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 !

Homo sapiens[2] **AAA59172** 110 acides aminés		MALWMRLLPLLALLALWGPDPAAAF VNQHLCGSHLVEALYLVCGERGFFY TPKTRREAEDLQVGQVELGGGPGAG SLQPLALEGSLQKRGIVEQCCTSICSL YQLENYCN
Mus musculus[3] **AAI45871** *108 acides aminés*		MALLVHFLPLLALLALWEPKPTQAF VKQHLCGPHLVEALYLVCGERGFFY TPKSRREVEDPQVEQLELGGSPGDL QTLALEVARQKRGIVDQCCTSICSL YQLENYCN
Octodon degus[4] **AAA40590** 109 acides aminés		MAPWMHLLTVLALLALWGPNSVQAY SSQHLCGSNLVEALYMTCGRSGFYR PHDRRELEDLQVEQAELGLEAGGLQ PSALEMILQKRGIVDQCCNNICTFN QLQNYCNVP

[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/degu-octodondegus.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 dissémilarité),
- 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**	**1+0+0+1+1+0+1+1+0+1=6**

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 10ème 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] x100). 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 :

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

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

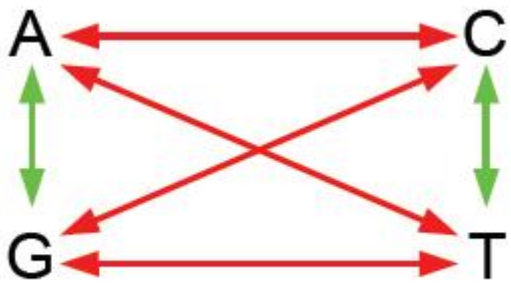
Il faut savoir qu'il existe une matrice (**matrice d'identité)** qui donne les valeurs de scores d'identité entre les séquences à comparer. Dans cette matrice, on attribue la valeur de 1 lorsque les deux nucléotides sont identiques et zéro s'ils ne le sont pas.

	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 :



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

Matrice de transition/transversion

Remarque *: Quelle matrice utiliser ?*

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

Recherche de segments identiques : La matrice de points

Elle permet une vue (méthode visuelle) englobant les similarités entre les régions des séquences à comparer.

Exemple de réalisation: On donne deux séquences x et y :

x=ACTCGGATT et y=AGCTCGGT

Cette méthode consiste à créer une matrice qui va contenir les deux séquences (la séquence x en horizontal et la séquence y en vertical) et de cocher les cases de cette matrice pour le seul cas où les nucléotides sont identiques (Match). Quand il n'ya pas identité on parle de Mismatch:

		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, constatons qu'il y a une diagonale formée de cinq cases. Donc le segment identique le plus long entre les deux séquences x et y contient cinq nucléotides identiques et consécutifs 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									

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

		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 Dottup[5].

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 x 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 élémentaires[6].

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

La fenêtre formée des cinq premiers nucléotides de la séquence A est : ATGTA. Il faut la comparer avec toutes les fenêtres possibles de taille égale à cinq retrouvées sur la séquence B. Ces séquences sont :

1. TATGT
2. ATGTG
3. TGTGA
4. GTGAA
5. TGAAT
6. GAATG

Remarque : Au-delà du nucléotide G en 6ème position dans la séquence B, on ne peut plus avoir une fenêtre de taille égale à cinq nucléotides.

La première comparaison concerne les deux motifs suivants :

Fenêtre de la séquence A = ATGTA

Fenêtre de la séquence B = TATGT

La comparaison de ces deux segments donne un score égal à zéro car il n'ya aucun nucléotide de la séquence A qui soit identique à celui de la séquence B quelque soit le site de comparaison :

Séquence A	A	T	G	T	A	Score global
Séquence B	T	A	T	G	T	
Scores élémentaires	0	0	0	0	0	S = 0

Remarque : Sur la matrice qui contient la totalité des deux séquences A et B, allez au à la case d'intersection qui rejoint le milieu du premier segment de A et celui de B pour insérer la valeur du score global :

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

En fixant le motif de la séquence A (ATGTA), vous passez au deuxième motif de la séquence B qui est ATGTG :

Séquence A	A	T	G	T	A	Score global
Séquence B	A	T	G	T	G	
Scores élémentaires	1	1	1	1	0	S = 4

La comparaison de la fenêtre de A avec les cinq fenêtres possibles de B donne les résultats suivants :

	ATGTA / TATGT	ATGTA / ATGTG	ATGTA / TGTGA	ATGTA / GTGAA	ATGTA / TGAAT	ATGTA / GAATG
Nucléotides identiques (score)	0	4	1	3	0	1

Ce qui donnera sur la matrice globale :

	A	**T**	**G**	**T**	**A**	**A**	T	G	C	A	T	G
T												
A												
T			**0**									
G			**4**									
T			**1**									
G			**3**									
A			**0**									
A			**1**									
T												
G												

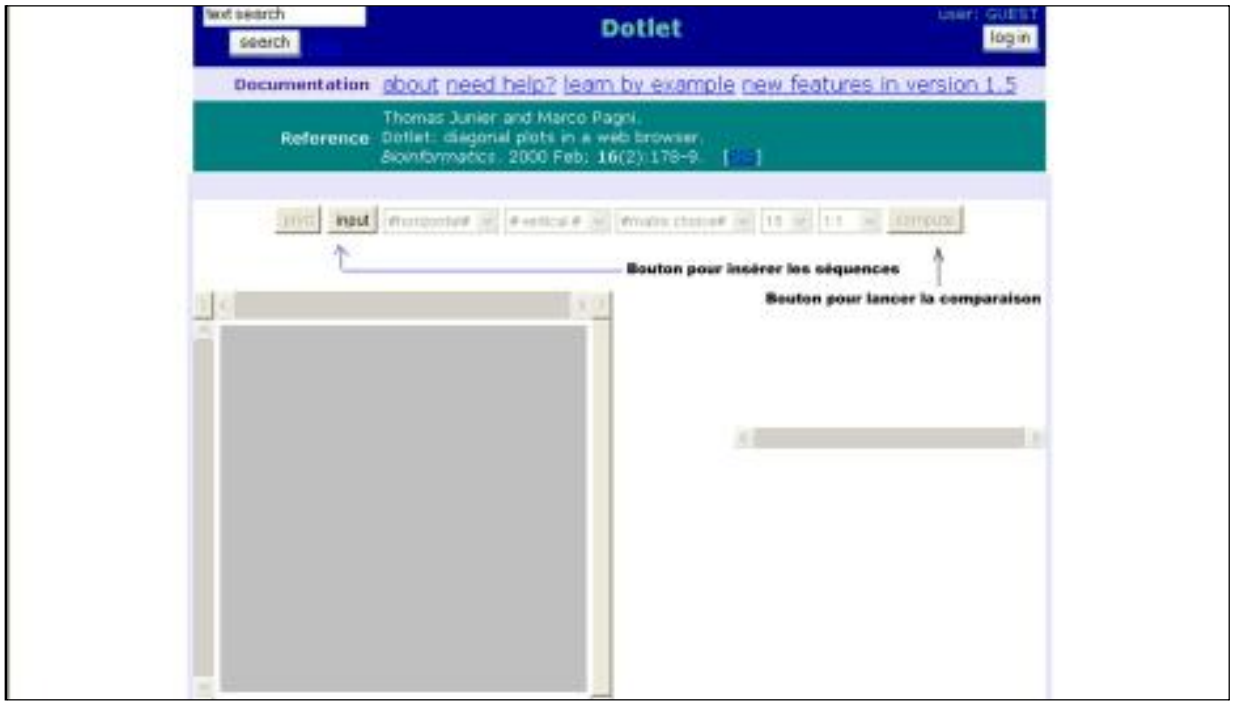
Une fois la comparaison effectuée avec toutes les fenêtres de B, nous incrémentons de un la séquence de A pour avoir la nouvelle fenêtre de cinq autres nucléotides qui sont : **TGTAA.**

C'est cette nouvelle fenêtre de A que nous allons devoir comparer avec les fenêtres de B que nous connaissons toutes maintenant.

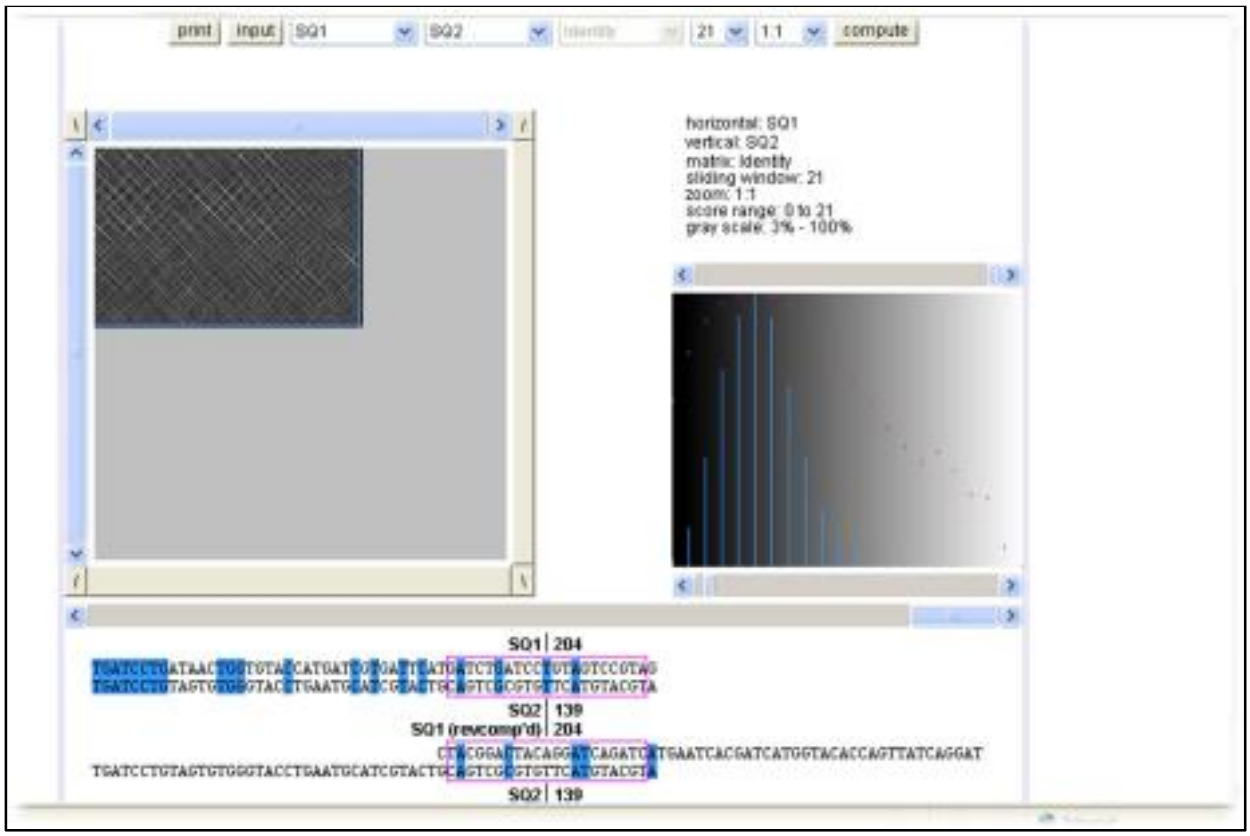
Le résultat final est :

	A	T	G	T	A	A	T	G	C	A	T	G	Interprétation
T													Il y a cinq segments formés de cinq nucléotides chacun entre les séquences A et B. Ces segments contiennent tous quatre nucléotides identiques: 1. ATGTA de la séqA et ATGGA de la séqB 2. TAATG de la séqA et GAATG de la séqB 3. TGTAA de la séqA et TGGAA de la séqB 4. TGCAT de la séqA et GGAAT de la séqB 5. GCATG de la séqA et GAATG de la séqB
A													
T			0	0	1	0	3	0	0	0			
G			**4**	0	1	0	1	3	0	0			
G			3	**4**	3	0	0	3	1	1			
A			0	3	3	0	0	0	**4**	1			
A			1	0	0	**4**	1	0	0	**4**			
T													
G													

Le programme dotlet[7] 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 Mx de taille n (n= 3 nucléotides par exemples) aura pour code Cx de sorte que :

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

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, **A**TG, **T**GC, **G**CC, **C**CT

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

C1 = $P_1x4^{(3-1)}$ + $P_2x4^{(3-2)}$ + $P_3x4^{(3-3)}$ + 1, P_1=3 (car le nucléotide est T), P_2 =0 (car le nucléotide est A) et P_3 =3 (car le nucléotide est T)

D'où

C1 = $P_1x4^{(2)}$ + $P_2x4^{(1)}$ + P_3 + 1

C1 = $3x4^{(2)}$ + $0x4^{(1)}$ + 3 + 1= 52

C1 = 52 correspond au triplet TAT

- Le motif **ATG** :

C2 = $P_1x4^{(3-1)}$ + $P_2x4^{(3-2)}$ + $P_3x4^{(3-3)}$ + 1, P_1=0(car le nucléotide est A), P_2 =3 (car le nucléotide est T) et P_3 =2 (car le nucléotide est G)

D'où

C2 = $P_1x4^{(2)}$ + $P_2x4^{(1)}$ + P_3 + 1

C2 = $0x4^{(2)}$ + $3x4^{(1)}$ + 2 + 1= 15

C2 = 15 correspond au triplet ATG

- Le motif **TGC** :

C3 = $P_1x4^{(3-1)}$ + $P_2x4^{(3-2)}$ + $P_3x4^{(3-3)}$ + 1, P_1=3(car le nucléotide est T), P_2 =2 (car le nucléotide est G) et P_3 =1 (car le nucléotide est C)

D'où

$C3 = P_1x4^{(2)} + P_2x4^{(1)} + P_3 + 1$

$C3 = 3x4^{(2)} + 2x4^{(1)} + 1 + 1 = 62$

C3 = 62 correspond au triplet ATG

- Le motif **GCC** :

$C4 = P_1x4^{(3-1)} + P_2x4^{(3-2)} + P_3x4^{(3-3)} + 1$, P_1=2(car le nucléotide est G), P_2 =1(car le nucléotide est C) et P_3 =1 (car le nucléotide est C)

D'où

$C4 = P_1x4^{(2)} + P_2x4^{(1)} + P_3 + 1$

$C4 = 2x4^{(2)} + 1x4^{(1)} + 1 + 1 = 38$

C4 = 38 correspond au triplet ATG

- Le motif **CCT** :

$C5 = P_1x4^{(3-1)} + P_2x4^{(3-2)} + P_3x4^{(3-3)} + 1$, P_1=1(car le nucléotide est C), P_2 =1(car le nucléotide est C) et P_3 =3 (car le nucléotide est T)

D'où

$C5 = P_1x4^{(2)} + P_2x4^{(1)} + P_3 + 1$

$C5 = 1x4^{(2)} + 1x4^{(1)} + 3 + 1 = 24$

C5 = 24 correspond au triplet ATG

Séquence	T	A	T	G	C	C	T
1er motif	**T**	**A**	**T**				
2ème motif		**A**	**T**	**G**			
3ème motif			**T**	**G**	**C**		
4ème motif				**G**	**C**	**C**	
5ème 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
1er motif	**A**	**G**	**A**				
2ème motif		**G**	**A**	**T**			
3ème motif			**A**	**T**	**G**		
4ème motif				**T**	**G**	**C**	
5ème 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 ya 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 :

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

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 1ère ligne et la 1ère 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)*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 **DEL**étion au cours de la quelle 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 6ème position face au gap de S1). Comme on peut supposer que c'est la séquence S2 qui a subi une mutation par **IN**sertion du nucléotide A par nécessité adaptative. Dans un cas ou dans l'autre une des deux séquences a subi une mutation (**IN**sertion ou **DEL**étion) ; 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 8ème 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	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 2ème 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)*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 nucléique basé sur l’identité n’est pas approprié pour le cas des systèmes protéiques. Ceci est du 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.

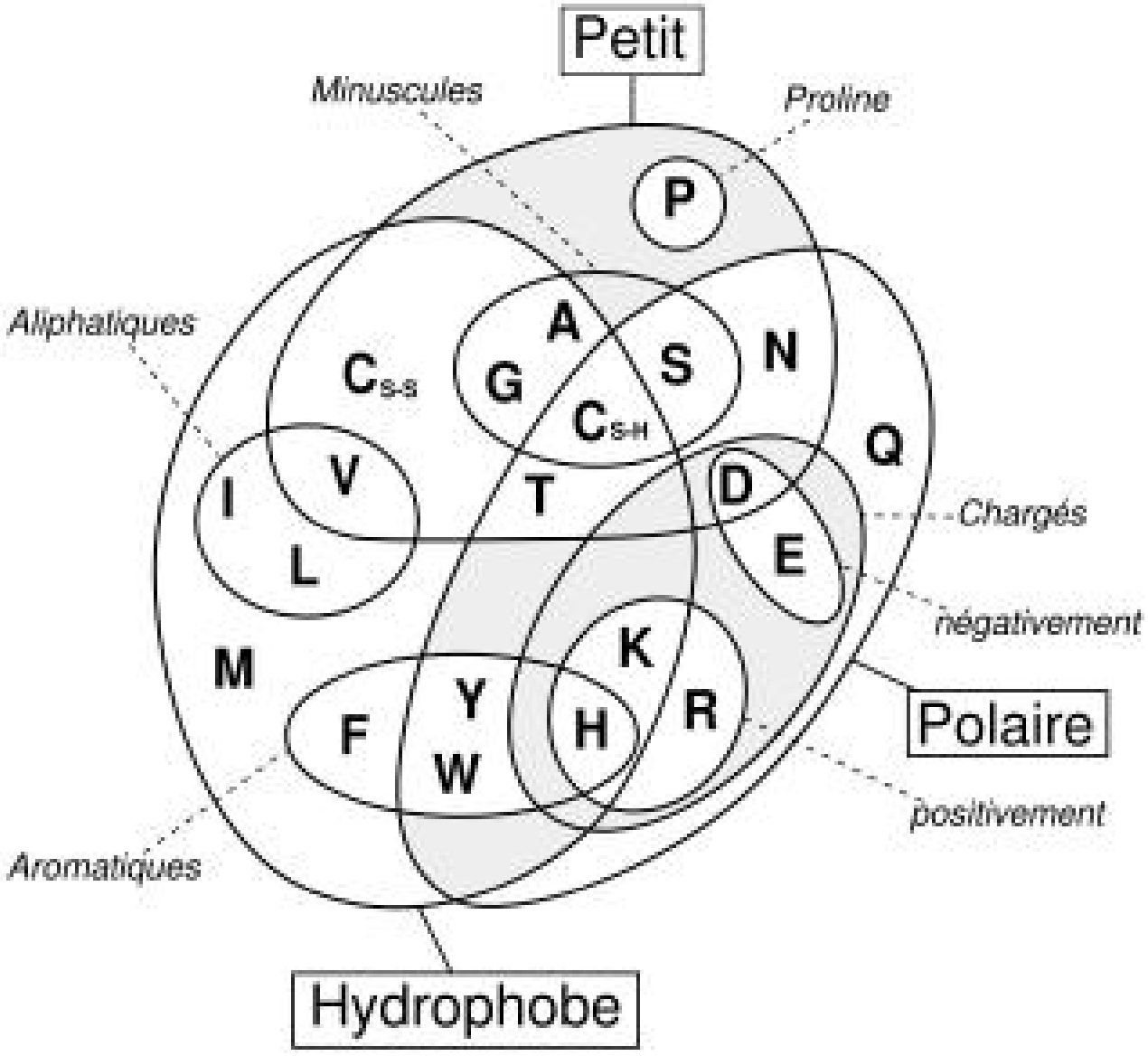
On peut donc classer les acides aminés en familles par rapport à leurs propriétés et obtenir ainsi un système de scores qui rende compte de l'affinité des résidus protéiques entre eux. C’est cette affinité qui permet à un acide aminé d’être substitué par un autre, et les deux structures protéiques ne 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’accession	Séquence : structure primaire
Mouche[8] *Musca domestica* 	AAY88830 319 aa	iareceeflaprgfagvqvspvtenvivanrpwweryqpisyklqtrsgtqqefsemcrrcnnvgiriy vdvllnhmaadqyqmavgtagsiadpaaksfpsvpyteldfhatceiwdwndryqvqncelvglk dldqsnewvrdclvefldhlvelgvagfrvdaakhmkasdleiiykrvrdlnvdhgfepnsrpfiyqev idhghetvskyeynllgavtefqfseeigrafrgnnqlkwlrnwgpqwgflpsdhalvfvdnhdnqr dggqvltyknskqykmatafalaypygitr imssfdftdrdqppphtne
Abeille Apis mellifera[9] 	BAA86909 493 aa	mmpaivlllalltlaageiahndphfapghdaivhlfewkwndiakeceqflgpvgfggvqvspvqe nividkrpwweryqpisykwitrsgtreqfidmvarcnkagvriyvdvimnhmsgdrndahgtgns rantynfdypqvpytvknfhprcavnnyndpsnvrncelvglhdldqsqeyvrsklvdflndlvaigv agfrvdaakhmwpsdlrtiysrvrnlnrthgfpndaqpyifqevidygneaiskreyngigaviefkys yeisnafrgnnnlkwlvnwgeqwgflpskdslvfvdnhdtqrdnpqiltykyskrykmavafmlshp fgtprimssfdfqskdqgppndgngnilspsihdnicsngwicehrwrqiynmvrfrnlvkgtkidnw wdngsnqiafsrgcsgfvafngdqydlkknlkvclppgqycdvisgnlekgrctgkivtvgsdgnani eigageedgvlaihvkakma

[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	Asparti, Asparagine	GAT, GAC, AAT, AAC
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, CTT, CTC, CTA, CTG
M	Met	Methionine	ATG
N	Asn	Asparagine	AAT, AAC
P	Pro	Proline	CCT, CCC, CCA, CCG
Q	Gln	Glutamine	CAA, CAG
R	Arg	Arginine	CGT, CGC, CGA, CGG, 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	Xxx	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 aminées 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 (Purcent 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'utilisation[10].

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.

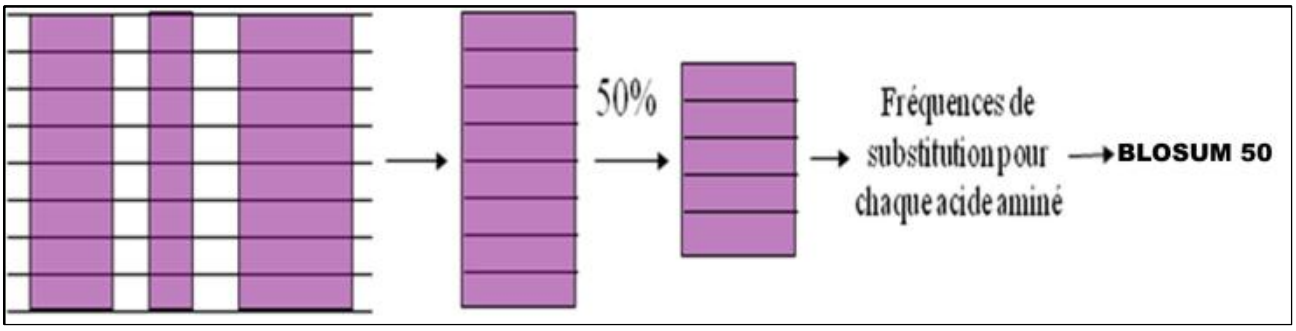
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2																			
R	-2	6																		
N	0	0	2																	
D	0	-1	2	4																
C	-2	-4	-4	-5	4															
Q	0	1	1	2	-5	4														
E	0	-1	1	3	-5	2	4													
G	1	-3	0	1	-3	-1	0	5												
H	-1	2	2	1	-3	3	1	-2	6											
I	-1	-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 blocs[11] :

- 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))

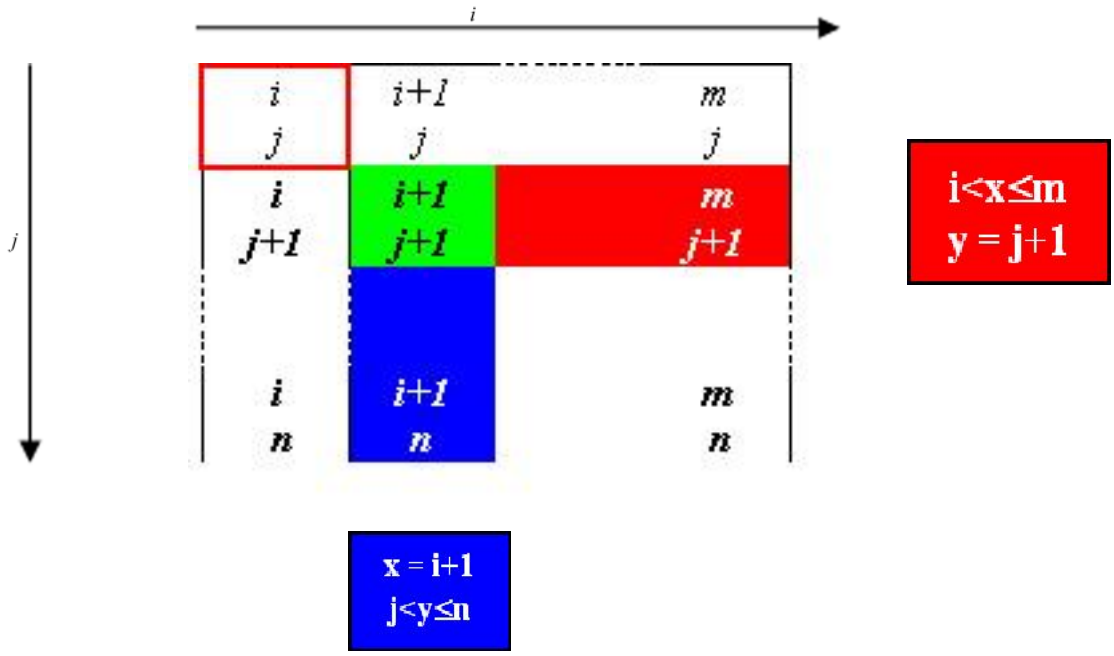
avec :

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

ou

$x = i+1$ 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

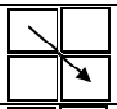
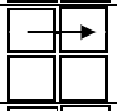
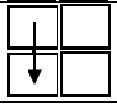
L'application de l'algorithme permet d'obtenir la matrice transformée suivante :

	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 : Alignement des deux séquences

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

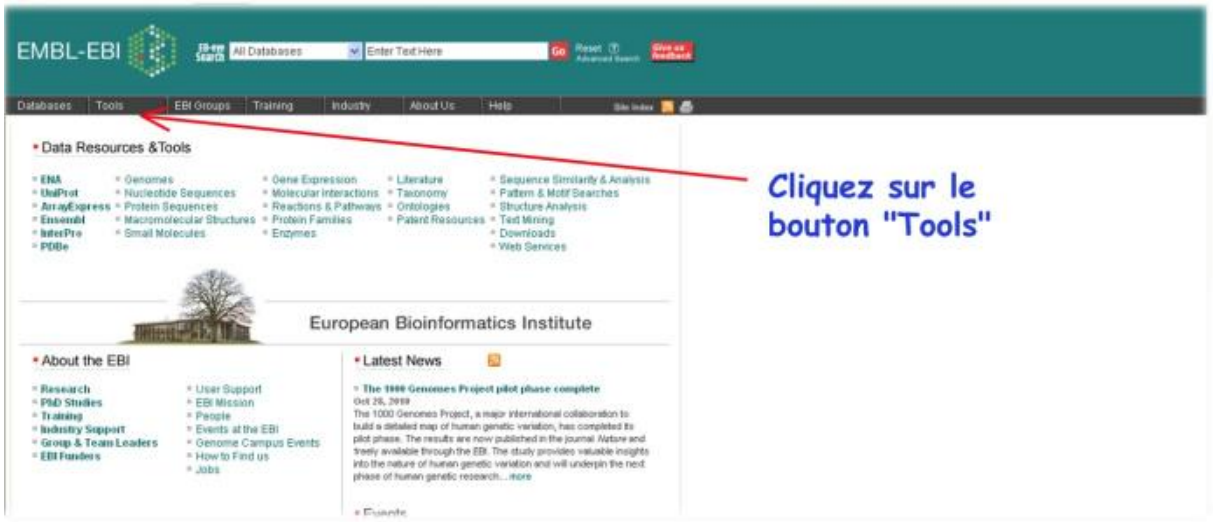
On peut supposer que la valine a été substituée en leucine dans la 2ème séquence par besoin d'adaptation de l'organisme à partir du quel a été isolée cette séquence. Le même raisonnement concernera les substitutions E-H et F-L.

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

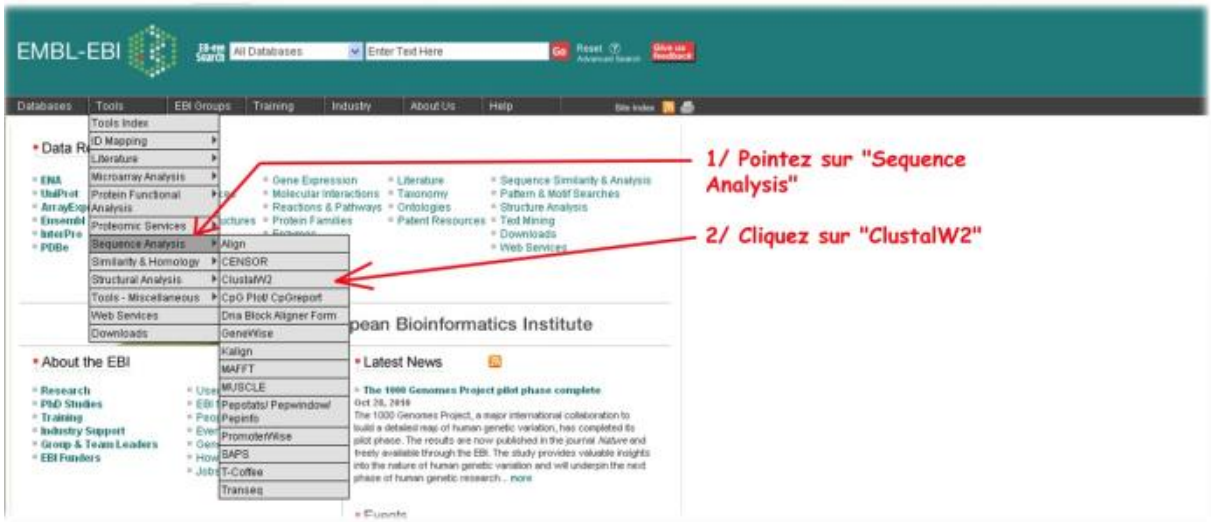
Les deux séquences sont : Séq1 de *Staphylococcus* et Séquence2 de *Micrococcus* pour l'ADN 16S.

Sé1	AGGTGATCCTGGCTCAGGATGAACGCTGGCGGCGTGCCTAATACATGCAAGTCGAGCGAACAGACGAGG AGCTTGCTCCTCTGACGTTAGCGGCGGACGGGTGAGTAACACGTGGATAACCTACCTATAAGACTGGGAT AACTTCGGGAAACCGGAGCTAATACCGGATAATAATATTGAACCCGCATGGTTCAATAGTGAAAGACGG TTTTGCTGTCACTTATAGATGGATCCGCGCCGCATTAGCTAGTTGGTAAGGTAACGGCTTACCAAGGC AACGATGCGTAGCCGACCTGAGAGGGTGATCGGCCACACTGGAACTGAGACACGGTCCAGACTCCTA CGGGAGGCAGCAGTAGGGAATCTTCCGCAATGGGCGAAAGCCTGACGGAGCAACGCCGCGTGAGT GATGAAGGTCTTCGGATCGTAAAACTCTGTTATTAGGGAAGAACAAATGTGTAAGTAACTATGCACG TCTTGACGGTACCTAATCAGAAAGCCACGGC
Séq2	GCCGCGTGAGGGATGACGGCCTTCGGGTTGTAAACCTCTTTCAGTAGGGAAGAAGCGAAAGTGACGGTAC CTGCAGAAGAAGCACCGGCTAACTACGTGCCAGCAGCCGCGGTAATACGTAGGGTGCGAGCGTTATCCGG AATTATTGGGCGTAAAGAGCTCGTAGGCGGTTTGTCGCGTCTGTCGTGAAAGTCCGGGGCTTAACCCCGGA TCTGCGGTGGGTACGGGCAGACTAGAGTGCAGTAGGGGAGACTGGAATTCCTGGTGTAGCGGTGGAATGC GCAGATATCAGGAGGAACACCGATGGCGAAGGCAGGTCTCTGGGCTGTAACTGACGCTGAGGAGCGAAAG CATGGGGAGCGAACAGGATTAGATACCCTGGTAGTCCATGCCGTAAACGTTGGGCACTAGGTGTGGGGACC ATTCCACGGTTTCCGCGCCGCAGCTAACGCATTAAGTGCCCCGCCTGGGGAGTACGGCCGCAAGGCTAAAAC TCAAA

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



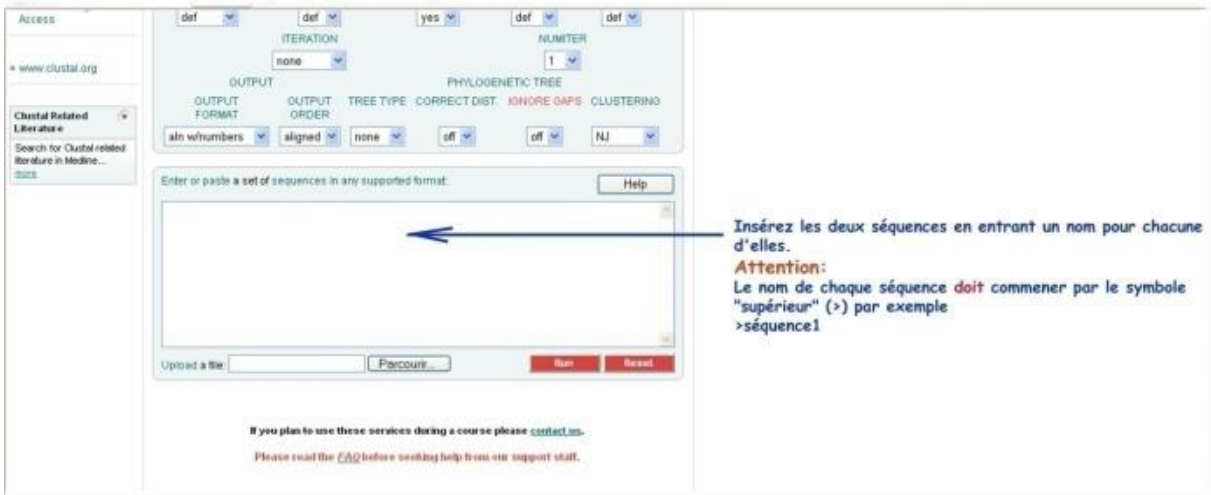
Etape 2 : Un menu déroulant va s'ouvrir. Il faut passer le pointeur de la souris sur "Sequence Analysis" pour découvrir un second menu. Sur ce second menu, cliquez sur " ClustalW2" :

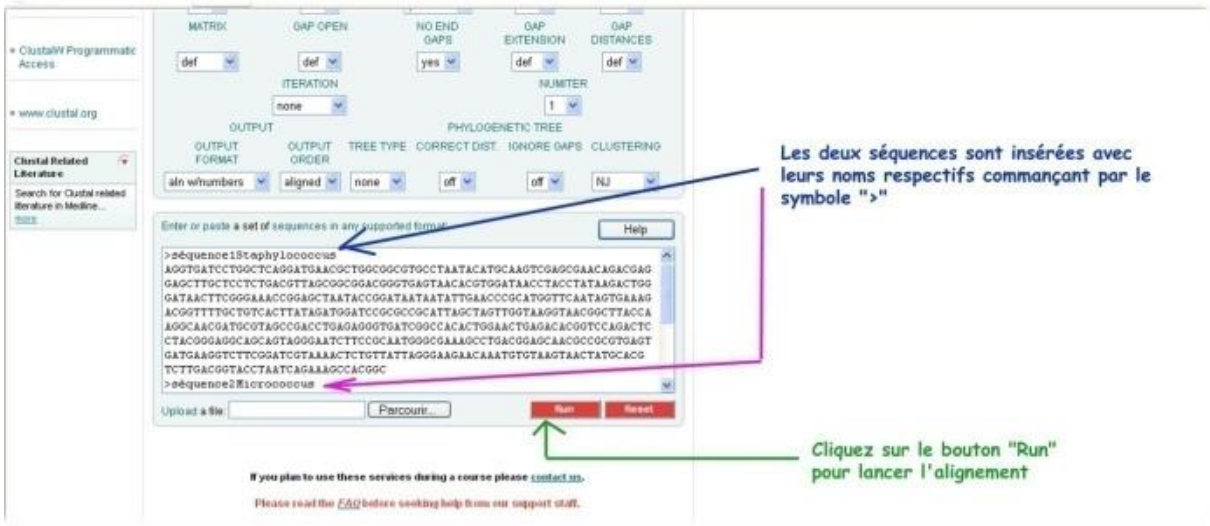


Etape 3 : En bas de la fenêtre qui s'ouvre, insérez les deux séquences en leur donnat chacune un nom significatif de la séquence. **CE NOM DOIT COMMENER PAR LE SYMBOLE "> " (Supérieur) :**

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

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

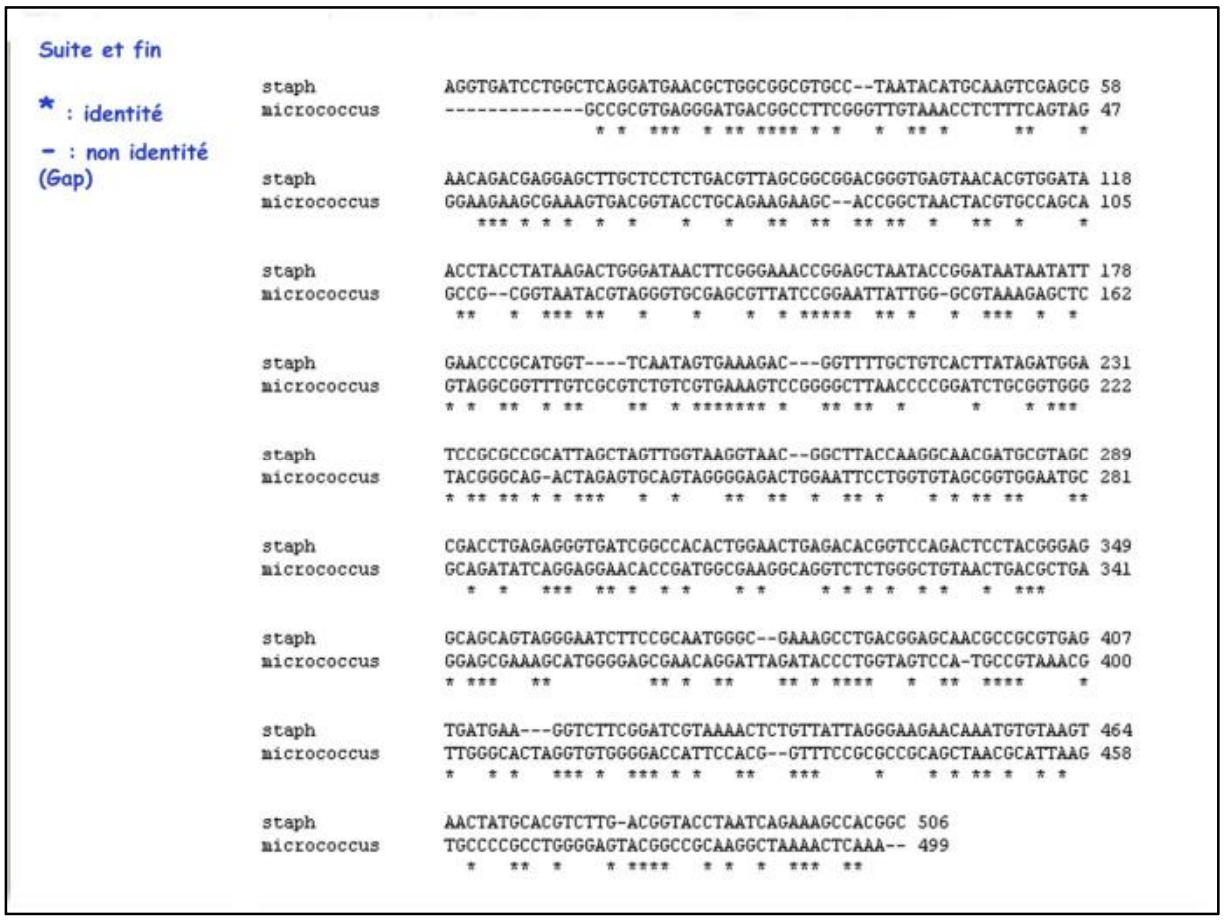




Etape 4 : Lecture du résultat. En début de la page des résultats, il y a lieu de constater la valeur du score global. Dans ce cas S = 1090.

Vous pouvez cliquer sur le bouton “Start Jalview” pour voir l’alignement en couleur et sur une même ligne.

En bas de la page des résultats vous pouvez voir l'alignement donné en plusieurs lignes (pour des raisons techniques et de formats) :

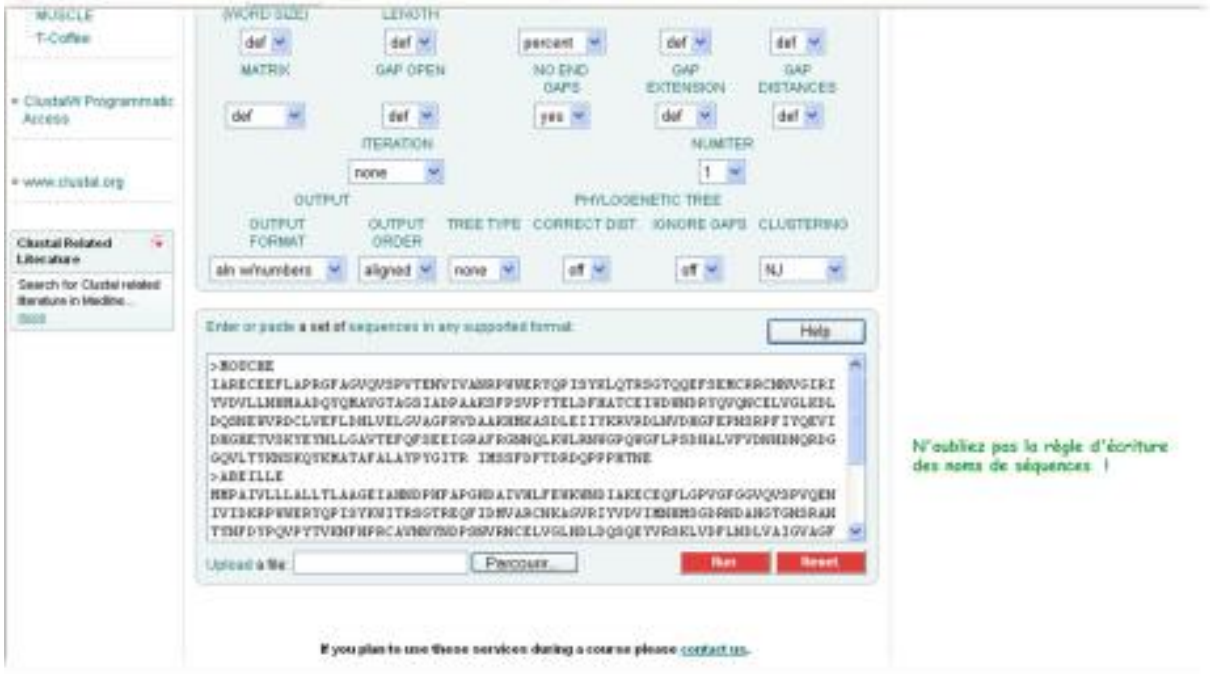


Application 2 : Il s'agit d'aligner, via Internet, deux séquences protéiques.

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

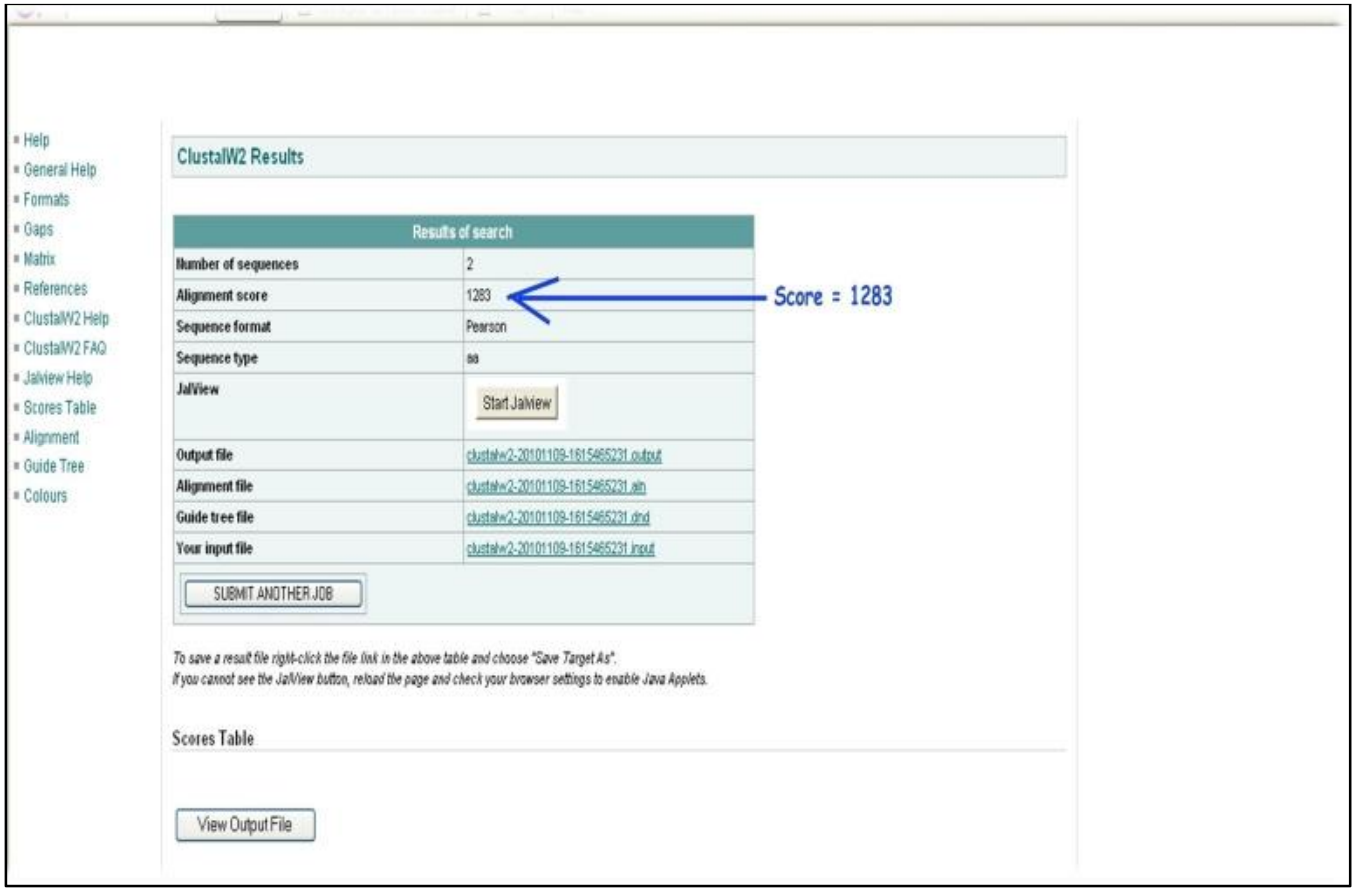
Nom systématique	Séquence : structure primaire
Mouche	IARECEEFLAPRGFAGVQVSPVTENVIVANRPWWERYQPISYKLQTRS GTQQEFSEMCRRCNNVGIRIYVDVLLNHMAADQYQMAVGTAGSIADP AAKSFPSVPYTELDFHATCEIWDWNDRYQVQNCELVGLKDLDQSNEW VRDCLVEFLDHLVELGVAGFRVDAAKHMKASDLEIIYKRVRDLNVDHGF EPNSRPFIYQEVIDHGHETVSKYEYNLLGAVTEFQFSEEIGRAFRGNNQ LKWLRNWGPQWGFLPSDHALVFVDNHDNQRDGGQVLTYKNSKQYK MATAFALAYPYGITR IMSSFDFTDRDQPPPHTNE
Abeille	MMPAIVLLLALLTLAAGEIAHNDPHFAPGHDAIVHLFEWKWNDIAKECE QFLGPVGFGGVQVSPVQENIVIDKRPWWERYQPISYKWITRSGTREQF IDMVARCNKAGVRIYVDVIMNHMSGDRNDAHGTGNSRANTYNFDYPQ VPYTVKNFHPRCAVNNYNDPSNVRNCELVGLHDLDQSQEYVRSKLVD FLNDLVAIGVAGFRVDAAKHMWPSDLRTIYSRVRNLNRTHGFPNDAQP YIFQEVIDYGNEAISKREYNGIGAVIEFKYSYEISNAFRGNNNLKWLVNW GEQWGFLPSKDSLVFVDNHDTQRDNPQILTYKYSKRYKMAVAFMLSH PFGTPRIMSSFDFQSKDQGPPNDGNGNILSPSIHDNICSNGWICEHRW RQIYNMVRFRNLVKGTKIDNWWDNGSNQIAFSRGCSGFVAFNGDQYD LKKNLKVCLPPGQYCDVISGNLEKGRCTGKIVTVGSDGNANIEIGAGEE DGVLAIHVKAKMA

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

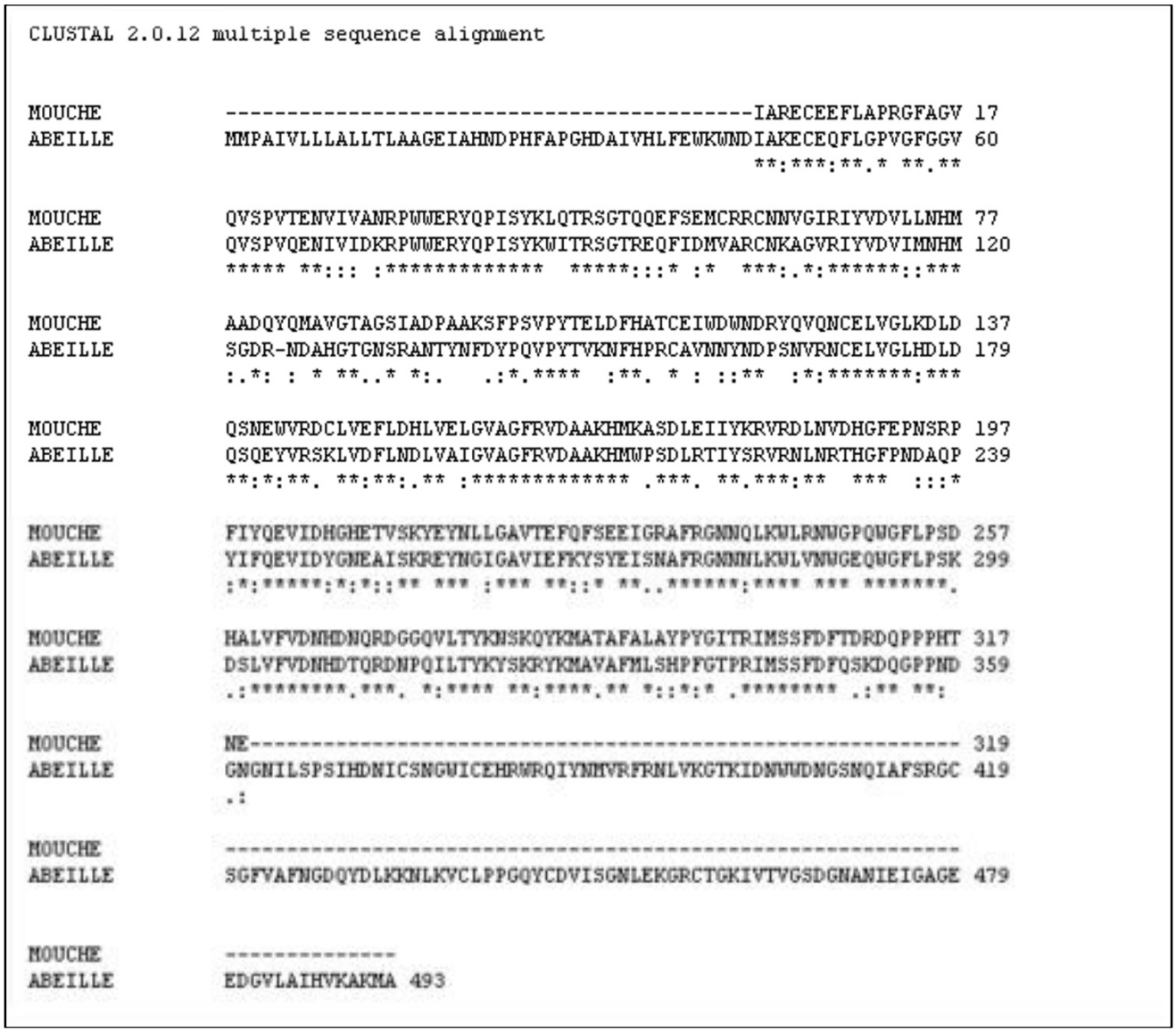


Une fois les deux séquences collées dans la zone et portant chacune un nom convenablement écrit, cliquez sur le bouton **"Run"** (En bas en rouge) :

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



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

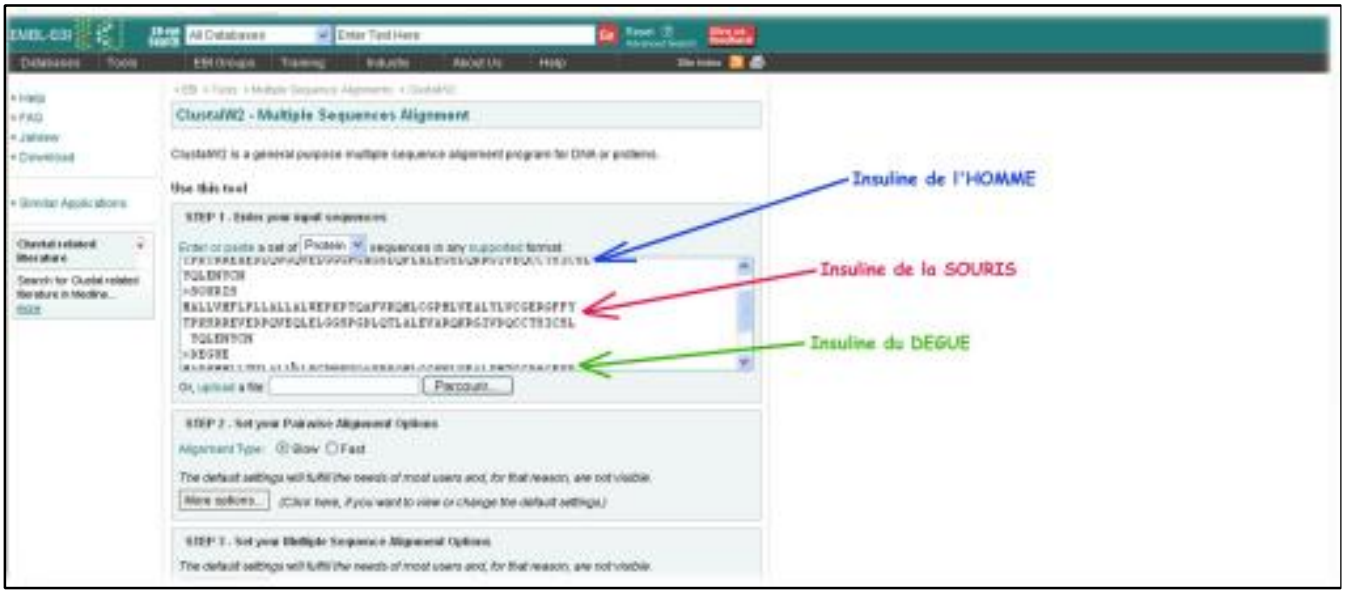


```
CLUSTAL 2.0.12 multiple sequence alignment


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

MOUCHE          QVSPVTENVIVANRPWWERYQPISYKLQTRSGTQQEFSEMCRRCNNVGIRIYVDVLLNHM 77
ABEILLE         QVSPVQENIVIDKRPWWERYQPISYKWITRSGTREQFIDMVARCNKAGVRIYVDVIMNHM 120
                ***** **::: :*************  *****:::* :*  ***:.*:******::***

MOUCHE          AADQYQMAVGTAGSIADPAAKSFPSVPYTELDFHATCEIWDWNDRYQVQNCELVGLKDLD 137
ABEILLE         SGDR-NDAHGTGNSRANTYNFDYPQVPYTVKNFHPRCAVNNYNDPSNVRNCELVGLHDLD 179
                :.*: : * **..* *:.   .:*.****  :**. * : ::**  :*:*******:***

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

MOUCHE          FIYQEVIDHGHETVSKYEYNLLGAVTEFQFSEEIGRAFRGNNQLKWLRNWGPQWGFLPSD 257
ABEILLE         YIFQEVIDYGNEAISKREYNGIGAVIEFKYSYEISNAFRGNNNLKWLVNWGEQWGFLPSK 299
                :*:*****:*:*::** *** :*** **::* **..******:**** *** *******.

MOUCHE          HALVFVDNHDNQRDGGQVLTYKNSKQYKMATAFALAYPYGITRIMSSFDFTDRDQPPPHT 317
ABEILLE         DSLVFVDNHDTQRDNPQILTYKYSKRYKMAVAFMLSHPFGTPRIMSSFDFQSKDQGPPND 359
                .:********.***. *:**** **:****.** *::*:* .******** .:** **:

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

MOUCHE          ------------------------------------------------------------
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 :

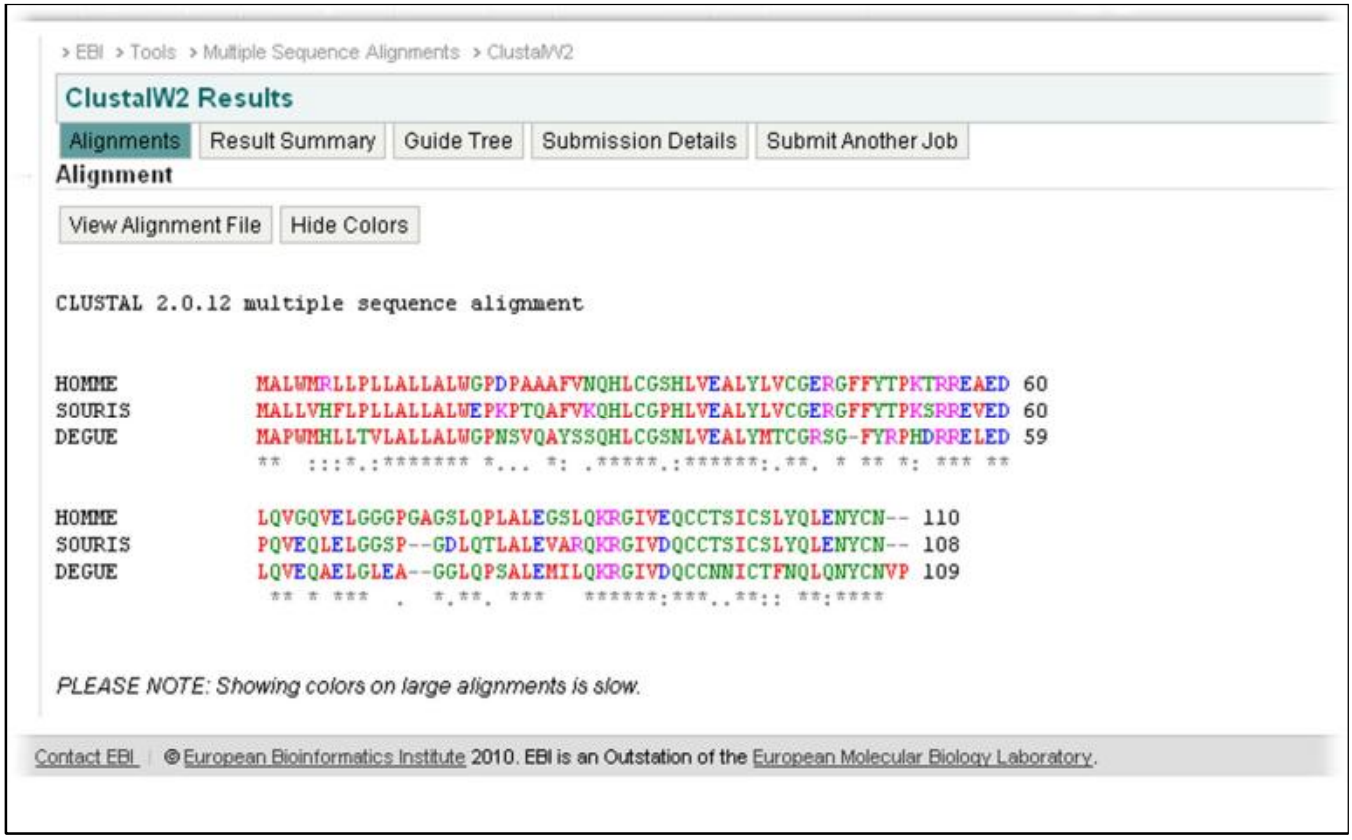
Homo sapiens **AAA59172** 110 acides aminés	MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFY TPKTRREAEDLQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSL YQLENYCN
Mus musculus **AAI45871** *108 acides aminés*	MALLVHFLPLLALLALWEPKPTQAFVKQHLCGPHLVEALYLVCGERGFFY TPKSRREVEDPQVEQLELGGSPGDLQTLALEVARQKRGIVDQCCTSICSL YQLENYCN
Octodon degus **AAA40590** 109 acides aminés	MAPWMHLLTVLALLALWGPNSVQAYSSQHLCGSNLVEALYMTCGRSGFYR PHDRRELEDLQVEQAELGLEAGGLQPSALEMILQKRGIVDQCCNNICTFN QLQNYCNVP

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.

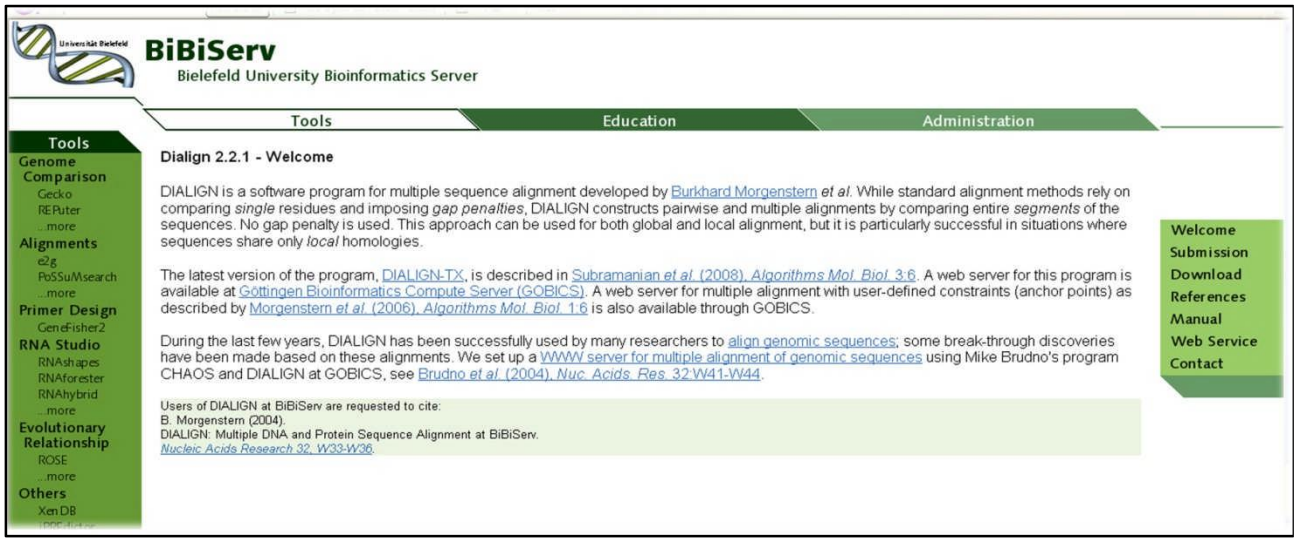


Remarque : Ce type d'alignement est également réalisable sur d'autres sites[12,13] ou avec des programmes téléchargeables. Il fera l'objet du prochain support pédagogique.

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