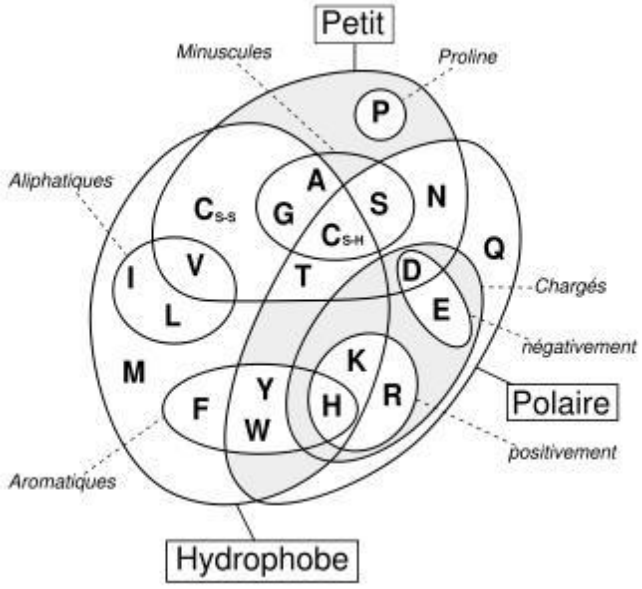
Alignement des séquences protéiques

Les matrices utilisées pour aligner les protéines sont totalement différentes de celles des acides nucléiques et ce en raison du nombre plus élevé des acides aminés (20 acides aminés et non 4 comme le cas des nucléotides) et de la nature physico-chimiques de ceux-ci. En effet, 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 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.



Afin d'aligner les protéines, des matrices dites de substitution sont mises à disposition. D'ailleurs, le site web https://www.genome.jp/aaindex/ récence actuellement 94 matrices de substitution pouvant 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. Les deux types de matrice 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.

Matrice de type PAM

Les matrices de type PAM (**point Accepted Mutation)**, ont été mises en place par le cherchreur **Margaret Dayhoff** après l'alignement d'environ 1300 séquences très semblables (> 85% d'identité) appartenant à 71 familles de protéines. Sont travail a montré que certaines mutations accumulées au cours de l'évolution n'ont pas altérées la fonction biologique des protéines proches *phylogénétiquement. Autrement dit,* 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 que la fonction de la protéine ne soit altérée, d'où la terminologie "**Mutations Ponctuelles Acceptées**". *Par conséquence,* les matrices PAM fonctionnent bien sur ce pourquoi elles ont été conçues : des séquences *phylogénétiquement* proches. Pour des séquences qui sont plus éloignées au sens de l'évolution, les matrices PAM fonctionnent moins bien.

Actuellement, il existe plusieurs matrices de type PAM parmi lesquelles on cite : **PAM250 (représentée ci-dessous).** Cette matrice donne la probabilité que 250 mutations soit acceptées pour 100 acides aminés. 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.

	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

Matrice PAM250

Matrice de type BLOSUM

Les matrices de type BLOSUM (***BLO****cks of Amino Acid* ***SU****bstitution* ***M****atrix*), ont été développées par Henikoff & Henikoff à partir de 2000 BLOCS d'alignement multiple générés de plus de 500 familles de protéines. Ces blocs ont été obtenus par alignement des **régions conservées de familles de protéines mais ne contenant pas d'insertions ou de délétions**. Puis à partir d'un ensemble de blocs est constitué un sous-ensemble qui contient les portions de séquences qui révèlent un pourcentage donné d'identité. Autrement dit, les séquences utilisées pour créer ces blocs sont regroupées (clustérisées) si leur identité dépasse un certain seuil dont découle la matrice BLOSUM. Ainsi, dans une matrice BLOSUM62, les séquences présentant plus de 62% d'identité ont été regroupées ensemble. Les matrices BLOSUM sont le type de matrice par défaut du logiciel "*Blastp*".