




INTRODUCTION : ce sont les mutations qui, au cours de l'évolution naturelle, causent des erreurs au moment de la réplication de l'ADN ; car l'évolution se fait par mutations successives. Ces erreurs peuvent être :

- Des substitutions (changement ponctuel d'un nucléotide par un autre). On parle de transition ou de transversion,
- des insertions (ajout d'un ou plusieurs nucléotides),
- ou alors des délétions (délétion d'une base ou d'un segment d'ADN).

Ces mutations n'affectent pas de la même manière les différentes espèces par rapport au site de mutation et par rapport au temps pendant lequel va se manifester cette mutation:

	Mutations (substitutions) par site et par milliard d'années	
	Homme	Souris
Histone 3	0	6,4
Interleukine-1	1,4	4,6
Interferon-γ	2,8	8,6

Il en découle alors des différences, plus ou moins importantes, dans les structures (primaire, secondaire, ...) de ces séquences, d'où la divergence et la biodiversité des espèces. De ce fait, deux notions évolutionnistes importantes ont vu le jour : **l'homologie et la similarité**. Par exemple, la protéine de l'insuline de l'homme, de la souris et le dègue du Chili (*Octodon degus*) exercent la même fonction biologique mais ont des structures primaires qui ne sont pas totalement identiques !

<p><i>Homo sapiens</i>² AAA59172 110 acides aminés</p>		<p>MALWMRLLPLLALLALWGPDPAAAF VNQHLCGSHLVEALYLVCGERGFFY TPKTRREAEDLQVGGQVELGGGPGAG SLQPLALEGSLQKRGIVEQCCTSICSL YQLENYCN</p>
<p><i>Mus musculus</i>³ AAI45871 108 acides aminés</p>		<p>MALLVHFLPLLALLALWEPKPTQAF VKQHLCGPHLVEALYLVCGERGFFY TPKSRREVEDPQVEQLELGGSPGDL QTLALEVARQKRGIVDQCCTSICSL YQLENYCN</p>
<p><i>Octodon degus</i>⁴ AAA40590 109 acides aminés</p>		<p>MAPWMHLLTVLALLALWGPNSVQAY SSQHLCGSNLVEALYMTCGRSGFYR PHDRRELEDLQVEQAELGLEAGGLQ PSALEMILQKRGIVDQCCNNICTFN QLQNYCNVP</p>

² <http://mimi40.m.i.pic.centerblog.net/c5xtto8q.jpg>

³ http://fr.academic.ru/pictures/frwiki/65/Apodemus_sylvaticus_bosmuis.jpg

⁴ <http://www.wildanimalsonline.com/mammals/degus-degus.jpg>

En bioinformatique, la comparaison des séquences (ADN, ARN et/ou protéines : ARNm, régions 5'UTR, les EST, des clones, ...) repose essentiellement sur la notion de l'**alignement**, et permet de déterminer le degré de ressemblance entre celles-ci (similitude ou identité en révélant des régions proches dans leurs séquences primaires). Cela peut alors indiquer que :

- La structure (primaire, secondaire ou tertiaire) des deux séquences est semblable,
- La fonction biologique est proche ou différente (dans le cas de la dissimilarité),
- L'origine des séquences alignées est commune ou éloignée (notion d'homologie), ...

Cependant, la comparaison pour l'obtention d'un alignement optimal entre deux séquences biologiques, nécessite néanmoins la mise en œuvre de procédures de calcul (algorithmes) et de modèles biologiques permettant de quantifier la notion de ressemblance entre ces séquences.

TRAITEMENT DES SEQUENCES NUCLEIQUES (ADN ou ARN)

Il existe différentes méthodes pour la détermination de segments identiques entre deux séquences biologiques (on parle alors de fenêtres, de motifs ou de mots) sur lesquelles une similitude significative peut exister.

Notion de score : Le score élémentaire (noté "s") est une entité numérique que l'on attribue à chaque couple de nucléotides des deux séquences à comparer. Il prend la valeur de 1 lorsque les deux nucléotides des deux séquences sont identiques, et la valeur de zéro sinon.

Exemple :

Séquence1	A	G	C	T	A	C	C	T	G	T	Score global : Total des scores
Séquence2	A	A	G	T	A	G	C	T	T	T	
Point de comparaison	1	2	3	4	5	6	7	8	9	10	
Score élémentaire (s)	1	0	0	1	1	0	1	1	0	1	

Dans cet exemple, constatez qu'au niveau du premier point de comparaison (ou site de comparaison), les deux séquences contiennent le même nucléotide A, donc le score élémentaire (s) à ce point prend la valeur de 1 (s = 1).

Au deuxième point de comparaison, la séquence 1 contient un G et la séquence 2 contient un A. Elles sont donc différentes en ce point d'où un score élémentaire de zéro (s = 0)...

Au 10^{ème} point de comparaison, les deux séquences contiennent le même nucléotide T donc un score élémentaire de 1.

Constatons que la somme des scores élémentaires est égale à six (s = 6). Donc il y a six points identiques entre les deux séquences ; soit 60% d'identité entre les deux séquences $([6/10] \times 100)$. On dit alors que le score global entre les deux séquences est égal à six. Le score a donc permis de quantifier la ressemblance entre les deux séquences.

La relation entre le score global (S) et les scores élémentaires (s) pour deux séquences est de la forme :

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

Question : Pourquoi avons-nous affecté la valeur de 1 dans le cas de l'identité et zéro dans le cas contraire ?

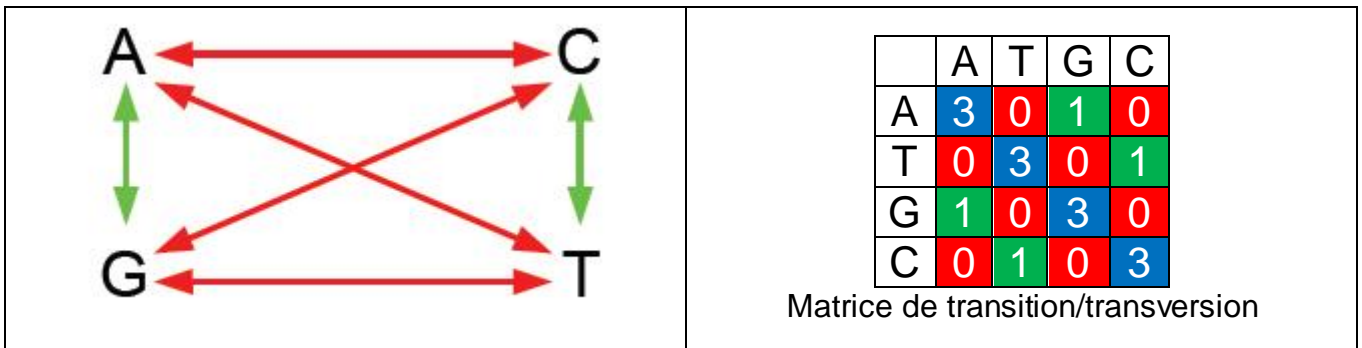
Il faut savoir qu'il existe une matrice (**matrice d'identité**) qui donne les valeurs de scores d'identité entre les séquences à comparer. Dans cette matrice, on attribue la valeur de 1 lorsque les deux nucléotides sont identiques et zéro s'ils ne le sont pas.

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

Matrice d'identité nucléique

Il existe une autre matrice de score, qui tient compte de l'analogie structurale entre purines (A et G) et pyrimidines (C, T et U) et affecte des scores en fonction de cette ressemblance : C'est la matrice de transition/transversion :

La substitution entre purines d'une part, et entre pyrimidines d'autre part est pondérée et n'a pas de score élémentaire nul au moment de la comparaison des séquences :



Remarque : Quelle matrice utiliser ?

En bioinformatique, on utilise beaucoup plus la matrice d'identité.

Remarque : Dans le cas où les deux séquences sont complètement identiques, le résultat est une diagonale :

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

La méthode du Dot-Plot

Le dot-plot est utile pour déterminer de combien d'exons est composé un gène en le comparant à son ARNm et pour avoir une idée de la taille des introns et des exons.

Il existe un logiciel de dotplot interactif, Dotlet qui nécessite JAVA. Si JAVA n'est pas installé sur vos machines, vous pouvez utiliser Dottup⁵.

Le principe du dot-plot est basé sur la comparaison de fenêtres de longueur fixe que l'on déplace le long des séquences.

Soit deux séquences A et B à comparer et l la longueur de la fenêtre. On détermine sur la séquence A une première fenêtre de longueur l que l'on va comparer avec toutes les fenêtres possibles de même longueur, obtenues à partir de la séquence B. Un incrément est alors appliqué pour déterminer une deuxième fenêtre sur la séquence A, puis l'on recommence le balayage des comparaisons sur la séquence B. Si l'on choisit un incrément de 1 et que les séquences ont respectivement une longueur de m et n éléments, on effectuera de l'ordre de $n \times m$ comparaisons de fenêtres différentes.

Pour chaque comparaison entre deux fenêtres, un score est obtenu et l'on mémorisera uniquement les comparaisons dont les scores sont jugés significatifs, c'est-à-dire supérieurs ou égaux à un seuil que l'on s'est fixé. Par exemple lorsque le score correspond au minimum à 80% d'identité avec l'utilisation d'une matrice unitaire nucléique comme matrice de scores élémentaires⁶.

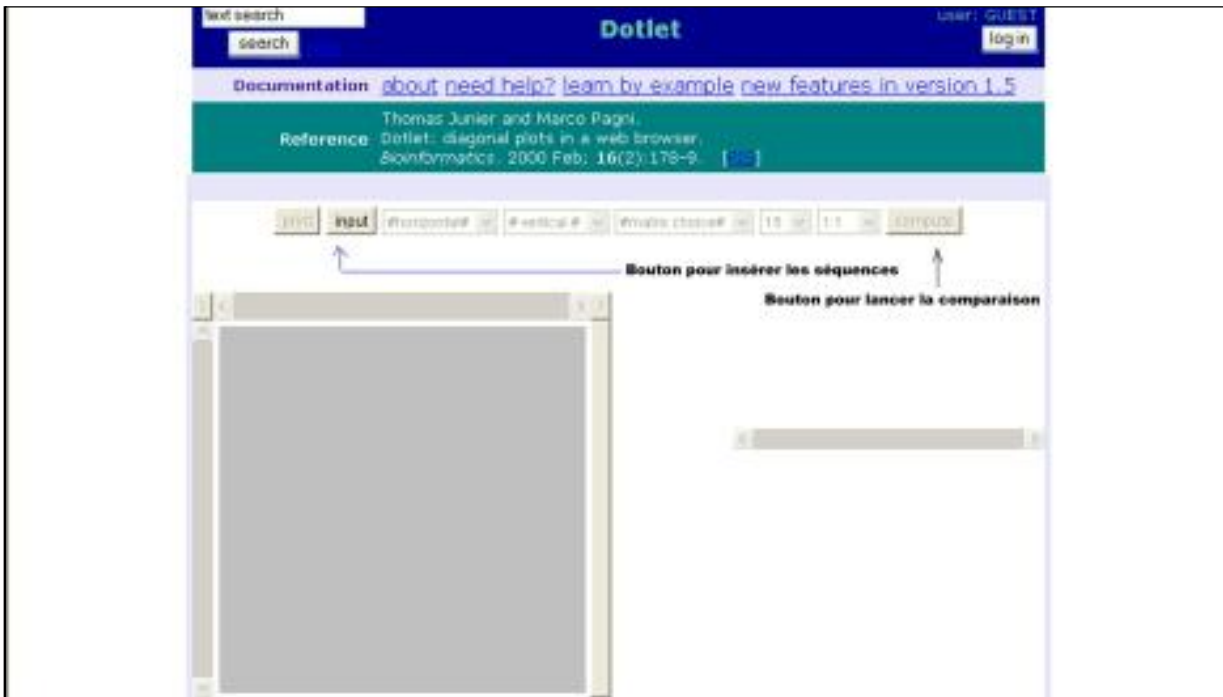
Considérons, par exemple, les deux séquences A et B suivantes :

Séq A = ATGTAATGCATG et Séq B = TATGTGAATG. La taille du motif (fenêtre) étant choisie égale à 5.

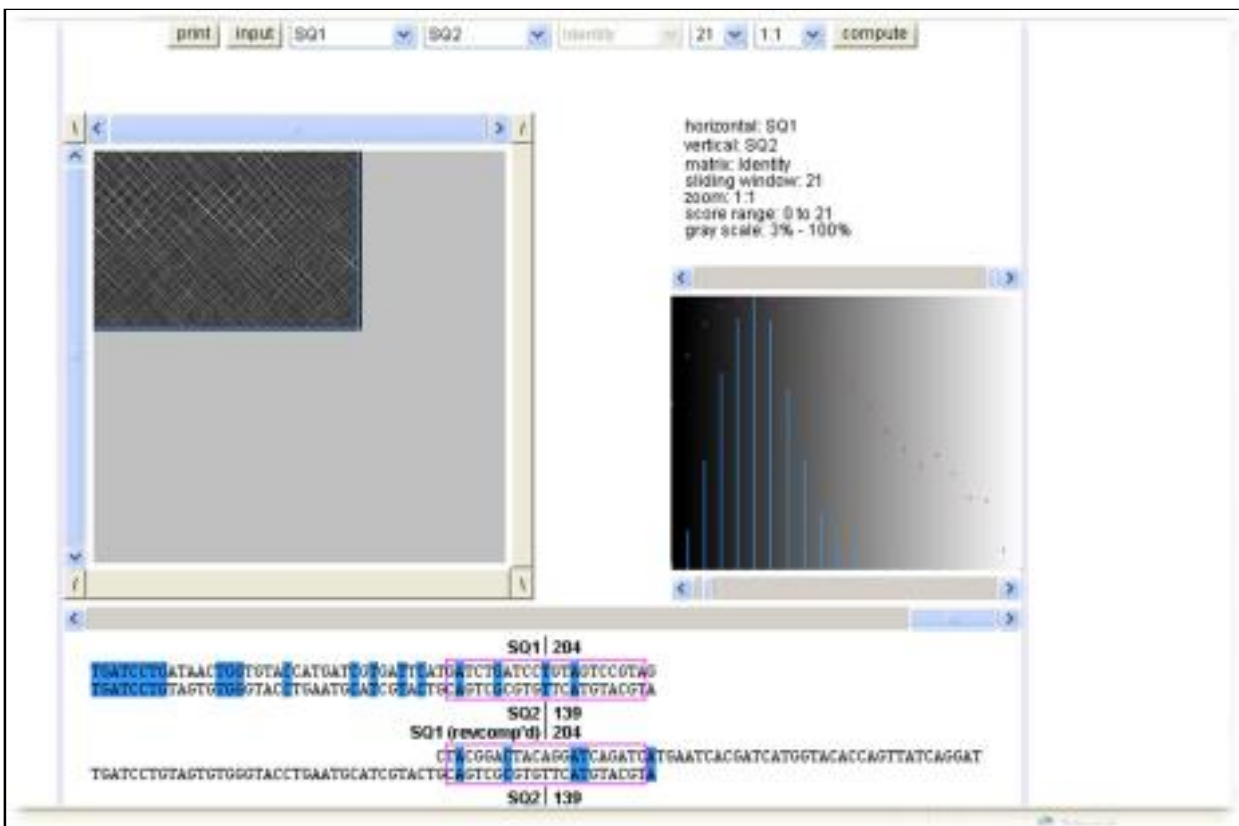
⁵ http://www.fil.univ-lille1.fr/~pupin/MRBS/comp_seq.html

⁶ http://genet.univ-tours.fr/gen001400_fichiers/chap5/genach5ec9.htm

Le programme dotlet⁷ permet de réaliser cette comparaison en ligne :



Une fois avoir inséré les deux séquences, le résultat obtenu sera :



⁷ <http://myhits.isb-sib.ch/cgi-bin/dotlet>

Recherche de segments identiques par codification numérique

Dans cette méthode, les séquences sont transformées en une suite d'entiers (ou codes) après les avoir décomposées en petits segments de taille fixe (taille = 3 nucléotides par exemple). Ces segments sont appelés motifs :

On considère la séquence suivante :

TATGCCT

A une position notée x de la séquence, le motif M_x de taille n ($n= 3$ nucléotides par exemples) aura pour code C_x de sorte que :

$$C_x = P_1x4^{(n-1)} + P_2x4^{(n-2)} + P_3x4^{(n-3)} + \dots + P_ix4^{(n-i)} + \dots + P_n + 1$$

$P_i = 0, 1, 2, 3$ respectivement pour les nucléotides A, C, G et T

Pour le cas où $n = 3$ nucléotides, les triplets qui seront codés à partir de la séquence sont :

TAT, ATG, TGC, GCC, CCT

- Le premier motif étant **TAT** il aura pour code :

$$C_1 = P_1x4^{(3-1)} + P_2x4^{(3-2)} + P_3x4^{(3-3)} + 1, P_1=3 \text{ (car le nucléotide est T), } P_2 = 0 \text{ (car le nucléotide est A) et } P_3 = 3 \text{ (car le nucléotide est T)}$$

D'où

$$C_1 = P_1x4^{(2)} + P_2x4^{(1)} + P_3 + 1$$

$$C_1 = 3x4^{(2)} + 0x4^{(1)} + 3 + 1 = 52$$

C1 = 52 correspond au triplet TAT

- Le motif **ATG** :

$$C_2 = P_1x4^{(3-1)} + P_2x4^{(3-2)} + P_3x4^{(3-3)} + 1, P_1=0 \text{(car le nucléotide est A), } P_2 = 3 \text{ (car le nucléotide est T) et } P_3 = 2 \text{ (car le nucléotide est G)}$$

D'où

$$C_2 = P_1x4^{(2)} + P_2x4^{(1)} + P_3 + 1$$

$$C_2 = 0x4^{(2)} + 3x4^{(1)} + 2 + 1 = 15$$

C2 = 15 correspond au triplet ATG

- Le motif **TGC** :

$$C_3 = P_1x4^{(3-1)} + P_2x4^{(3-2)} + P_3x4^{(3-3)} + 1, P_1=3 \text{(car le nucléotide est T), } P_2 = 2 \text{ (car le nucléotide est G) et } P_3 = 1 \text{ (car le nucléotide est C)}$$

D'où

$$C3 = P_1 \times 4^{(2)} + P_2 \times 4^{(1)} + P_3 + 1$$

$$C3 = 3 \times 4^{(2)} + 2 \times 4^{(1)} + 1 + 1 = 62$$

C3 = 62 correspond au triplet ATG

- **Le motif GCC :**

$$C4 = P_1 \times 4^{(3-1)} + P_2 \times 4^{(3-2)} + P_3 \times 4^{(3-3)} + 1, P_1=2(\text{car le nucléotide est G}), P_2=1(\text{car le nucléotide est C}) \text{ et } P_3=1(\text{car le nucléotide est C})$$

D'où

$$C4 = P_1 \times 4^{(2)} + P_2 \times 4^{(1)} + P_3 + 1$$

$$C4 = 2 \times 4^{(2)} + 1 \times 4^{(1)} + 1 + 1 = 38$$

C4 = 38 correspond au triplet ATG

- **Le motif CCT :**

$$C5 = P_1 \times 4^{(3-1)} + P_2 \times 4^{(3-2)} + P_3 \times 4^{(3-3)} + 1, P_1=1(\text{car le nucléotide est C}), P_2=1(\text{car le nucléotide est C}) \text{ et } P_3=3(\text{car le nucléotide est T})$$

D'où

$$C5 = P_1 \times 4^{(2)} + P_2 \times 4^{(1)} + P_3 + 1$$

$$C5 = 1 \times 4^{(2)} + 1 \times 4^{(1)} + 3 + 1 = 24$$

C5 = 24 correspond au triplet ATG

Séquence	T	A	T	G	C	C	T
1 ^{er} motif	T	A	T				
2 ^{ème} motif		A	T	G			
3 ^{ème} motif			T	G	C		
4 ^{ème} motif				G	C	C	
5 ^{ème} motif					C	C	T
Code Cx	52	15	62	38	24		

La deuxième séquence est : AGATGCC. Le résultat de son transcodage donne :

Séquence	A	G	A	T	G	C	C
1 ^{er} motif	A	G	A				
2 ^{ème} motif		G	A	T			
3 ^{ème} motif			A	T	G		
4 ^{ème} motif				T	G	C	
5 ^{ème} motif					G	C	C
Code Cx	9	36	15	62	38		

Comparons les deux séquences selon leur transcodage :

Séquence 1	52	15	62	38	24				
Séquence 2	9	36	15	62	38				
Résultat	Il y a 3 motifs identiques entre les deux séquences : 15, 62 et 38								

Remarque : la comparaison peut également se faire sur des motifs de tailles plus grandes.

L'alignement des séquences nucléiques: La programmation dynamique

Pourquoi vouloir réaliser des alignements ?

L'alignement, comme nous allons le voir dans les exemples suivants, permet de mesurer la similitude entre les séquences. S'il y a similitude, cela signifie qu'il est possible que les deux séquences présentent la même fonction biologique, ou du moins les deux séquences présente une structure fortement similaire. Ce type d'information est nécessaire dans la mesure où, généralement, nous avons à faire à une séquence inconnue. Sa comparaison avec des séquences de structure et de fonction connues permet de tirer un maximum d'informations quant à la structure et la fonction de la séquence inconnue.

Dans certains cas, on peut même confirmer si la séquence inconnue est un gène ou une portion de gène après l'avoir aligné avec des séquences de structure génique connue (régions codantes : codons d'initiation et de terminaison, sites d'épissage, zones de fixation des ribosomes).

L'algorithme de Needleman et Wunsch : Il permet de réaliser un alignement global entre deux séquences nucléiques. Son expression est de la forme :

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

Exemple :

Supposons que nous désirons calculer un alignement global des deux séquences suivantes de taille m et n respectivement:

S1 = TAAGTCCG m=8 et S2 = TACGTACG n=8

Remarque : Ici, les deux séquences sont de même longueur (8 résidus chacune). On peut calculer un alignement entre deux séquences de tailles inégales.

Pour calculer l'alignement entre les deux séquences S1 et S2, quatre étapes sont nécessaires :

Etape 1 : Calcul de la matrice initiale

Il s'agit d'insérer les deux séquences S1 et S2 dans une matrice de sorte que S1 soit à l'horizontal et S2 à la verticale du tableau, puis remplir les cases par 1 (identité des deux nucléotides de S1 et de S2) ou 0 (sinon) :

	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

Etape 2 : Calcul de la matrice transformée : Initialisation de la matrice

Construisons une nouvelle matrice à m+2 colonnes et n+2 lignes, dans laquelle la 1^{ère} ligne et la 1^{ère} colonne seront initialisées à zéro :

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

L'application de l'algorithme de Needleman et Wunsh permet de remplir les cases de cette matrice. Le résultat est le suivant :

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

Etape 3 : Parcours de la matrice transformée

Parcourir la matrice transformée depuis le plus haut score calculé (ici s=6) jusqu'au score le plus petit (ici s=1)

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

Remarque : le développement de cette procédure de calcul et de parcours aura lieu en séance des travaux dirigés.

Etape 4 : Alignement des deux séquences et calcul de score

Séquence S1	T	A	A	G	T	—	C	C	G
						*		*	
Séquence S2	T	A	A	G	T	A	C	—	G

Exploitation de l'alignement :

- Le score global de cet alignement est de 7.
- Le pourcentage de l'identité entre les deux séquences S1 et S2 est : $\%id = (7/9) \times 100 = 77,78\%$
- Le trou retrouvé entre les nucléotides T et C de la séquence S1 est un GAP ou InDel : il signifie qu'à ce point, la séquence S1 a subi une mutation par **DELétion** au cours de laquelle le nucléotide A est perdu par nécessité évolutive et d'adaptation à l'environnement ; en même temps, il est conservé dans la séquence S2 (à la 6^{ème} position face au gap de S1). Comme on peut supposer que c'est la séquence S2 qui a subi une mutation par **INsertion** du nucléotide A par nécessité adaptative. Dans un cas ou dans l'autre une des deux séquences a subi une mutation (**INsertion** ou **DELétion**) ; ce point est appelé **INDEL** pour dire INSERTION dans la séquence S2 ou DELETION dans la séquence S1. La même interprétation concerne le deuxième gap retrouvé 8^{ème} position : il s'agit d'une délétion du nucléotide C dans la séquence S2 ou de l'insertion de C dans la séquence S1.

Un autre exemple d'alignement : S1 = CAATGGCCGA et S2 = CATTGGCCG

Etape 1 : Calcul de la matrice initiale

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

Etape 2 : Calcul de la matrice transformée : Initialisation de la matrice

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

Etape 3 : Parcours de la matrice transformée

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

Etape 4 : Alignement des deux séquences et calcul du score

Avant de procéder à l'alignement, à proprement dire, revenons à la matrice de l'étape 3 :

		C	A	A	T	G	G	C	C	G	A
	0	0	0	0	0	0	0	0	0	0	0
C	0	1	1	1	1	1	1	1	1	1	1
Délétion après ce nucléotide	A	0	1	2	2	2	2	2	2	2	2
T	0	1	2	2	3	3	3	3	3	3	3
T	0	1	2	2	3	3	3	3	3	3	3
G	0	1	2	2	3	4	4	4	4	4	4
G	0	1	2	2	3	4	5	5	5	5	5
C	0	1	2	2	3	4	5	6	6	6	6
C	0	1	2	2	3	4	5	6	7	7	7
G	0	1	2	2	3	4	5	6	7	8	8

Vous constatez que le 2^{ème} nucléotide de la séquence S2 (c'est à dire le nucléotide A) correspond à deux nucléotides de la séquence S1 qui sont A et A ; or il ne doit correspondre qu'à un seul nucléotide. La deuxième case (couleur verte) en face du nucléotide A de la séquence S2 correspond donc à une mutation : Cela suppose qu'une **DELETION** a eu lieu après le nucléotide A de S2.

Séquence S1	C	A	A	T	—	G	G	C	C	G	A
			*		*						*
Séquence S2	C	A	—	T	T	G	G	C	C	G	—

- Le score global de cet alignement est de 8.
- Le pourcentage de l'identité entre les deux séquences S1 et S2 est : %id = $(8/11) \times 100 = 72,73\%$



L'alignement des séquences protéiques par la programmation dynamique

Les matrices protéiques : Notons tout d'abord que les matrices protéiques utilisées pour réaliser des alignements sont totalement différentes de celles des acides nucléiques (matrice d'identité et matrice de transition/transversion) et ce en raison du nombre 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, le système nucléaire basé sur l'identité n'est pas approprié pour le cas des systèmes protéiques. Ceci est dû au fait que certains acides aminés peuvent être remplacés par d'autres (à cause de leurs propriétés physicochimiques surtout) sans altérer le rôle et la fonction biologique de la protéine.

On peut donc classer les acides aminés en familles par rapport à leurs propriétés et obtenir ainsi un système de scores qui rend compte de l'affinité des résidus protéiques entre eux. C'est cette affinité qui permet à un acide aminé d'être substitué par un autre, et les deux structures protéiques ne seront pas identiques à ce point où la substitution a eu lieu mais on dira que les deux séquences sont **SIMILAIRES** et la fonction de la protéine reste conservée.

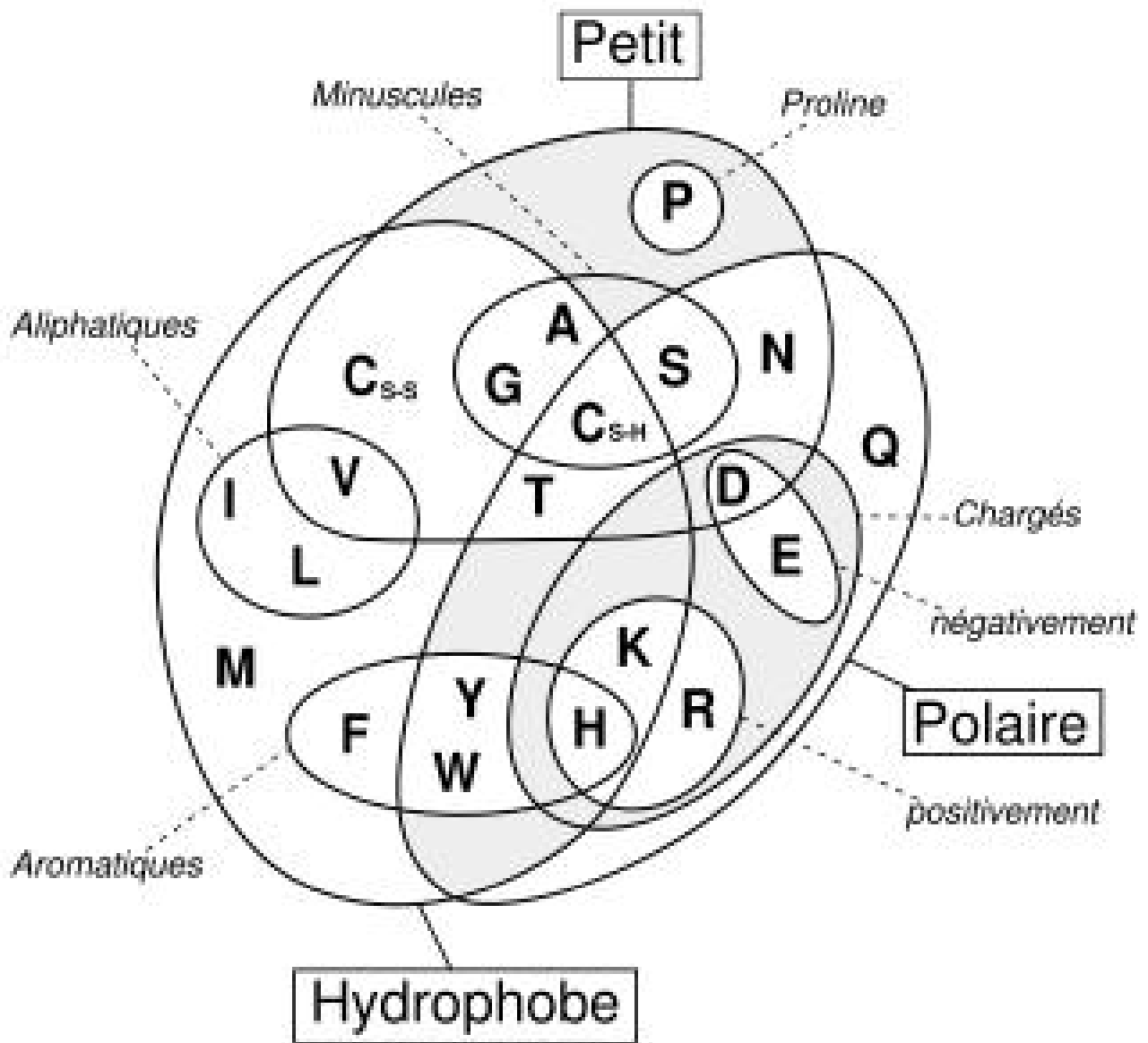
Dans l'exemple qui suit, on dispose de la structure primaire de deux séquences enzymatiques ayant la même fonction biologique, c'est-à-dire toutes les deux sont des amylases. Mais on se rend vite compte que l'amylase de la mouche ne ressemble pas à celle de l'abeille, et pourtant elles assurent toutes les deux l'hydrolyse de l'amidon.

Nom systématique	Numéro d'accèsion	Séquence : structure primaire
<p><i>Mouche⁸ domestica</i> <i>Musca</i></p> 	<p>AAAY88830</p> <p>319 aa</p>	<p>iareceeflaprgfagvqvspvtenvivanrpwweryqpisyklqtrsgtqqefsemcrrcnnvgiriy vdvllnhmaadqyqmavgtagsiadpaaksfpsvpyteldfhatceiwdndryqvqncelvglk dlldqsnewvrclvefldhlvelgvagfrvdaakhmkasdleiiykrvrdlnvdhgfepnsrpfyqev idhghetvskyeynllgavtefqfseeigrafrgnnqlkwlrnwgpqwgflpsdhalvfdndhnqr dggqvlytknskqykmatafalaypygitr imssfdftrdrdppphntne</p>
<p><i>Abeille Apis mellifera⁹</i></p> 	<p>BAA86909</p> <p>493 aa</p>	<p>mmpaivllalltlaageiahndphfapghdaivhlfewkwndiakeceqflgvpvgggvqvspvqe nividkrpwweryqpisykwitrgtreqfidmvarcnkagvriyvdimnhmsgdrndahgtgns rantynfdypqvpytvknfhprcavnnyndpsnrvncelvglhdldqsqeyvrsklvdfndlvaigv agfrvdaakhmwpsdlrtiysvrnlrthgfpndaqpyifqevidygneaiskreyngigaviefkys yeisnfrgnnlkwlvnwgeqwgflpskdsldvfdndhdtqrndnpqiltykyskrykmavafmlshp fgtprimssfdqskdqgppndgngnilspsihdnicngwicehrwrqiynmvfrnlvkgtkidnw wdngsnqiafsrcsgfvafngdqydlkknkvclppgqycdvisgnlekgrctgkivtvgsdgnani eigageedgvlaihvkakma</p>

⁸ <http://www.insecte.org/recherche.php3?recherche=mouche+>

⁹ <http://www.rios-galegos.com/abella2.jpg>

Les acides aminés de même classe peuvent se substituer par simple mutation acceptable et répondre ainsi aux contraintes de la sélection évolutive. Il en découle alors des structures protéiques non identiques mais similaires :



Remarque : les acides aminés sont codés par une seule lettre au lieu de trois :

Code international des acides aminés selon l'IUPAC (INTERNATIONAL UNION OF PURE AND APPLIED CHEMISTRY)

Symbole	Code 3c	Acide aminé	Codons
A	Ala	Alanine	GCT, GCC, GCA, GCG
B	Asp, Asn	Aspartic, Asparagine	GAT, GAC, AAT, AAG
C	Cys	Cysteine	TGT, TGC
D	Asp	Aspartic	GAT, GAC
E	Glu	Glutamic	GAA, GAG
F	Phe	Phenylalanine	TTT, TTC
G	Gly	Glycine	GGT, GGC, GGA, GGG
H	His	Histidine	CAT, CAC
I	Ile	Isoleucine	ATT, ATC, ATA
K	Lys	Lysine	AAA, AAG
L	Leu	Leucine	TTG, TTA, GTT, GTC, GTA, GTG
M	Met	Methionine	ATG
N	Asn	Asparagine	AAT, AAC
P	Pro	Proline	CCT, CCG, CCA, CCG
Q	Gln	Glutamine	CAA, CAG
R	Arg	Arginine	CGT, CGC, CGA, CCG, AGA, AGG
S	Ser	Serine	TCT, TCC, TCA, TCG, AGT, AGC
T	Thr	Threonine	ACT, ACC, ACA, ACG
V	Val	Valine	GTT, GTC, GTA, GTG
W	Trp	Tryptophan	TGG
X	Xcc	Unknown	-
Y	Tyr	Tyrosine	TAT, TAC
Z	Glu, Gln	Glutamic, Glutamine	GAA, GAG, CAA, CAG
*	End	Stop	TAA, TAG, TGA

Les matrices protéiques peuvent être classées en deux catégories :

- Une catégorie qui regroupe les matrices issues d'études montrant le caractère de substitution des acides aminés au cours de l'évolution (matrices liées à l'évolution). Elles représentent les échanges 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 physico-chimiques des acides aminés : caractère hydrophile ou hydrophobe des protéines, la structure secondaire ou tertiaire des protéines. **Exemple1** : matrice d'hydrophobicité basée sur des mesures d'énergie libre de transfert de l'eau à l'éthanol des acides aminés (levitt, 1976). **Exemple2** : Matrice de structure secondaire, basée sur la propension (tendance) d'un acide aminé à être dans une conformation donnée (Levinn, 1986).

Ce sont les matrices liées à l'évolution qui seront utilisées pour réaliser les alignements des séquences protéiques.

La matrice PAM250 (Percent Accepted Mutation): La matrice de mutation de Dayhoff.

La plus courante, cette famille de matrices probabilistes a été calculée à partir d'une étude sur une famille de 71 protéines très semblables, que l'on pouvait facilement aligner. Chaque élément de la matrice représente alors la probabilité qu'un acide aminé se transforme en un autre dans un temps d'évolution donné. La matrice créée est une matrice 1PAM, on obtient une matrice XPAM en la multipliant par elle-même. Les probabilités associées sont alors les probabilités de mutation en un temps plus long. En prenant compte des fréquences relatives

de mutation et en prenant le logarithme de chaque élément de la matrice, on construit la matrice PAM-X, utilisable directement dans les programmes. La matrice PAM-250 s'est avérée être optimale par rapport au problème biologique ce qui explique sa très grande fréquence d'utilisation¹⁰.

Les matrices de type PAM dérivent d'alignements globaux de protéines très semblables et représentent les échanges possibles et acceptables d'un acide aminé par un autre au cours de l'évolution des protéines : Les acides aminés entrant dans la composition d'une protéine peuvent avoir les mêmes propriétés physico-chimiques ou presque et la structure 3d va donc dépendre de ces caractéristiques. Cette similarité des propriétés physico-chimiques est donc suffisante pour permettre la substitution (la mutation) entre ces acides aminés sans pour autant perturber la fonction de la protéine.

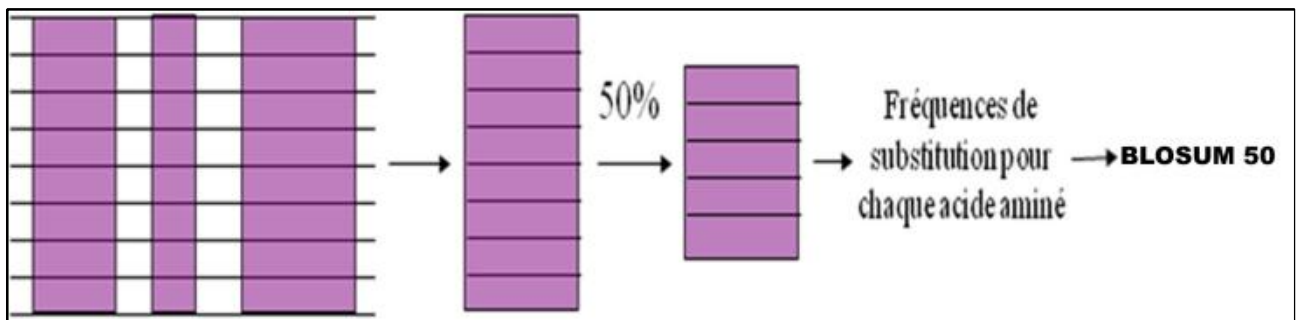
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2																			
R	-2	6																		
N	0	0	2																	
D	0	-1	2	4																
C	-2	-4	-4	-5	4															
Q	0	1	1	2	-5	4														
E	0	-1	1	3	-5	2	4													
G	1	-3	0	1	-3	-1	0	5												
H	-1	2	2	1	-3	3	1	-2	6											
I	-1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

¹⁰ <http://sequencage.outness.net/logi1.html>

La matrice BLOSUM 62 (Henikoff & Henikoff, 1992)

Le degré de substitution des acides aminés a été mesuré en observant des blocs d'acides aminés issus de protéines plus éloignées. Chaque bloc est obtenu par l'alignement multiple sans insertion/délétion de courtes régions très conservées. Ces blocs sont utilisés pour regrouper tous les segments de séquences ayant un pourcentage d'identité minimum au sein de leur bloc. On en déduit des fréquences de substitution pour chaque paire d'acides aminés et l'on calcule ensuite une matrice logarithmique de probabilité dénommée **BLOSUM** (BLOcks SUBstitution Matrix). A chaque pourcentage d'identité correspond une matrice particulière. Ainsi la matrice BLOSUM50 est obtenue en utilisant un seuil d'identité de 50%. Henikoff et Henikoff, (1992) ont réalisé un tel traitement à partir d'une base contenant plus de 2000 blocs¹¹ :

- observation de blocs d'acides aminés issus de protéines relativement éloignées ;
- chaque bloc provient d'alignements multiples sans insertions / délétions de courtes régions conservées ;
- les blocs sont utilisés pour regrouper tous les segments de séquences ayant un pourcentage d'identité minimum au sein de leur bloc ;
- on en déduit des fréquences de substitution pour chaque paire d'acides aminés ;
- on calcule une matrice logarithmique de probabilité ;
- à chaque pourcentage d'identité correspond une matrice :
- BLOSUM50 avec un seuil d'identité de 50 % ;
- BLOSUM62 avec un seuil d'identité de 62 %.



¹¹ http://www.med.univ-angers.fr/discipline/bio_cel/Maitrise/Bioinfo/matrice_blosum.htm

L'algorithme de Needleman et Wunsch pour le cas des protéines :

L'équation suivante résume le principe de calcul d'une case de la matrice transformée :

$$S(i,j) = se(i,j) + \max(S(x,y))$$

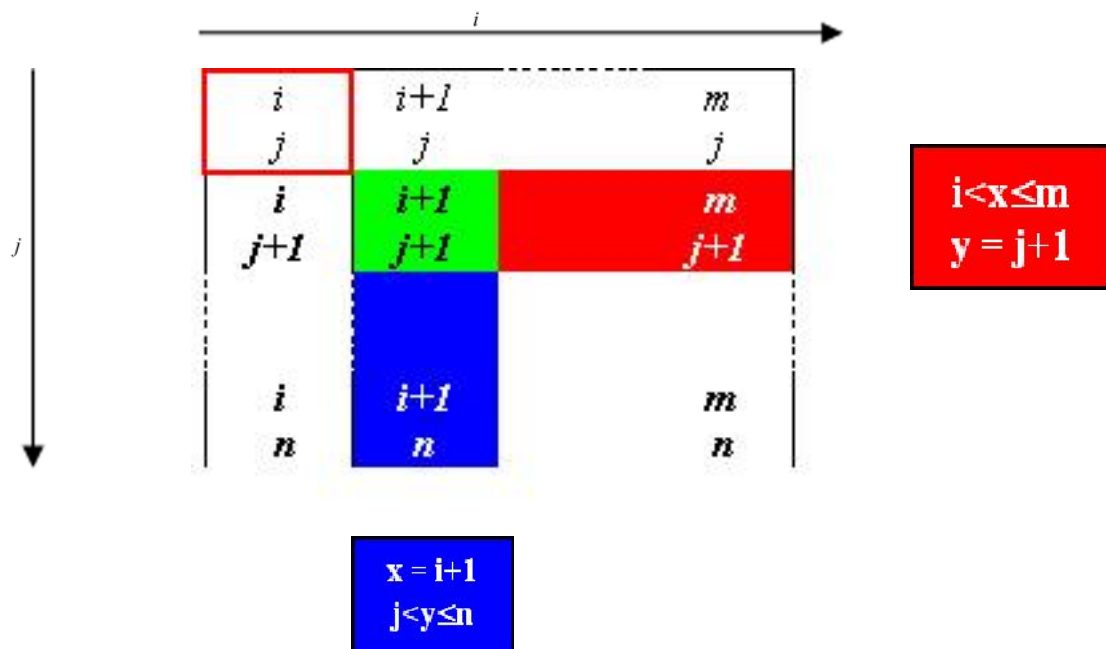
avec :

$$i < x \leq m \text{ et } y = j+1$$

ou

$$x = i+1 \text{ et } j < y \leq n$$

- **S(i,j)** est le score somme de la case d'indice i et j,
- **se** le score élémentaire de la case d'indice i et j de la matrice initiale
- m et n sont les longueurs des deux séquences



Exemple d'alignement : On considère les deux séquences suivantes

Séq 1 = VTEERDAF et Séq 2 = LTSHEAL

Etape 1 : Construction de la matrice initiale à partir de PAM250

	V	T	E	E	R	D	A	F
L	2	-2	-3	-3	-3	-4	-2	2
T	0	3	0	0	-1	0	1	-2
S	-1	1	0	0	0	0	1	-3
H	-2	-1	1	1	2	1	-1	-2
E	-2	0	4	4	-1	3	0	-5
A	0	1	0	0	-2	0	2	-4
L	2	-2	-3	-3	-3	-4	-2	2

Etape 2 : Calcul de la matrice transformée

A ce stade, il faut garder les valeurs de la dernière colonne et de la dernière ligne :

	V	T	E	E	R	D	A	F
L								2
T								-2
S								-3
H								-2
E								-5
A								-4
L	2	-2	-3	-3	-3	-4	-2	2

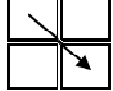
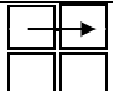

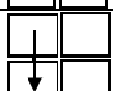

L'application de l'algorithme permet d'obtenir la matrice transformée suivante :

	V	T	E	E	R	D	A	F
L	14	7	6	6	4	0	0	2
T	10	12	9	9	6	4	3	-2
S	8	10	9	9	7	4	3	-3
H	6	7	9	8	9	5	1	-2
E	2	4	8	8	3	7	2	-5
A	2	3	2	2	0	2	4	-4
L	2	-2	-3	-3	-3	-4	-2	2

Etape 3 : Parcours de la matrice transformée

Le parcours s'effectue du plus haut score vers le plus petit. Si les trois cases ont des valeurs de scores égales, alors le chemin vers la diagonale est favorisé :

	V	T	E	E	R	D	A	F
L	14	7	6	6	4	0	0	2
T	10	12	9	9	6	4	3	-2
S	8	10	9	9	7	4	3	-3
H	6	7	9	8	9	5	1	-2
E	2	4	8	8	3	7	2	-5
A	2	3	2	2	0	2	4	-4
L	2	-2	-3	-3	-3	-4	-2	2

		Substitution
		insertion dans i délétion dans j
		insertion dans j délétion dans i

Etape 4 : Aligement des deux séquences

Séq1	V	T	—	E	E	R	D	A	F
Séq2	L	T	*	H	E	*	*	A	L

Le score de cet alignement est : $s= 49$

Il y a trois identités : T-T, E-E et A-A et trois similarités (substitutions): V-L, E-H et F-L

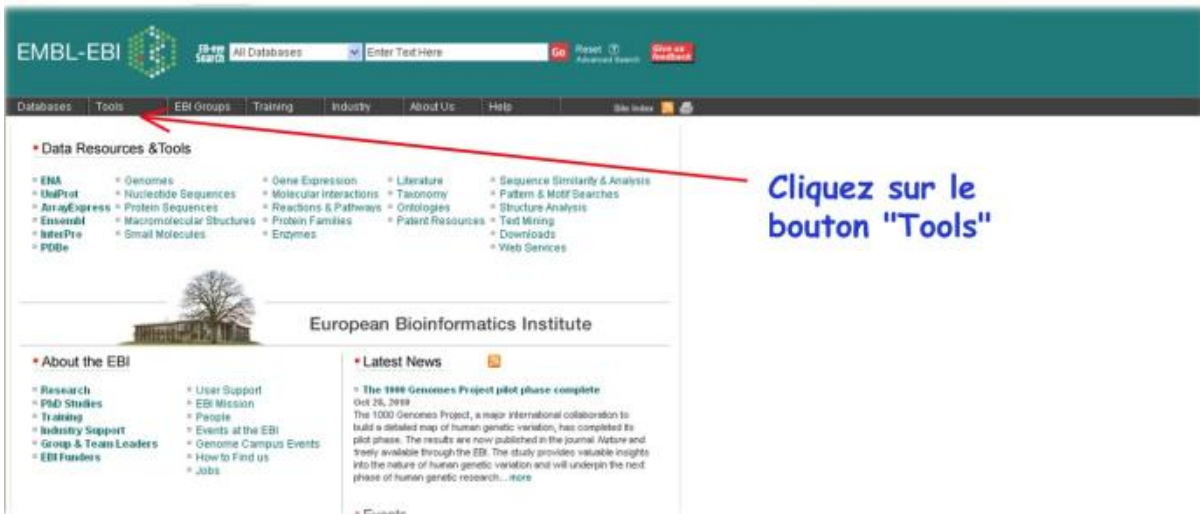
On peut supposer que la valine a été substituée en leucine dans la 2^{ème} séquence par besoin d'adaptation de l'organisme à partir du quel a été isolée cette séquence. Le même raisonnement concernera les substitutions E-H et F-L.

Application 1 : Il s'agit d'aligner, via Internet, deux séquences nucléiques.

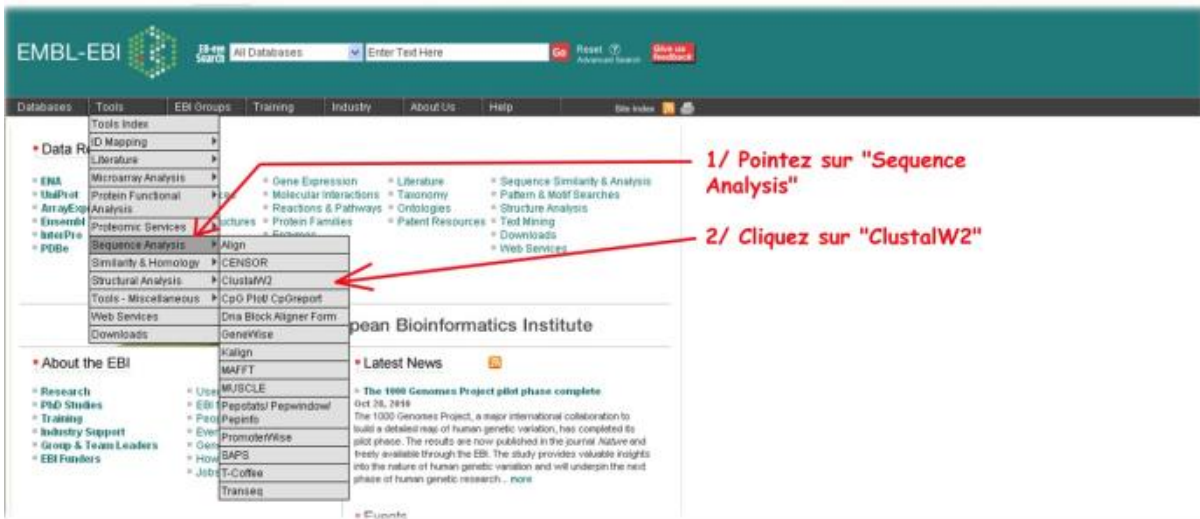
Les deux séquences sont : Séq1 de *Staphylococcus* et Séquence2 de *Micrococcus* pour l'ADN 16S.

Sé1	AGGTGATCCTGGCTCAGGATGAACGCTGGCGGCGTGCCTAATACATGCAAGTCGAGCGAACAGACGAGG AGCTTGCTCCTCTGACGTTAGCGGCGGACGGGTGAGTAACACGTGGATAACCTACCTATAAGACTGGGAT AACTTCGGGAAACCGGAGCTAATACCGGATAATAATATTGAACCCGCATGGTTCAATAGTGAAGACGG TTTTGCTGTCACCTATAGATGGATCCGCGCCGCATTAGCTAGTTGGTAAGGTAACGGCTTACCAAGGC AACGATGCGTAGCCGACCTGAGAGGGTGTATCGGCCACACTGGAAGTGAAGACACGGTCCAGACTCCTA CGGGAGGCAGCAGTAGGGAATCTTCCGCAATGGGCGAAAGCCTGACGGAGCAACGCCGCGTGAGT GATGAAGGTCTTCGGATCGTAAACTCTGTTATTAGGGAAGAACAATGTGTAAGTAACTATGCACG TCTTGACGGTACCTAATCAGAAAGCCACGGC
Séq2	GCCGCGTGAGGGATGACGGCCTTCGGGTTGTAAACCTCTTTCAGTAGGGAAGAAGCGAAAGTGACGGTAC CTGCAGAAGAAGCACCGGCTAACTACGTGCCAGCAGCCGCGGTAATACGTAGGGTGCAGCGTTATCCGG AATTATTGGGCGTAAAGAGCTCGTAGGCGGTTTGTTCGCGTCTGTCGTGAAAGTCCGGGGCTTAACCCCGGA TCTGCGGTGGGTACGGGCAGACTAGAGTGCAGTAGGGGAGACTGGAATTCCTGGTGTAGCGGTGGAATGC GCAGATATCAGGAGGAACACCGATGGCGAAGGCAGGTCTCTGGGCTGTAAGTACGCTGAGGAGCGAAAG CATGGGAGCGAACAGGATTAGATACCCTGGTAGTCCATGCCGTAACGTTGGGCACTAGGTGTGGGGACC ATTCCACGGTTTCCGCGCCGCAGCTAACGCATTAAGTGCCCCGCTGGGGAGTACGGCCGCAAGGCTAAAAC TCAA

Etape 1 : Allez à la page <http://www.ebi.ac.uk/> et cliquez sur le bouton « Tools » du menu horizontal (en haut à gauche) :



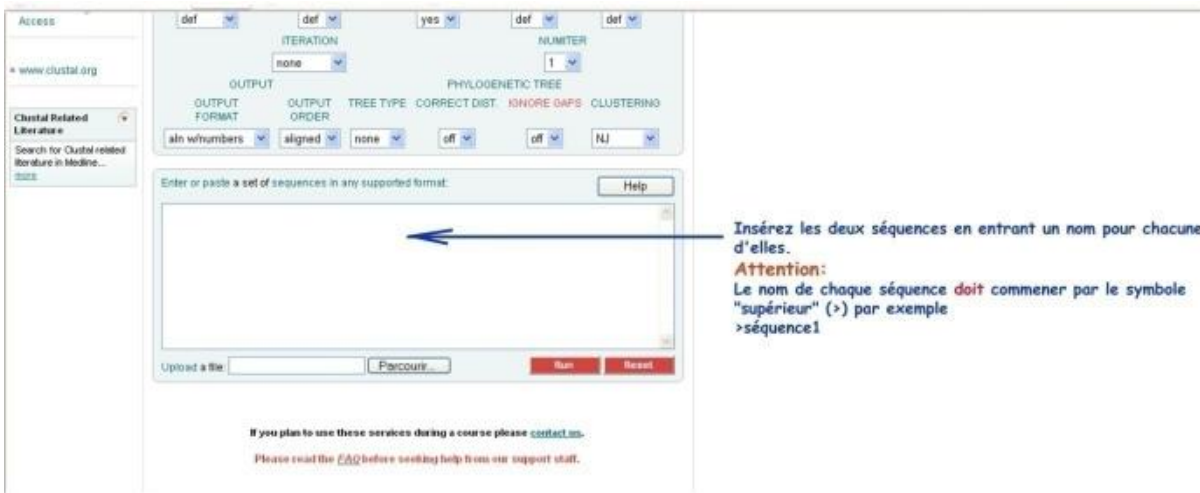
Etape 2 : Un menu déroulant va s'ouvrir. Il faut passer le pointeur de la souris sur "Sequence Analysis" pour découvrir un second menu. Sur ce second menu, cliquez sur "ClustalW2" :

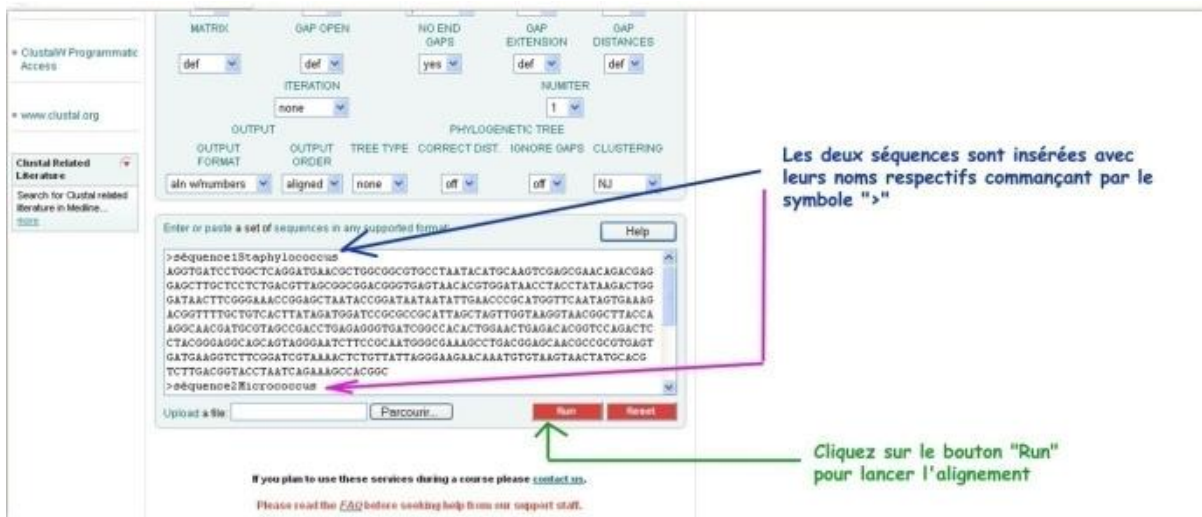


Etape 3 : En bas de la fenêtre qui s'ouvre, insérez les deux séquences en leur donnant chacune un nom significatif de la séquence. **CE NOM DOIT COMMENCER PAR LE SYMBOLE ">" (Supérieur) :**

Par exemple vous pouvez écrire comme nom : **>séquence1Staphylococcus**

Cliquez sur le bouton "Run" pour lancer l'alignement.





Etape 4 : Lecture du résultat. En début de la page des résultats, il y a lieu de constater la valeur du score global. Dans ce cas $S = 1090$.

Vous pouvez cliquer sur le bouton "Start Jalview" pour voir l'alignement en couleur et sur une même ligne.

Application 2 : Il s'agit d'aligner, via Internet, deux séquences protéiques.

Ces deux séquences sont celles de l'amylose de la mouche et de l'abeille.

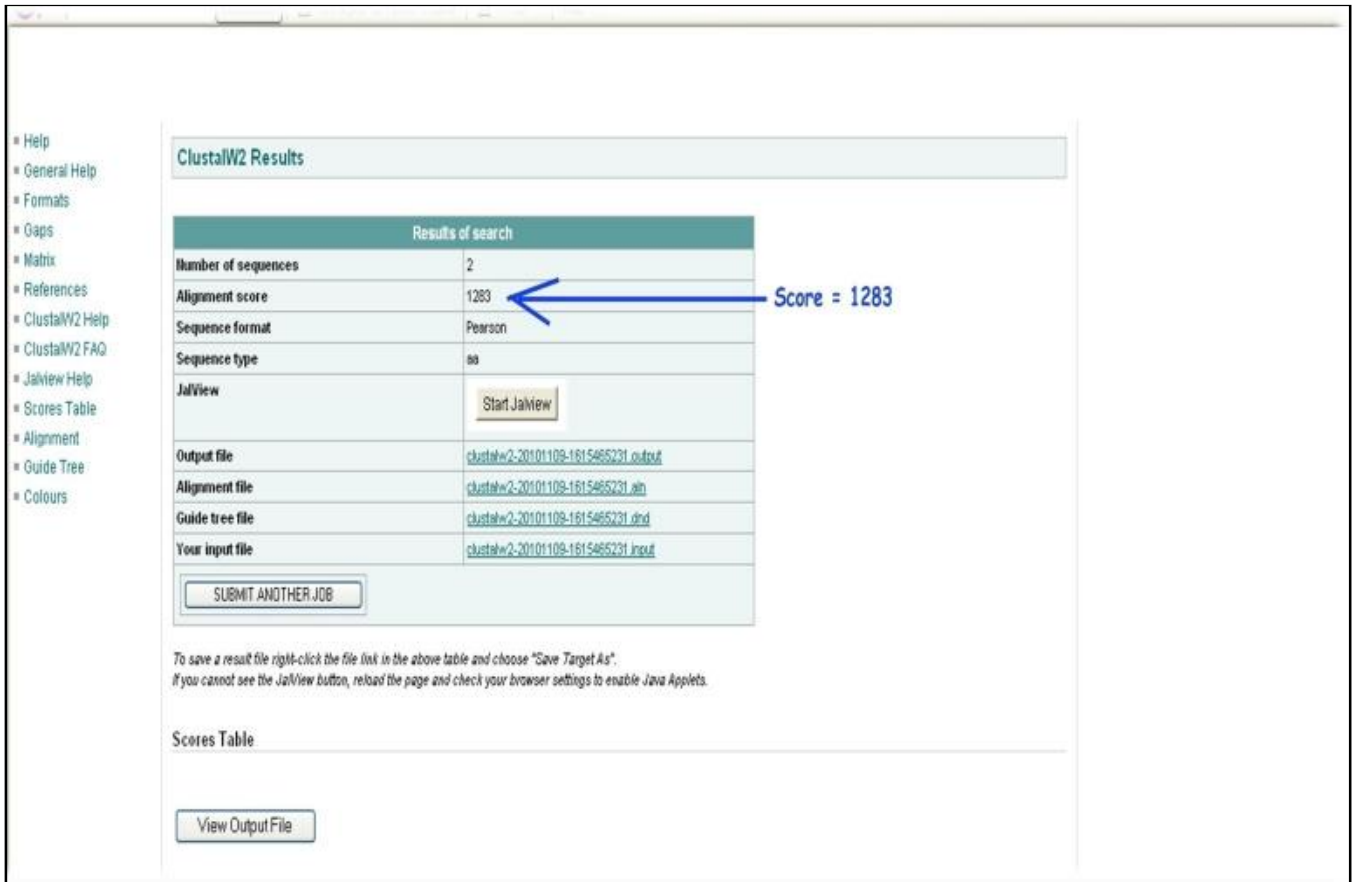
Nom systématique	Séquence : structure primaire
Mouche	IARECEEFLAPRGFAGVQVSPVTENVIVANRPWWERYQPISYKLQTRS GTQQEFSEMCRRCNNVGIRIYVDVLLNHMAADQYQMAVGTAGSIADP AAKSFPSVPYTELDFHATCEIWDWNDRYQVQNCENVLGLKDLQDQSNW VRDCLVEFLDHLVELGVAGFRVDAAKHMKASDLEIYKRVRLNVDHGF EPNSRPFYQEVIDHGHETVSKYEYNLLGAVTEFQFSEEIGRAFRRGNNQ LKWLRNWGPQWGFLPSDHALVFDNHDNQRDGGQVLTYSKSKQYK MATAFALAYPYGTR IMSSFDFTRDQPPHTNE
Abeille	MMPAIVLLLALLTLAAGEIAHNDPHFAPGHDAIVHLFEWKWNDIAKECE QFLGPVGGVQVSPVQENVIDKRPWWERYQPISYKWI TRSGTREQF IDMVARCNKAGVRIYVDVIMNHMSGDRNDAHGTGNSRANTYNFDYPQ VPYTVKNFHPRAVNNYNDPSNVRNCELVGLHDLQDQSEYVRSKLV FLNDLVAIGVAGFRVDAAKHMWPSDLRTIYSRVRNLNRTHGFPNDAQP YIFQEVIDYGNIAISKREYNGIGAVIEFKYSYEISNAFRGNNNLKWLNVW GEQWGFLPSKDSL VFDNHD TQRDNPQILTYKYSKRYKMAVAFMLSH PFGTRIMSSFDFQSKDQPPNDGNGNILSPSIHDNICSNGWICEHRW RQIYNMVRFRNLVKGTKIDNWWDNNGSNQIAFSRGC SGFVAFNGDQYD LKKNLKVCLPPGQYCDVISGNLEKGRCTGKI VTVGSDGNANIEIGAGEE DGVLAHV KAKMA

Etape 1 : Allez directement à la page <http://www.ebi.ac.uk/Tools/clustalw2/index.html> et insérez les deux séquences protéiques.

The screenshot shows the ClustalW2 web interface. On the left, there is a sidebar with 'ClustalW Programmatic Access' and 'Clustal Related Literature'. The main area contains various configuration options for the alignment process, including 'WORD SIZE', 'LENGTH', 'MATRIX', 'GAP OPEN', 'GAP EXTENSION', 'ITERATION', 'NUMBER', 'OUTPUT', and 'PHYLOGENETIC TREE'. A text area at the bottom contains the two protein sequences from the table above, with the first sequence labeled '>MOUCHE' and the second '>ABEILLE'. A green note on the right says 'N'oubliez pas la règle d'écriture des noms de séquences.' (Don't forget the sequence naming rule). At the bottom, there are 'Upload a file', 'Process', 'Run', and 'Reset' buttons.

Une fois les deux séquences collées dans la zone et portant chacune un nom convenablement écrit, cliquez sur le bouton **“Run”** (En bas en rouge) :

Etape 2 : Dans la première partie de la page web, vous pouvez lire le résultat du score global après alignement des deux séquences : $S=1283$



The screenshot displays the ClustalW2 Results page. On the left, there is a navigation menu with links such as Help, General Help, Formats, Gaps, Matrix, References, ClustalW2 Help, ClustalW2 FAQ, Jalview Help, Scores Table, Alignment, Guide Tree, and Colours. The main content area is titled "ClustalW2 Results" and contains a table with the following data:

Results of search	
Number of sequences	2
Alignment score	1283
Sequence format	Pearson
Sequence type	aa
JalView	<input type="button" value="Start JalView"/>
Output file	clustalw2-20101109-1615465231.output
Alignment file	clustalw2-20101109-1615465231.aln
Guide tree file	clustalw2-20101109-1615465231.dnd
Your input file	clustalw2-20101109-1615465231.input

Below the table, there is a button labeled "SUBMIT ANOTHER JOB". A blue arrow points from the text "Score = 1283" to the value "1283" in the "Alignment score" row. Below the table, there is a note: "To save a result file right-click the file link in the above table and choose 'Save Target As'. If you cannot see the JalView button, reload the page and check your browser settings to enable Java Applets." Below this note, there is a section titled "Scores Table" and a button labeled "View Output File".

La suite de la page donne l'alignement des deux protéines :

```
CLUSTAL 2.0.12 multiple sequence alignment

MOUCHE      -----IARECEEFLAPRGFAGV 17
ABEILLE     MMPAIVLLLLALLTLAAGEIAHNDPHFAPGHDAIVHLFEWKWNDIAKECEQFLGPVGFGGV 60
                                         **:***:*. * **.*

MOUCHE      QVSPVTENVIVANRPWERYQPISYKLQTRSGTQQEFSEMCRRCNNVGIIRIYVDVLLNHM 77
ABEILLE     QVSPVQENIVIDKRPWERYQPISYKWITRSGTREQFIDMVARCNKAGVRIYVDVIMNHM 120
***** **: : :***** *****: : * : * **.: :*****: :***

MOUCHE      AADQYQMAVGTAGSIADPAAKSFPSVPTYELDFHATCEIWDWDRYQVQNCVLVGLKLDL 137
ABEILLE     SGDR-NDAHGTGNSRANTYNFDYPQVPYTVKNFHPRCVANNYNDPSNVRNCELVGLHLDL 179
.: : : * **.* *:. .:*.***** :**.* : : : ** :*:*****:***

MOUCHE      QSNEWVRDCLVEFLDHLVELGVAGFRVDAAKHMKASDLEIIYKRVRDLNVDHGFEPNSRP 197
ABEILLE     QSQEYVRSKLVDFLNDLVAIGVAGFRVDAAKHMWPSDLRTIYSRVRNLNRTHGFPNDAQP 239
**:*:*. **:*:*.** :***** .***. **.***:*. ** ** : : *

MOUCHE      FIYQEVIDHGHETVSKYEYNLLGAVTEFQFSEEIGRAFRGNNQLKULRNUGFPQUGFLPSD 257
ABEILLE     YIFQEVIDYGNEAISKREYNGIGAVIEFKYSYEISNAFRGNNLKVLVNUGEQUGFLPSK 299
: :*****: : : : ** ** : ** ** : * **..*****:**** ** *****.

MOUCHE      HALVFVDNHDNQRDGGQVLTYKNSKOYKMAAFALAYPYGITRIMSSFDFTDRDQPPPH 317
ABEILLE     DSLVFVDNHDNQRDNPQILTYKSKRYKMAVAFMLSHPPGTTRIMSSFDQSKDQGPND 359
.:*****.***. *:**** **:****. ** *: : : * .***** .: ** **:

MOUCHE      NE----- 319
ABEILLE     GNGNILSPSIHDNICSNGWICEHRWRQIYNHVRFRNLVKGTKIDNWUDNGSNQIAFSRGC 419
.:

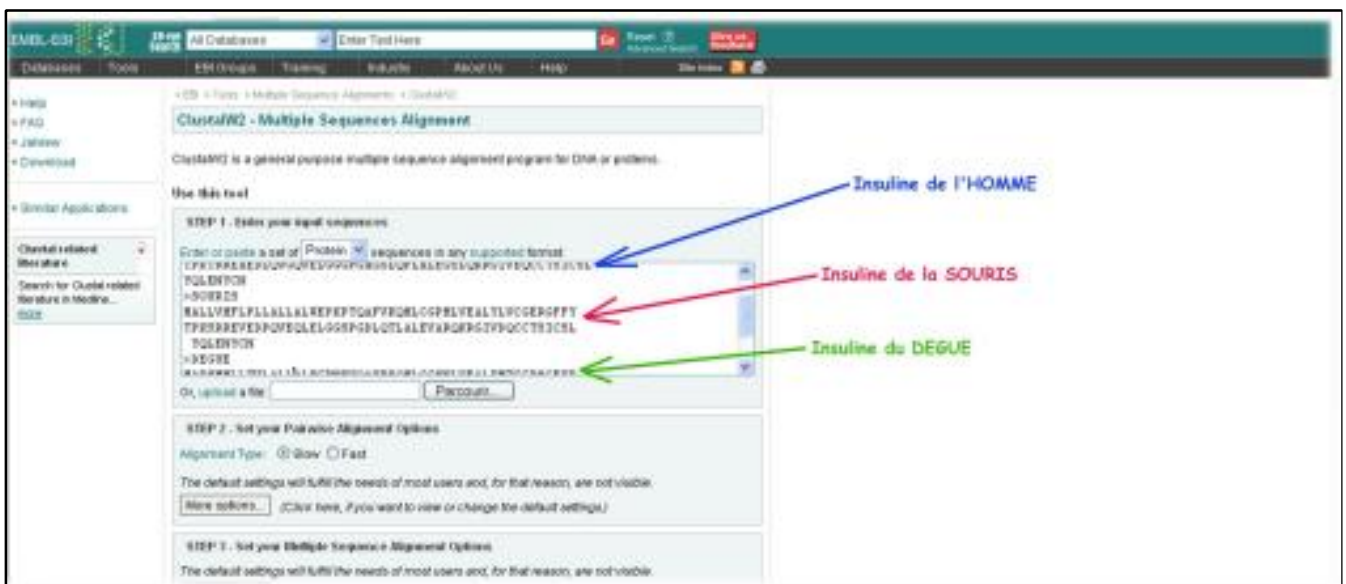
MOUCHE      ----- 479
ABEILLE     SGFVAFNGDQYDLKKNLKVCLPPGQYCDVISGNLEKGRCTGKIVTVGSDGNANIEIGAGE 479

MOUCHE      -----
ABEILLE     EDGVLAIHVKAKMA 493
```


Application 3 : Nous allons nous initier à travers cette application aux alignements multiples (trois séquences et plus). Les séquences qui nous intéressent sont :

<p><i>Homo sapiens</i></p> <p>AAA59172</p> <p>110 acides aminés</p>	<p>MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFY TPKTRREAEDLQVGGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSL YQLENYCN</p>
<p><i>Mus musculus</i></p> <p>AAI45871</p> <p>108 acides aminés</p>	<p>MALLVHFLPLLALLALWEPKPTQAFVKQHLCGPHLVEALYLVCGERGFFY TPKSRREVEDPQVEQLELGGSPGDLQTLALEVARQKRGIVDQCCTSICSL YQLENYCN</p>
<p><i>Octodon degus</i></p> <p>AAA40590</p> <p>109 acides aminés</p>	<p>MAPWMHLLTVLALLALWGPNSVQAYSSQHLCGSNLVEALYMTCGRSGFYR PHDRRELEDLQVEQAELGLEAGGLQPSALEMILQKRGIVDQCCNNICTFN QLQNYCNVP</p>

Etape 1 : Allez à la page <http://www.ebi.ac.uk/Tools/clustalw2/index.html> et insérez les trois séquences protéiques (Insuline de l'homme, de la souris et du dègue du Chili).



Une fois les trois séquences collées dans la zone et portant chacune un nom convenablement écrit, cliquez sur le bouton **“Run”**.

Etape 2 : Le résultat de l’alignement des trois séquences protéiques montre des zones de forte similarité et des zones qui contiennent des gaps.

> EBI > Tools > Multiple Sequence Alignments > ClustalW2

ClustalW2 Results

Alignments | Result Summary | Guide Tree | Submission Details | Submit Another Job

Alignment

View Alignment File | Hide Colors

CLUSTAL 2.0.12 multiple sequence alignment

```
HOMME      MALWMRLPLLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED 60
SOURIS     MALLVHFLPLLALLALWEPKPTQAFVKQHLCGPHLVEALYLVCGERGFFYTPKSRREVED 60
DEGUE      MAPWMHLLTVLALLALWGPNSVQAYSSQHLCGSNLVEALYMTCCRSG-FYRPHDRRELED 59
**  :::*.:***** *... *: .*****.:*****:..*, * ** *: *** **

HOMME      LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN-- 110
SOURIS     PQVEQLELGGSP--GDLQTLALEVARQKRGIVDQCCTSICSLYQLENYCN-- 108
DEGUE      LQVEQAELEGLA--GGLQPSALEMILQKRGIVDQCCNNICTFNQLQNYCNVP 109
** * *** . *,**_ *** *****:***_**:: **:****
```

PLEASE NOTE: Showing colors on large alignments is slow.

Contact EBI | © European Bioinformatics Institute 2010. EBI is an Outstation of the European Molecular Biology Laboratory.

Remarque : Ce type d'alignement est également réalisable sur d'autres sites^{12,13} ou avec des programmes téléchargeables. Il fera l'objet du prochain support pédagogique.

Il faut retenir que les programmes les plus utilisés pour ce type d'alignements sont :

MultAlin http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_multalinan.html

Dialign <http://bibiserv.techfak.uni-bielefeld.de/dialign/>

¹² <http://align.genome.jp/sit-bin/clustalw>

¹³ http://www.ensam.inra.fr/biochimie/td/analyses_sequences/alignements_multiples.html

ClustalW

http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_clustalw.html



Pôle BioInformatique Lyonnais
Network Protein Sequence Analysis
NPS@ is the IBCP contribution to PBIL in Lyon, France

[HOME](#) | [NPS@](#) | [SRS](#) | [HELP](#) | [REFERENCES](#) | [NEWS](#) | [MPSA](#) | [ANTHEPROT](#) | [Gemo3D](#) | [SubMe](#) | [Positions](#) | [PBIL](#)

Wednesday, August 27th 2008. Computer upgrade for increase speed of computation ([see news](#))
Tuesday, April 27th 2010. A filesystem corruption on 24th and 27th causes NPS@ service downtime.

CLUSTALW

[\[Abstract\]](#) | [\[NPS@ help\]](#) | [\[Original server\]](#)

Paste a protein sequence databank in Pearson/Fasta format below : [help](#)

All sequence names must be different !

Output width :