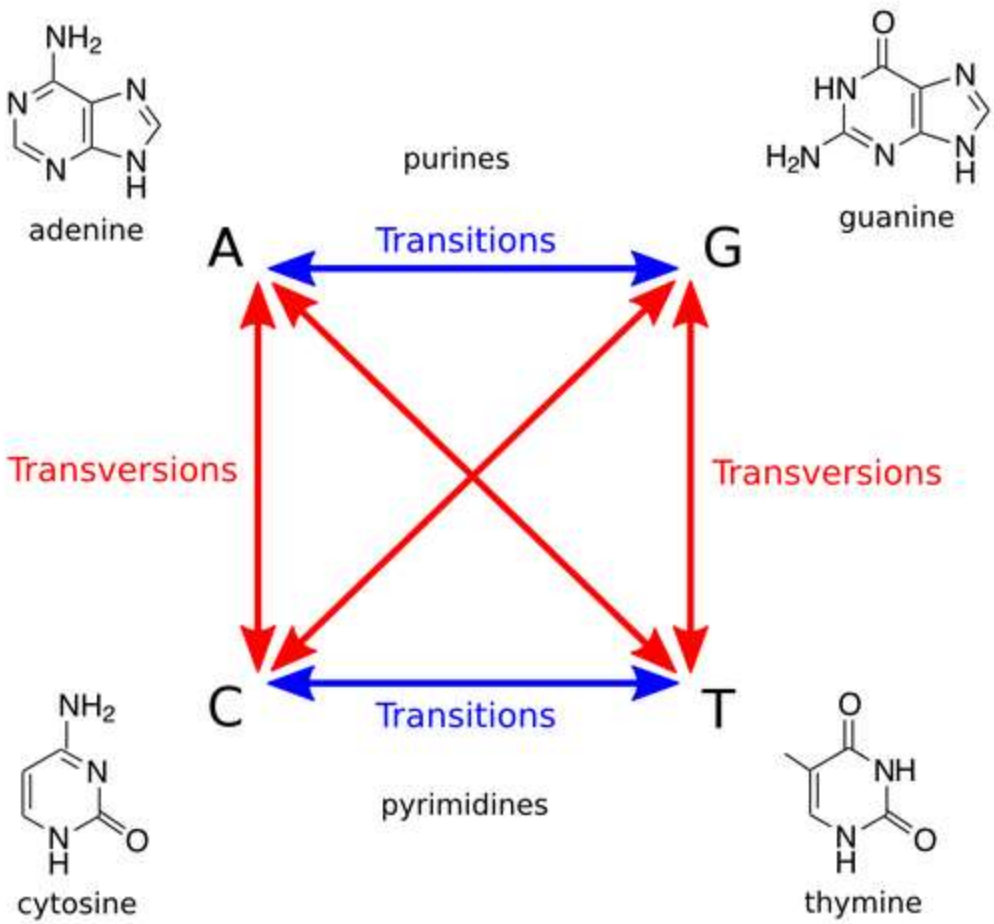
	A	T	G	C
A	1	0	0	0
T	0	1	0	0
G	0	0	1	0
C	0	0	0	1

2- Matrice de transition/transversion

Les bases azotées (Adénine, Guanine, Cytosine et Tymine) sont classées en 2 familles:
Les purines (Adénine et Guanine) et les pyrimidines (Thymine et Cytosine).

Une mutation de type **transition** est une substitution entre deux bases appartenant à la même famille. C'est à dire une purine devient une autre purine, ou alors une pyrimidine devient une autre pyrimidine.

Une mutation de type **transversion** est associée à un changement de famille. C'est une purine qui se transforme en pyrimidine ou l'inverse.


adenine
purines
guanine
A
Transitions
G
Transversions
Transversions
C
Transitions
T
cytosine
pyrimidines
thymine

Les mutations de type transition (A-G et T-C) se produisent plus souvent que celles de type Transversions (A-C, A-T, G-C et G-T).

La **matrice transition/transversion** représente cette information en tenant compte de l'analogie structurale entre purines (A et G) et pyrimidines (C et T) et attribue des scores en fonction de cette ressemblance.

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

3- Matrice de points (Dot-plot)

Méthode graphique permettant une comparaison visuelle et rapide de deux séquences sans l'utilisation des scores associés.

Dans cette matrice, les deux séquences sont représentées sur les deux axes x et y.

Un point (dot) est tracé pour chaque correspondance (identité) entre les deux séquences.

Une suite de points en diagonale révèle les régions (segments ou motifs) identiques entre les deux séquences et ce dans les deux sens.

Soit les deux séquences
s=ACTCGGATT
t=AGCTCGGT

		Séquence s								
		A	C	T	C	G	G	A	T	T
Séquence t	A	X						X		
	G					X	X			
	C		X		X					
	T			X					X	X
	C		X		X					
	G					X	X			
	G					X	X			
	T			X					X	X

Sur cette matrice, nous pouvons constater qu'il existe une diagonale formée de cinq cases.

Le segment identique le plus long entre ces deux séquences contient cinq nucléotides qui sont **CTCGG.**

		Séquence s								
		A	C	T	C	G	G	A	T	T
Séquence t	A									
	G									
	C		X							
	T			X						
	C				X					
	G					X				
	G						X			
	T									

Alignement des séquences nucléiques

Algorithme de Needleman et Wunsch

L'algorithme de Needleman-Wunsch est un algorithme qui effectue un alignement global maximal entre deux séquences nucléiques et/ou protéiques

$$S(i,j) = \text{Max} \begin{cases} s(i-1, j-1) + s(i,j) \\ s(i-1, j) \\ s(i, j-1) \end{cases}$$

j : numéro des cases dans l'axe des abscisses (horizontal)
i : numéro des cases dans l'axe dans des ordonnées (vertical)
S(i,j) = score de **Needleman-Wunsch** dans la case i,j
s (i,j)= score élémentaire dans la case i,j

Le score élémentaire (s) varie en fonction de la matrice à utiliser
Dans le cas d'identité entre deux nucléotides : s=1 si on utilise la matrice d'identité et 3 si on utilise la matrice de transition transversion).

L’alignement entre deux séquences nucléiques s’effectue en quatre étapes complémentaires :

1ère étape : Construction de la matrice initiale :

•Les deux séquences S1 et S2 sont insérées dans une matrice dite **initiale**
•La séquence S1 soit à l’horizontal (axe des abscisses) et la séquence S2 à la verticale du tableau (axe des ordonnées).
•Les cases de cette matrice doivent être remplies par des scores élémentaires (s) selon la matrice choisie (matrice d’identité ou de transition transversion).

Exemple :
S1 : TAAGTCCG
S2 : TACGTACG

Remplir la matrice initiale en utilisant la **matrice d’identité**

Matrice initiale

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

2ème étape : Construction de la matrice transformée

il faut créer une deuxième matrice à i+2 colonnes et j+2 lignes.
la 1ère ligne et la 1ère colonne seront initialisées à zéro comme suit :

		j	1	2	3	4	5	6	7	8
		S1	T	A	A	G	T	C	C	G
i	S2	0	0	0	0	0	0	0	0	0
1	T	0	1							
2	A	0								
3	C	0								
4	G	0								
5	T	0								
6	A	0								
7	C	0								
8	G	0								

C'est à ce niveau qu'on applique l'algorithme de **Needleman-Wunsch** afin d'aligner les deux séquences S1 et S2

$$S(i,j) = \text{Max} \begin{cases} s(i-1, j-1) + s(i,j) \\ s(i-1, j) \\ s(i, j-1) \end{cases}$$

En commençant par la case (1,1), l'algorithme est appliqué comme suit :

$$S(1,1) = \text{Max} \begin{cases} s(0,0) + s(1,1) = 0 + 1 = \mathbf{1} \\ s(0,1) = 0 \\ s(1,0) = 0 \end{cases}$$

		T	A	A	G	T	C	C	G
	0	0	0	0	0	0	0	0	0
T	0	1	1	1	1	1	1	1	1
A	0	1	2	2	2	2	2	2	2
C	0	1	2	2	2	2	3	3	3
G	0	1	2	2	3	3	3	3	4
T	0	1	2	2	3	4	4	4	4
A	0	1	2	3	3	4	4	4	4
C	0	1	2	3	3	4	5	5	5
G	0	1	2	3	4	4	5	5	6

Le maximum entre 1, 0 et 0 c'est bien 1. Donc le score à mettre dans la case i= 1 et j=1c'est **1.**

L’application de l’algorithme de **Needleman-Wunsch** permet de remplir la matrice transformée comme suit :

		T	A	A	G	T	C	C	G
	0	0	0	0	0	0	0	0	0
T	0	**1**	1	1	1	1	1	1	1
A	0	1	2	2	2	2	2	2	2
C	0	1	2	2	2	2	3	3	3
G	0	1	2	2	3	3	3	3	4
T	0	1	2	2	3	4	4	4	4
A	0	1	2	3	3	4	4	4	4
C	0	1	2	3	3	4	5	5	5
G	0	1	2	3	4	4	5	5	6

3ème étape : traceback (traçage en arrière)

le parcours de la matrice transformée commence par le plus haut score, vers le plus haut score parmi les trois cases (i-1, j-1) (i-1,j) et (i, j-1) et ainsi de suite jusqu'à la case (1,1).
Dans cet exemple, on commence par la case (8,8) ayant le plus haut score = 6.
Les scores des 3 cases (7,7), (7,8) et (8,7) sont les mêmes (5). Dans ce cas, le parcours en diagonal est recommandé (vers la case (7,7)).
Le parcours qui en résulte est le suivant :

	j	1	2	3	4	5	6	7	8
		T	A	A	G	T	C	C	G
i	0	0	0	0	0	0	0	0	0
1 T	0	1	1	1	1	1	1	1	1
2 A	0	1	2	2	2	2	2	2	2
3 C	0	1	2	2	2	2	3	3	3
4 G	0	1	2	2	3	3	3	3	4
5 T	0	1	2	2	3	4	4	4	4
6 A	0	1	2	3	3	4	4	4	4
7 C	0	1	2	3	3	4	5	5	5
8 G	0	1	2	3	4	4	5	5	6

4ème étape : Génération de l'alignement et calcul des scores

En suivant le parcours de la matrice transformée, les nucléotides en diagonal représentent soit appariement (identité **|**) ou une substitution (**:**).

S1	**T**	**A**	**A**	**G**	**T**	**—**	**C**	**C**	**G**
	|	**|**	**:**	**|**	**|**	*****	**|**	*****	**|**
S2	**T**	**A**	**C**	**G**	**T**	**A**	**C**	**—**	**G**

Le trou (**—**) retrouvé entre les nucléotides T et C de la séquence S1 est un GAP ou InDel (représenté en * lors de l'alignement).
À ce niveau, la séquence S1 a subit une mutation par **DELétion** au cours de la quelle un nucléotide a été perdu par nécessité évolutive ou adaptation à l'environnement. En même temps, ce nucléotide A a été conservé dans la séquence S2 (à la 6ème position).

On peut supposer que la séquence S2 a subi une mutation par **INsertion** du nucléotide A par nécessité adaptative ou adaptation à l'environnement.

•Communément; ce point est appelé INDEL (ou Gap).

•Une substitution a eu lieu au niveau de la troisième position où le nucléotide A de la séquence S1 a été substitué par un C dans la séquence S2 et ce par nécessité évolutive et d'adaptation à l'environnement.

Calcul des scores :

Le pourcentage d'identité (%id) = (nombre d'identités / taille de la séquence après alignement)

Dans cet exemple %id= (6/9)* 100 = 66.66%

Le pourcentage des gaps : = (nombre gaps / taille de la séquence après alignement)*100

= (2/9)*100= 22.22%

Le pourcentage des substitutions : (nombre de substitutions / taille de la séquence après alignement)*100

= (1/9)*100= 11.11%.

Exercice:
Soit les deux séquences:
S1: ACGGCTAT
S2: ACTGTAT

Établir un alignement global entre ces séquences en se basant sur la matrice d'identité?
Calculer le pourcentage d'identité, de gaps et de substitution?

	S1	A	C	G	G	C	T	A	T
S2	0	0	0	0	0	0	0	0	0
A	0	1	1	1	1	1	1	1	1
C	0	1	2	2	2	2	2	2	2
T	0	1	2	2	2	2	3	3	3
G	0	1	2	3	3	3	3	3	3
T	0	1	2	3	3	3	4	4	4
A	0	1	2	3	3	3	4	5	5
T	0	1	2	3	3	3	4	5	6

S1	A	C		G	G	C	T	A	T
S2	A	C	T	G			T	A	T

Alignement des séquences protéiques

Les matrices utilisées pour aligner les protéines sont totalement différentes de celles des acides nucléiques :

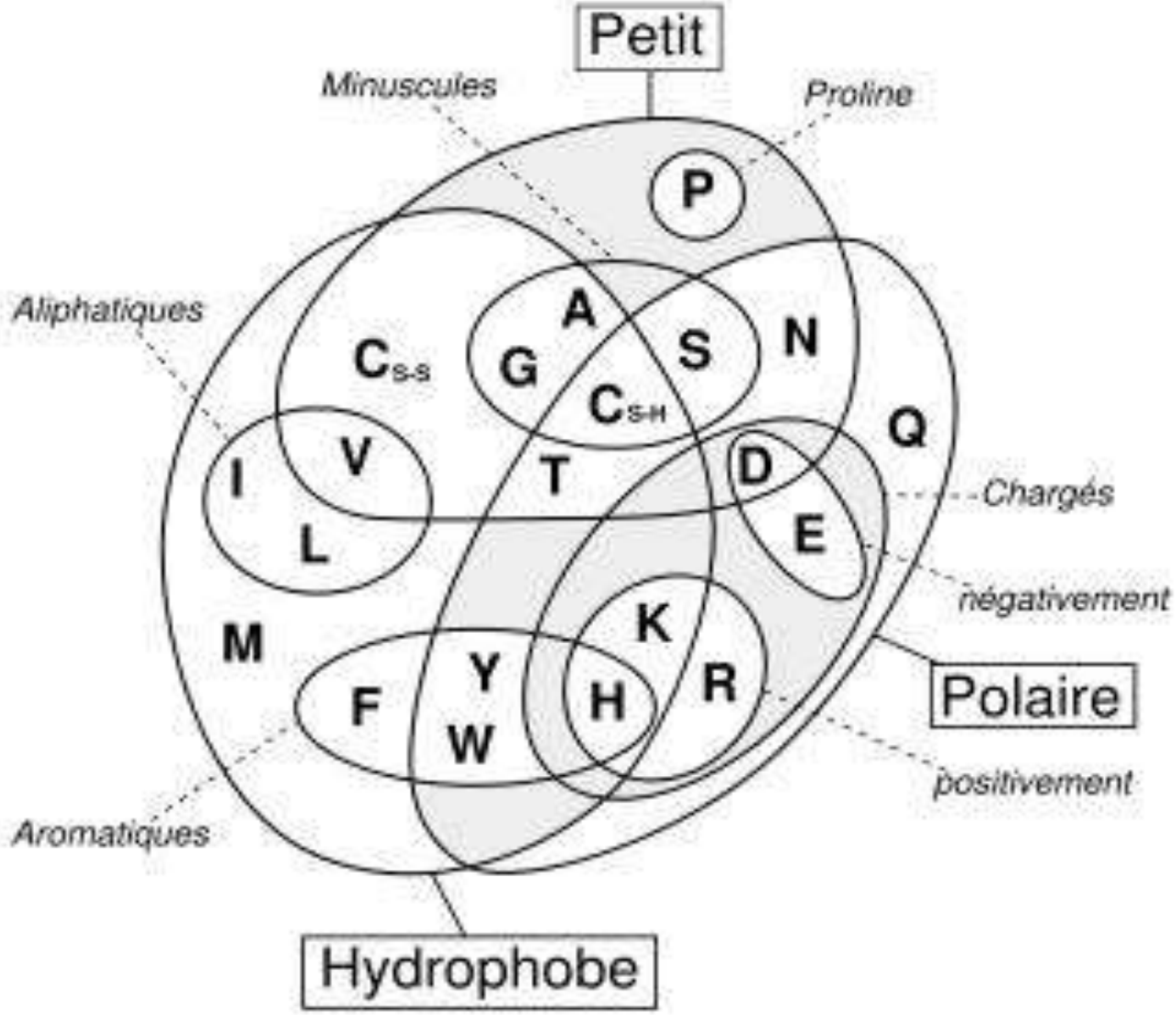
Nombre plus élevé des acides aminés (20 acides aminés et non 4 comme le cas des nucléotides)

Nature physico-chimiques de ceux-ci.

Certains acides aminés peuvent être substitués par d'autres sans altérer la fonction biologique de la protéine et ce à cause de leurs propriétés **physicochimiques** ou **stériques** très proches.

Les acides aminés de même classe peuvent être substitués par simple mutation **acceptable** et répondre ainsi aux contraintes de la sélection évolutive.

Il en découle alors des structures protéiques non identiques mais **similaires** assurant la même **fonction** biologique.


Petit
Minuscules
Proline
P
Aliphatiques
A
G
S
N
C_{S-S}
C_{S-H}
Q
V
I
T
D
Chargés
L
E
négativement
K
M
Y
F
H
R
Polaire
W
positivement
Aromatiques
Hydrophobe

Afin d'aligner les protéines, des matrices dites de **substitution** sont mises à disposition.

Ces matrices peuvent être regroupées en deux catégories :

•Une catégorie qui regroupe les matrices liées à l'évolution. Ces matrices ont été mises au point suite aux études montrant le caractère de substitution (mutation) des acides aminés au cours de l'évolution. Elles représentent les **substitutions possibles et acceptables** d'un acide aminé par un autre lors de l'évolution des protéines.

•La deuxième est basée plus particulièrement sur les caractéristiques physicochimiques
•des acides aminés.

Les matrices plus connues et fréquemment utilisées pour aligner des protéines sont :
la matrice **PAM** et la matrice **BLOSUM**.

Ces matrices utilisent des scores basés sur la comparaison entre la fréquence observée des substitutions et leur fréquence attendue.

La différence entre les deux types de matrices vient du jeu de données sur lequel les fréquences sont observées.

Application de l'algorithme de Needleman et Wunsch dans le cas des protéines

Dans le cas où l'utilisateur ne souhaite pas utiliser une des matrices de substitutions décrites précédemment (PAM, BLOSUM..etc.) Il est primordial de définir au préalable 3 scores (**identité**, **substitution** et **gap**) afin d'appliquer convenablement l'algorithme de **Needleman et Wunsch.**

$$S(i,j) = \text{Max} \begin{cases} S(i-1,j-1) + se(i,j) \\ S(i-1,j) + gap \\ S(i,j-1) + gap \end{cases}$$

Exemple d'alignement : On considère les deux séquences suivantes :
S1 = MPRCLCQR
S2 = PYRCKCR

Aligner ces deux séquences en utilisant l'algorithme de **Needleman et Wunsch**

On admet que le score **d'identité = 3**, le score de **substitution = -1** et le score de **gap = -2**

1ere étape : Construction de la matrice initiale

Remplir la matrice initiale en mettant comme score élémentaire **3** si les deux acides aminés sont identiques et **-1** pour une substitution.

	M	**P**	**R**	**C**	**L**	**C**	**Q**	**R**
P	-1	3	-1	-1	-1	-1	-1	-1
Y	-1	-1	-1	-1	-1	-1	-1	-1
R	-1	-1	3	-1	-1	-1	-1	3
C	-1	-1	-1	3	-1	3	-1	-1
K	-1	-1	-1	-1	-1	-1	-1	-1
C	-1	-1	-1	3	-1	3	-1	-1
R	-1	-1	3	-1	-1	-1	-1	3

2ème étape : construction de la matrice transformée

Dans un deuxième temps, il faut créer une deuxième matrice à i+2 colonnes et j+2 lignes, dans laquelle la 1ère ligne et la 1ère colonne seront initialisées non pas à zéro mais en utilisant le score des gaps (-2) comme suit :

i \ j		1	2	3	4	5	6	7	8
		M	**P**	**R**	**C**	**L**	**C**	**Q**	**R**
1	0	-2	-4	-6	-8	-10	-12	-14	-16
2	**P**	-2							
3	**Y**	-4							
4	**R**	-6							
5	**C**	-8							
6	**K**	-10							
7	**C**	-12							
	R	-14							

C'est à ce niveau qu'on applique l'algorithme de **Needleman-Wunsch**

$$S(i,j) = \text{Max} \begin{cases} S(i-1,j-1) + se(i,j) \\ S(i-1,j) + gap \\ S(i,j-1) + gap \end{cases}$$

Si on commence par la case (1,1), l'algorithme est appliqué comme suit :

$$S(1,1) = \text{Max} \begin{cases} S(0,0) + se(1,1) = 0 + -1 = -1 \\ S(0,1) + -2 = -4 \\ S(1,0) + -2 = -4 \end{cases}$$

$$S(i,j) = \text{Max} \begin{cases} S(i-1,j-1) + se(i,j) \\ S(i-1,j) + gap \\ S(i,j-1) + gap \end{cases}$$

i \ j		1 M	2 P	3 R	4 C	5 L	6 C	7 Q	8 R
	0	-2	-4	-6	-8	-10	-12	-14	-16
1 P	-2	-1	1	-1	-3	-5	-7	-9	-11
2 Y	-4	-3	-1	0	-2	-4	-6	-8	-10
3 R	-6	-5	-3	2	0	-2	-4	-6	-5
4 C	-8	-7	-5	0	5	3	1	-1	-3
5 K	-10	-9	-7	-2	3	4	2		
6 C	-12								
7 R	-14								

	j	1	2	3	4	5	6	7	8
		M	P	R	C	L	C	Q	R
i	0	-2	-4	-6	-8	-10	-12	-14	-16
1	-2	-1	1	-1	-3	-5	-7	-9	-11
2	-4	-3	-1	0	-2	-4	-6	-8	-10
3	-6	-5	-3	2	0	-2	-4	-6	-5
4	-8	-7	-5	0	5	3	1	-1	-3
5	-10	-9	-7	-2	3	4	2	0	-2
6	-12	-11	-9	-4	1	2	7	5	3
7	-14	-13	-11	-6	-1	0	5	6	8

P	
Y	
R	
C	
K	
C	
R	

3ème étape : traceback (traçage en arrière)

Le parcours de la matrice transformée commence par le plus haut score, vers le plus haut score parmi les trois cases (i-1, j-1) (i-1,j) et (i, j-1) et ainsi de suite jusqu'à la case (1,1). Dans cet exemple, on commence par la case (7,8) ayant le plus haut score = 8. Dans ce cas, les scores des 3 cases (6,7) (6,8) et (7,7) sont respectivement 5, 3 et 6. Donc, le parcours de la matrice sera vers la case (7,7) où le score est le plus grand soit 6. Le parcours final de la matrice transformée est le suivant :

		j	1	2	3	4	5	6	7	8
			M	**P**	**R**	**C**	**L**	**C**	**Q**	**R**
i		0	-2	-4	-6	-8	-10	-12	-14	-16
1	**P**	-2	-1	1	-1	-3	-5	-7	-9	-11
2	**Y**	-4	-3	-1	0	-2	-4	-6	-8	-10
3	**R**	-6	-5	-3	2	0	-2	-4	-6	-5
4	**C**	-8	-7	-5	0	5	3	1	-1	-3
5	**K**	-10	-9	-7	-2	3	4	2	0	-2
6	**C**	-12	-11	-9	-4	1	2	7	5	3
7	**R**	-14	-13	-11	-6	-1	0	5	6	8

4ème étape : génération de l'alignement et calcul des scores

S1	M	P	–	R	C	L	C	Q	R
	*	I	*	I	I	:	I	*	I
S2	–	P	Y	R	C	K	C	–	R

le premier et le deuxième acides aminés de la séquence S1 (M et P respectivement), sont en horizontal dans le parcours de la matrice transformée. Cela signifie que soit la séquence S1a subit une insertion de l'acide aminé M (car M a un plus faible score par rapport à P) ou alors la séquence S2 à subit une délétion de cette acide aminé au cours de l'évolution.

Calcul des scores :

Le pourcentage d'identité (%id) = (nombre d'identités / taille de la séquence après alignement)

%id= (5/9)* 100 = 55.55%

Le pourcentage des substitutions : (nombre de substitutions / taille de la séquence après alignement)*100

= (1/9)*100= 11.11%.

Le pourcentage des gaps : = (nombre gaps / taille de la séquence après alignement)*100

= (3/9)*100= 33.33%

Score d'alignement : (nombre d'identités * score d'identité) +
(nombre de substitutions*score des substitutions) + (nombre de gaps * score des gaps)

(5*3)+(1*-1)+(3*-2)=8

Matrices de substitution

Matrice de type PAM (Point Accepted Mutation ou Mutation Ponctuelle Acceptées

En 1978, Margret Dayhoff réalise des alignements de 1300 séquences protéiques appartenant à 71 familles très semblables (> 85% d'identité) .

Par la suite, il a compté le nombre de substitutions et d'identités entre chaque paire d'acide aminé aligné.

Son travail a montré que certaines mutations accumulées au cours de l'évolution n'ont pas altérées la fonction biologique des protéines proches phylogénétiquement

Ce type de matrice donne la probabilité que, suite à une mutation par substitution au cours de l'évolution, un acide aminé quelconque peut remplacer un autre acide aminé sans altérer la fonction de la protéine

Les matrices de type PAM fonctionnent bien avec des séquences phylogénétiquement proches

Actuellement, il existe plusieurs matrices de type PAM parmi lesquelles **PAM250**

Cette matrice donne la probabilité que 250 mutations soit acceptées pour 100 acides aminés.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2																			
R	-2	6																		
N	0	0	2																	
D	0	-1	2	4																
C	-2	-4	-4	-5	4															
Q	0	1	1	2	-5	4														
E	0	-1	1	3	-5	2	4													
G	1	-3	0	1	-3	-1	0	5												
H	-1	2	2	1	-3	3	1	-2	6											
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5										
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6									
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5								
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6							
F	-4	-4	-4	-6	-4	-5	-5	-5	-2	1	2	-5	0	9						
P	1	0	-1	-1	-3	0	-1	-1	0	-2	-3	-1	-2	-5	6					
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	3				
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-2	0	1	3			
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17		
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	2	4

Une valeur faible dans cette matrice (exemple : W / C = -8) signifie qu'il est peu probable d'observer la substitution d'un tryptophane par une cystéine sans perte significative de la fonction de la protéine.

Au contraire, une valeur forte (exemple : Y / F = 7) signifie qu'il est probable d'observer la substitution d'une tyrosine par une phénylalanine.

La diagonale reflète le taux de conservation des résidus. Notons que certains résidus rares ont un score de conservation très important.

Il existe d'autres matrices de type PAM, parmi lesquelles **PAM001, PAM050, PAM100…etc.**

Le choix du type de matrice est primordiale pour réussir l'alignement.

Le taux de différence entre les deux séquences à aligner doit être pris en considération avant de choisir la matrice.

Matrice de type BLOSUM (***BLO****cks of Amino Acid* ***SU****bstitution* ***M****atrix*)

Ces matrices ont été développées par Henikoff & Henikoff à partir de 2000 BLOCKS d'alignement multiple générés de plus de 500 familles de protéines

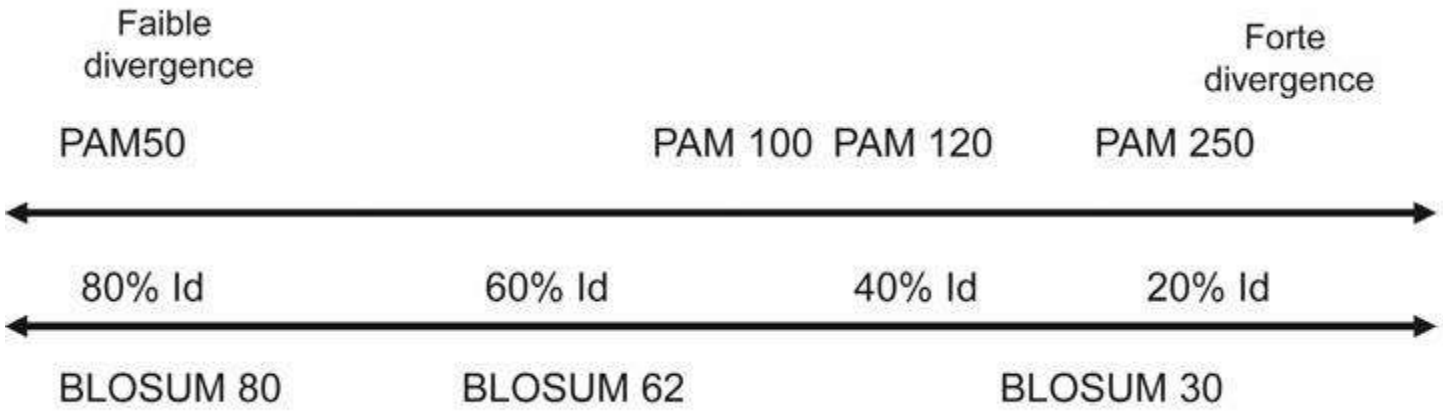
Les blocs ont été obtenus par alignement des **régions conservées des familles de protéines mais ne contenant pas d'insertions ou de délétions**.

Les séquences utilisées pour créer ces blocs sont regroupées (clustérisées) si leur identité dépasse un certain seuil dont découle la matrice BLOSUM

Dans une matrice BLOSUM62, les séquences présentant plus de 62% d'identité ont été regroupées ensemble

La matrice BLOSUM80 a été calculée sur des blocs de >=80% d'identité

Les matrices BLOSUM sont le type de matrice par défaut du logiciel "*Blastp*".



Un indice élevé pour une matrice PAM décrit une distance d'évolution élevée (similarité faible)

Un indice élevé pour une matrice BLOSUM décrit au contraire une forte similarité de séquences donc une distance d'évolution faible