Alignement de deux séquences protéiques

Les acides aminés composant une protéine peuvent avoir des propriétés physico-chimiques similaires.

➡ La structure 3D dépend de ces caractéristiques

Une similitude au niveau de ces propriétés sera suffisante pour permettre la substitution d'un acide aminé en un autre sans perturber la fonction de la protéine (par exemple, échange de l'acide aminé hydrophobe valine en leucine).

Lors de la comparaison de deux séquences protéiques, nous devons prendre en compte ces similitudes et pas seulement les identités.

Comment quantifier la similitude entre deux acides aminés ?

- calculer une distance entre acides aminés basée sur leurs caractéristiques
- estimer la fréquence de substitution de l'acide aminé X en Y au cours de l'évolution

Les deux approches donnent une matrice (20,20) symétrique par rapport à la diagonale. Cependant, les matrices les plus utilisées ont été obtenues par la seconde approche et sont appelées « matrices de substitution »

Approches basée sur les caractéristiques des a.a.

Basée sur le code génétique : une substitution d'un a.a. en un autre se produit d'autant plus rarement que cela nécessite un plus grand nombre de mutations au niveau ADN.

Matrice génétique (Fitch, 1966)

Identité : +3
1 mutation ADN = 2 nt identiques : +2
2 mutations ADN = 1 nt identique : +1
3 mutations ADN = 0 nt identique : 0

le code génétique

Première lettre (côté 5')	Deuxième lettre: U		C		A		G		Troisième lettre (côté 3')
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys	C
	UUA	Leu	UCA	Ser	UAA	Stop	UGA	Stop	A
	UUG	Leu	UCG	Ser	UAG	Stop	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC	Leu	CCC	Pro	CAC	His	CGC	Arg	C
	CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg	A
	CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg	G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser	C
	AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg	A
	AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg	G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly	C
	GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly	A
	GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly	G
	codon d'initiation				codon de terminaison				

Basée sur les propriétés physico-chimiques des a.a. :

- composition, polarité, volume moléculaire (Grantham, 1974)
- volume et polarité (Miyata *et al.*, 1979)
- paramêtres de Chou et Fasman (structures secondaires), polarité et hydrophobicité (Rao, 1987)

Approches basée sur les fréquences de substitutions des a.a. au cours de l'évolution

Principe :

- les séquences homologues ont conservées des fonctions similaires
- deux a.a. se ressembleront d'autant plus que la fréquence de substitution observée est grande puisque ces substitutions n'auront pas modifié la fonction de la protéine
- il est possible d'estimer la fréquence avec laquelle un a.a. est remplacé par un autre au cours de l'évolution à partir de séquences homologues alignées

Principales approches :

- Comparaison directe des séquences (alignement global) : matrices PAM (Dayhoff, 1978)
- Comparaison des domaines protéiques (régions les plus conservées) : matrices **BLOSUM** (Henikoff et Henikoff, 1992)
- Alignement des séquences en comparant leur structure secondaire ou tertiaire

Matrices PAM

PAM : Point/Percent Accepted Mutation (Dayhoff, 1978)

Elle rend compte de deux processus :

- l'apparition de substitutions
- leur passage au travers du crible de la sélection.

Construction :

• 71 familles de protéines (environ 1300 séquences, Choix des séquences : très proches minimum 85% d'identité entre chaque paire de séquences de manière à éviter la présence de substitutions multiples).

• alignements multiples (global)

• reconstruction des arbres phylogénétiques et inférences des séquences ancêtres (1 572 mutations)

• pour chaque a.a. on compte le nombre de fois où il a été substitué

Exemple : pour *Val* combien de fois il est resté inchangé, et combien de fois il a été substitué par *Ala*, par *Arg*, …

•Cumul des mutations « acceptées » par la sélection naturelle au sein des différentes familles

• Calcul de la mutabilité des a.a. (propension d'un a.a. à être remplacé, f(Val->X))

• Calcul de la matrice de probabilités de mutation à partir des 2 étapes précédentes

• Pour chaque a.a., ex: Val -> Ala = mutabilité(Val) * cumul(Val->Ala) / nb(Val)

• Calcul de la matrice Lods (Log odd ratios) : PAM1

```
                                   1                   2                   3
                  1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9…
KAPPA
 1 HUMAN EU       T - V A A P S V F I F P P S D E Q - L K S - G T A S V V C L L N N F Y P - R E…
 2 MOUSE MOPC 21  A - D A A P T V S I F P P S S E Q - L T S - G G A S V V C F L N N F Y P - K D
 3 QAT S211       A - N A A P T V S I F P P S T Z Z - L A T - G G A S V V C L M N K.F Y P – R D
 4 84 RA881T 4135 D - P V A P T V L I F P P A A D Q - V A T - G T V T I V C V A N K Y F P - - D
 B9 RA881T        D P P I A P T V L L F P P S A D Q - L T T - Z T V T I V C V A N K F R P - D D

LAMBDA
 6 HUMAN SH       Q P K A A P S V T L F P P S S E E - L Q A - N K A T L V C L I S D F Y P - G A
 7 PIG            Q P K A A P T V N L F P P S S E E - L G T - N K A T L V C L I S D F Y P - G A
 8 MOUSE MOPC 104E Q P K S S P S V T L F P P S S E E - L T E - N K A T L V C T I T O F Y P - G V
 9 MOUSE MOPC 315 Q P K S T P T L T V F P P S S E E - L K E - N K.A T L V C L I S N F S P - G S
 …
```

Matrices PAM

PAM : Point/Percent Accepted Mutation (Dayhoff, 1978)

Construction :

- 71 familles de protéines
- alignements multiples (global)
- reconstruction des arbres phylogénétiques et inférences des séquences ancêtres (1 572 mutations)
- pour chaque a.a. on compte le nombre de fois où il a été substitué

  Exemple : pour *Val* combien de fois il est resté inchangé, et combien de fois il a été substitué par *Ala*, par *Arg*, …

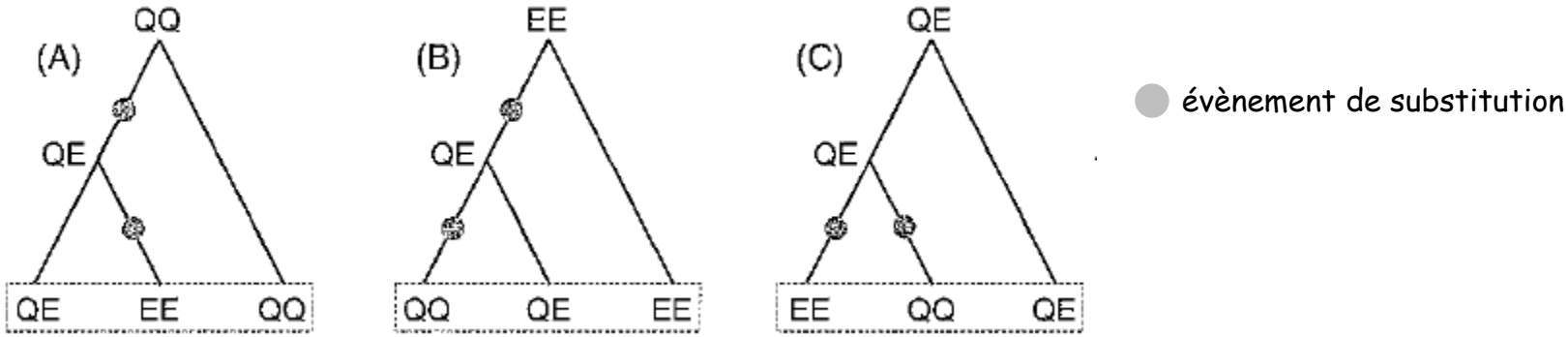
- Cumul des mutations « acceptées » par la sélection naturelle au sein des différentes familles
- Calcul de la mutabilité des a.a. (propension d'un a.a. à être remplacé, f(Val->X))
- Calcul de la matrice de probabilités de mutation à partir des 2 étapes précédentes
  - Pour chaque a.a., ex: Val -> Ala = mutabilité(Val) * cumul(Val->Ala) / nb(Val)
- Calcul de la matrice Lods (Log odd ratios) : PAM1

Principes généraux de la construction de la PAM

➢ Comptage du nombre de substitution pour chaque paire d'acides aminés :

Procédure :

- pour chaque famille alignement multiple des séquences et construction d'un arbre au moyen de la méthode du maximum de parcimonie qui permet de reconstruire les séquences ancêtres à chaque nœud de l'arbre.



Extrait de Perrière et Brochier-Armanet (2010) Concepts et méthodes en phylogénie moléculaire.

- Alignement de chaque séquence avec sa séquence ancêtre et comptage du nombre n_{ij} de substitution de l'acide aminé *i* vers l'acide aminé *j* et du nombre de conservations : obtention de la matrice 20x20 de résultats brutes qui est symétrique car on fait l'hypothèse de réversabilité $n_{ij} = n_{ji}$.

Matrice de cumul des mutations acceptées (x10)

	ala	arg	asn	asp	cys	gln	glu	gly	his	ile	leu	lys	met	phe	pro	ser	thr	trp	tyr	val
A																				
R	30																			
N	109	17																		
D	154	0	532																	
C	33	10	0	0																
Q	93	120	50	76	0															
E	266	0	94	831	0	422														
G	579	10	156	162	10	30	112													
H	21	103	226	43	10	243	23	10												
I	66	30	36	13	17	8	35	0	3											
L	95	17	37	0	0	75	15	17	40	253										
K	57	477	322	85	0	147	104	60	23	43	39									
M	29	17	0	0	0	20	7	7	0	57	207	90								
F	20	7	7 0	0	0	0	17	20	90	167	0	17								
P	345	67	27	10	10	93	40	49	50	7	43	43	4	7						
S	772	137	432	98	117	47	86	450	26	20	32	168	20	40	269					
T	590	20	169	57	10	37	31	50	14	129	52	200	28	10	73	696				
W	0	27	3	0	0	0	0	0	3	0	13	0	0	10	0	17	0			
Y	20	3	36	0	30	0	10	0	40	13	23	10	0	260	0	22	23	6		
V	365	20	13	17	33	27	37	97	30	661	303	17	77	10	50	43	186	0	17	

Matrices PAM

PAM : Point/Percent Accepted Mutation (Dayhoff, 1978)

Construction :

- 71 familles de protéines
- alignements multiples (global)
- reconstruction des arbres phylogénétiques et inférences des séquences ancêtres (1 572 mutations)
- pour chaque a.a. on compte le nombre de fois où il a été substitué

  Exemple : pour *Val* combien de fois il est resté inchangé, et combien de fois il a été substitué par *Ala*, par *Arg*, …

- Cumul des mutations « acceptées » par la sélection naturelle au sein des différentes familles
- Calcul de la mutabilité des a.a. (propension d'un a.a. à être remplacé, f(Val->X))
- Calcul de la matrice de probabilités de mutation à partir des 2 étapes précédentes
  - Pour chaque a.a., ex: Val -> Ala = mutabilité(Val) * cumul(Val->Ala) / nb(Val)
- Calcul de la matrice Lods (Log odd ratios) : PAM1

Matrices PAM

PAM : Point/Percent Accepted Mutation (Dayhoff, 1978)

Construction :

- 71 familles de protéines
- alignements multiples (global)
- reconstruction des arbres phylogénétiques et inférences des séquences ancêtres (1 572 mutations)
- pour chaque a.a. on compte le nombre de fois où il a été substitué

  Exemple : pour *Val* combien de fois il est resté inchangé, et combien de fois il a été substitué par *Ala*, par *Arg*, …

- Cumul des mutations « acceptées » par la sélection naturelle au sein des différentes familles
- Calcul de la mutabilité des a.a. (propension d'un a.a. à être remplacé, f(Val->X))
- Calcul de la matrice de probabilités de mutation à partir des 2 étapes précédentes
  - Pour chaque a.a., ex: Val -> Ala = mutabilité(Val) * cumul(Val->Ala) / nb(Val)
- Calcul de la matrice Lods (Log odd ratios) : PAM1

Principes généraux de la construction de la PAM

- Calcul de la *mutabilité* de chaque acide aminé *i*

La mutabilité est défini comme le rapport entre le nombre de substitutions affectant l'acide aminé *i* et le nombre d'acide aminé *i* observé dans les données :

$$m_i = \frac{\sum_{i \neq j} n_{ij}}{\sum_j n_{ij}}$$

- Prise ne compte du temps : calcul de la matrice de probabilité PAM1

Matrice particulière car correspond à un intervalle de temps *t* au cours duquel les acides aminés auront muté. La PAMk correspond à un intervalle de temps ou k% des acides aminés examinés auront muté. La matrice « de base » construite par Dayhoff correspond à la PAM1 dans laquelle seulement 1% des acides aminés étudiés auront mutés. Le facteur correctif ρ prenant en compte le temps a été calculé par Dayhoff de telle manière que les fréquences de conservation des acides aminés (termes diagonaux de la matrice) représentent une conservation de 99%. Les termes de la PAM1 sont donnés par :

$$\boxed{q_{ij_{i \neq j}} = \rho m_i \frac{n_{ij}}{\sum_{i \neq j} n_{ij}}}$$

et

$$\boxed{q_{ii} = 1 - \rho m_i}$$

Matrices PAM

Calcul de la matrice Lods (Log odd ratios) :

Permet de faire la somme des scores élémentaires pour un alignement plutôt que le produit des probabilités : log (a*b) = log a + log b

$$w_{i,j} = \log \frac{q_{ij}}{p_{ij}}$$

où:
q_{ij} est la fréquence observée de substitution de l'acide aminé *i* en *j*
p_{ij} est la fréquence théorique de substitution de l'acide aminé *i* en *j*

PAM1 : Normalisée pour avoir 1 mutation acceptée pour 100 a.a.

→ Temps qu'il faut pour qu'une mutation se fixe dans la population
= Distance évolutive conceptuelle : 1 PAM

Hypothèse : les probabilités de mutations sont indépendantes

PAM2 = PAM1 x PAM1
Matrice pour une distance évolutive de 2 PAM

De même, PAM40 = $PAM1^{40}$, PAM120 = $PAM1^{120}$, PAM250 = $PAM1^{250}$

Alignement de deux séquences protéiques
Matrices de substitution

La matrice PAM250

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	12																			
S	0	2																		
T	-2	1	3																	
P	-3	1	0	6																
A	-2	1	1	1	2															
G	-3	1	0	-1	1	5														
N	-4	1	0	-1	0	0	2													
D	-5	0	0	-1	0	1	2	4												
E	-5	0	0	-1	0	0	1	3	4											
Q	-5	-1	-1	0	0	-1	1	2	2	4										
H	-3	-1	-1	0	-1	-2	2	1	1	3	6									
R	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6								
K	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5							
M	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6						
I	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5					
L	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6				
V	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4			
F	-4	-3	-3	-5	-4	-5	-3	-6	-5	-5	-2	-4	-5	0	1	2	-1	9		
Y	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10	
W	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17

S – small hydrophilic

N- acid, acid amide, hydrophylic

H - basic

V – small hydrophobic

F- aromatic

Matrices PAM

Remarques :

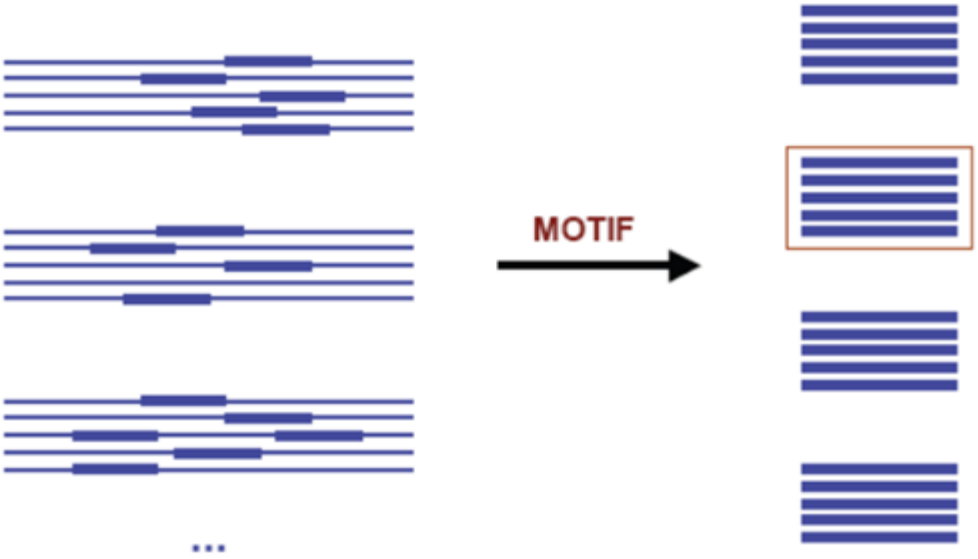
- Matrice calculée à partir de séquences ayant moins de 15% de divergence
- Biais dans la sélection des protéines (petites protéines globulaires)
- Actualisées : 16 130 séquences appartenant à 2 621 familles de protéines

Matrices BLOSUM (Henikoff et Henikoff, 1992)

BLOSUM : BLOcks SUbstituion Matrix

Principe :
- Obtention à partir de blocs de séquences alignées (alignement multiple sans brèche)
- Pour une paire d'a.a. : log (fréquence observée / fréquence attendue)

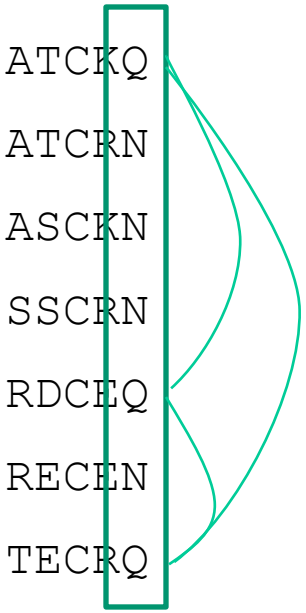
Avantages par rapport aux matrices PAM :
- contrairement aux matrices PAM, les matrices BLOSUM pour différentes distances évolutives sont obtenues directement avec des séquences plus ou moins divergentes
- l'utilisation de blocs plutôt que de séquences complètes : modélise les contraintes uniquement sur les régions conservées
- obtenues à partir d'un plus grand jeu de données (>2000 blocks, > 500 familles)



Matrices BLOSUM

Principe général du calcul :

1. Calcul de la fréquence observée de chaque paire



$q_{Q,N} = 12/21$
$q_{N,N} = 6/21$
$q_{Q,Q} = 3/21$

Si *n* séquences alignées alors :
n(n-1)/2 comparaisons

	A	C	D	E	K	N	Q	R	S	T
A										
C										
D										
E										
K										
N						6/21	12/21			
Q						12/21	3/21			
R										
S										
T										

Matrices BLOSUM

2. Calcul de la log odd matrice

$$w_{i,j} = \log \frac{q_{ij}}{p_{ij}}$$

où: q_{ij} est la fréquence observée de substitution de l'acide aminé i en j
p_{ij} est la fréquence théorique de substitution de l'acide aminé i en j

<u>Calcul de la fréquence théorique de chaque paire i,j</u>
Paires observées sont celles de la population :
Ex: 12 QN, 6 NN, et 3 QQ
La probabilité de Q dans une paire est :
p_Q = ((3+(12/2))/21 = 9/21
(3 paires avec Q aux deux positions et 12 paires avec Q à une position)
La probabilité de N dans une paire est :
p_N = ((6+(12/2))/21 = 12/21

```
ATCKQ
ATCRN
ASCKN
SSCRN
RDCEQ
RECEN
TECRQ
```

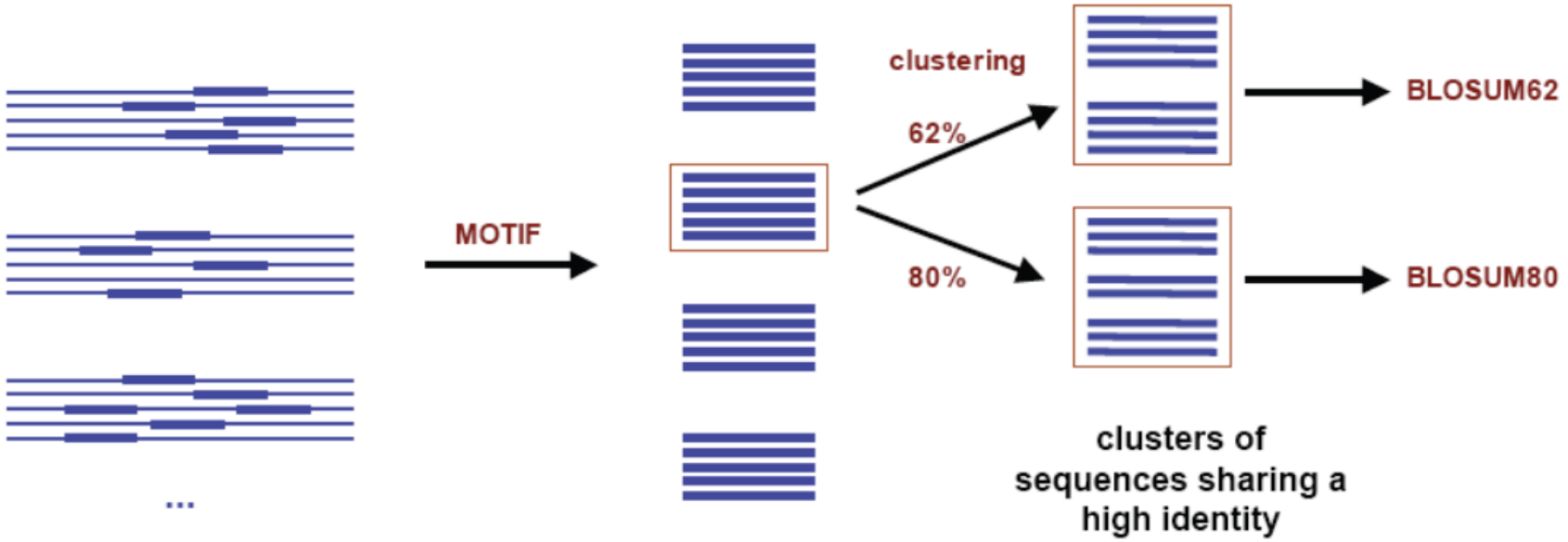
De manière générale :

$$p_i = q_{ii} + \sum_{i \neq j} \frac{q_{ij}}{2} \quad \text{et} \quad \begin{cases} p_{ii} = p_i^2 \\ p_{ij} = p_i p_j + p_j p_i = 2 p_i p_j \end{cases}$$

(une mutation peut aller dans les deux sens donc on considère p_{QN} et p_{NQ})
On a donc la fréquence théorique de substitution de Q vers N donnée par :
$p_{QN} = p_Q * p_N + p_N * p_Q = 2(p_Q * p_N) = 2(9/21 * 12/21) = 0.49$
La fréquence observée de substitution de Q vers N données par :
q_{QN} = 12/21 = 0,571

Matrices BLOSUM

Problème de la représentation des séquences dans les bases de données : certaines espèces et/ou familles de protéines sont plus fréquentes que d'autres -> biais.
Pour réduire ce biais du à l'échantillon dans le calcul de l'estimation des fréquences de substitutions d'un acide aminé X vers Y, les séquences sont regroupées en fonction de leur pourcentage d'identité. Leur importance dans le calcul sera ensuite pondéré par un poids basé sur leur nombre.
Par exemple, ci-dessous, les séquences dont l'identité est ≥ 62% et les séquences dont l'identité est ≥ 80%.
Pour les différents regroupements des matrices sont construites, le numéro de la BLOSUM indique que les estimations des fréquences de substitution ont été réalisées en regroupant les séquences ayant un pourcentage d'identité ≥ au numéro.



Matrices BLOSUM

Exemple : construction BLOSUM 50

On regroupe les séquences dont l'identité est ≥ 50

ATCKQ
ATCRN
ASCKN
SSCRN

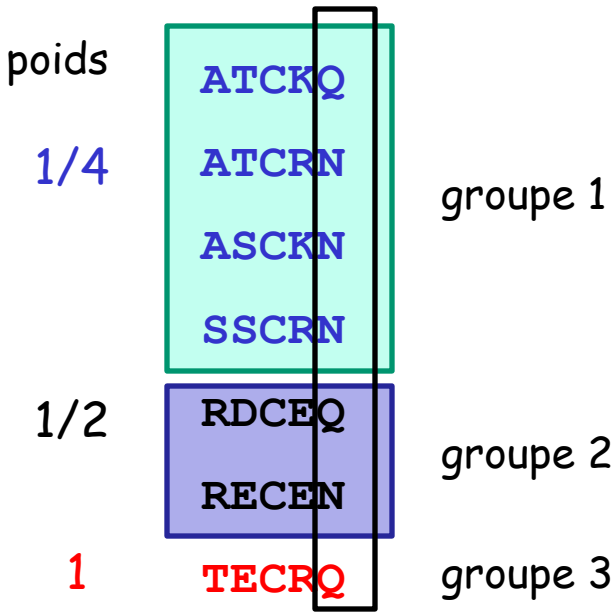
RDCEQ
RECEN

TECRQ

%identité	1	2	3	4	5	6	7
1	100	60	60	20	20	20	40
2		100	60	60	20	40	40
3			100	60	20	40	20
4				100	20	40	40
5					100	60	40
6						100	40

Matrices BLOSUM

Exemple : construction BLOSUM 50



	A	C	D	E	K	N	Q	R	S	T
A										
C										
D										
E										
K										
N						3/8	14/8			
Q						14/8	7/8			
R										
S										
T										

On ne compare plus qu'entre groupes et en pondérant par le poids du groupe:

Q -> N : 1 fois entre groupes 1 et 2 -> $\frac{1}{4} * \frac{1}{2}$

N -> Q : 3 fois entre groupes 1 et 2 -> $\frac{3}{4} * \frac{1}{2}$, 3 fois entre groupes 1 et 3 -> $\frac{3}{4} * 1$, 1 fois entre groupes 2 et 3 -> $\frac{1}{2} * 1$

Donc :

$$q_{Q,N} = \frac{1}{4} * \frac{1}{2} + \frac{3}{4} * \frac{1}{2} + \frac{3}{4} * 1 + \frac{1}{2} * 1 = 14/8$$

$$q_{N,N} = \frac{3}{4} * \frac{1}{2} = 3/8$$

$$q_{Q,Q} = \frac{1}{4} * \frac{1}{2} + \frac{1}{4} * 1 + \frac{1}{2} * 1 = 7/8$$

Alignement de deux séquences protéiques
Matrices de substitution

La matrice BLOSUM62

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	9																			
S	-1	4																		
T	-1	1	5																	
P	-3	-1	-1	7																
A	0	1	0	-1	4															
G	-3	0	-2	-2	0	6														
N	-3	1	0	-2	-2	0	6													
D	-3	0	-1	-1	-2	-1	1	6												
E	-4	0	-1	-1	-1	-2	0	2	5											
Q	-3	0	-1	-1	-1	-2	0	0	2	5										
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8									
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5								
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5							
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5						
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4					
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4				
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4			
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6		
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7	
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11

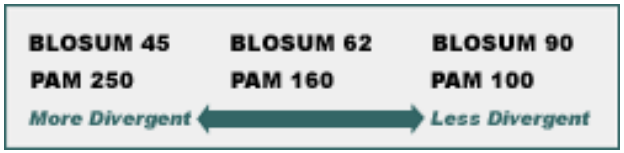
Alignement de deux séquences protéiques
Matrices de substitution

Famille de matrices correspondant à différentes distances évolutives entre les séquences :

PAM120 et BLOSUM80 : estimation des fréquences de substitution entre acides aminés pour des séquences proches dans l'évolution (courtes distances)

PAM250 et BLOSUM45 : estimation des fréquences de substitution entre acides aminés pour des séquences distantes dans l'évolution (longues distances)

PAM160 et BLOSUM62 : estimation des fréquences de substitution entre acides aminés pour des séquences ayant des distances évolutives intermédiaires.

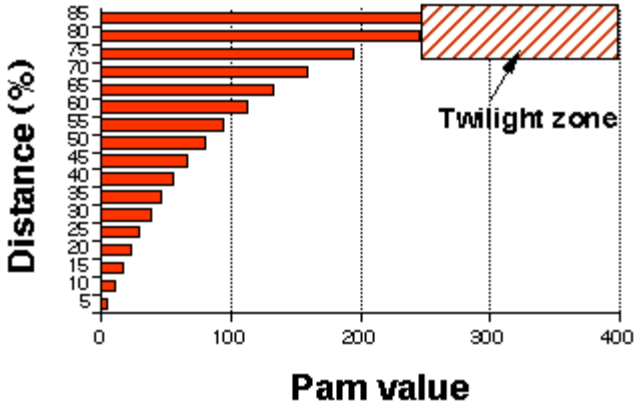


Source figure : ebi.ac.uk

longueur séquence	matrice	ouverture de gap	extension de gap
≥300	BLOSUM50	-10	-2
85-300	BLOSUM62	-7	-1
50-85	BLOSUM80	-16	-4
≥300	PAM250	-10	-2
85-300	PAM120	-16	-4

Recommandations (à adapter)

distance %	PAM
1	1
25	30
50	80
80	246



Source figure : Infobiogen.fr

Alignement de deux séquences protéiques

Ces matrices sont utilisées comme paramètres dans :

- les programmes d'alignement de deux séquences
- les recherches par similitude dans les bases de données
- les programmes d'alignement multiple

Dans le cas des alignements de deux séquences, elles remplacent les scores élémentaires correspondant à l'identité et à la substitution.

Pour calculer le score de la cellule (i,j) à partir de celui de la cellule (i-1,j-1), le poids $w(x_i,y_j)$ sera donné par la valeur de la substitution de l'acide aminé X en Y dans la matrice de substitution utilisée. Ce poids sera positif si l'échange des deux acides aminés a été favorisé au cours de l'évolution (acide aminés similaires) et il sera négatif si cette substitution a été contre sélectionnée. Ce système de score n'est donc pas si différent de celui utilisé pour la comparaison de séquences d'acides nucléiques dans lequel, l'identité recevait un score positif et la substitution un score négatif.

Donc quand on compare deux séquences protéiques :

- le pourcentage d'identité correspond au pourcentage d'acides aminés identiques
- le pourcentage de similitude (similarité) correspond au pourcentage d'acides aminés identiques et positifs (valeurs positives dans la matrice de substitution).

Alignement de deux séquences protéiques

Quelle matrice doit-on utiliser ?

Les matrices BLOSUM sont le plus souvent proposées comme matrices par défaut car les fréquences de substitution sont directement calculées à partir de l'alignement.

La BLOSUM62 est utilisée comme matrice par défaut car elle offre un bon compromis quand les distances évolutives entre les séquences ne sont pas connues.

La BLOSUM80 donnera de meilleurs résultats pour des séquences proches dans l'évolution. Elle tend à trouver des alignements courts fortement similaires.

La BLOSUM45 donnera de meilleurs résultats pour des séquences éloignées dans l'évolution. Elle trouvera de plus longs alignements locaux de faible conservation.

Effet du choix de la matrice de substitution

```
# Aligned_sequences: 2
# 1: PDC1_MAIZE
# 2: ILVB_ARATH
# Matrix: EBLOSUM62
# Gap_penalty: 12
# Extend_penalty: 2
#
# Length: 692
# Identity:     133/692 (19.2%)
# Similarity:   244/692 (35.3%)
# Gaps:        104/692 (15.0%)
# Score: -14

                                                  10
PDC1_M METLLAG--------------------------------NPANGVAKPT
       :                                      ::    .
ILVB_A MAAATTTTTTSSSISFSTKPSPSSSKSPLPISRFSLPFSLNPNKSSSSSR
                10        20        30        40        50

        20        30        40             50
PDC1_M CNGVGALPVANSHAIIATPAAAAATLAPAGAT-----LGRH---------
         :. .   ..  :.. :      : .:   :     . :
ILVB_A RRGIKSSSPSSISAVLNTTTNVTTTPSPTKPTKPETFISRFAPDQPRKGA
                60        70        80        90       100

               60        70        80        90       100
PDC1_M --LARRLVQIGASDVFAVPGDFNLTLLDYLIAEPGLTLVGCCNELNAGYA
         :   : . :   ::: ::  .. .   :     .  :   .:    .:
ILVB_A DILVEALERQGVETVFAYPGGASMEIHQALTRSSSIRNVLPRHEQGGVFA
               110       120       130       140       150

              110       120       130       140       150
PDC1_M ADGYARSRGV-GACAVTFTVGGLSVLNAIAGAYSENLPVVCIVGGPNSND
       :.::::: :  : :  :   :  .... .: :  ...:.: : :
ILVB_A AEGYARSSGKPGICIATSGPGATNLVSGLADALLDSVPLVAITGQVPRRM
               160       170       180       190       200

               160       170       180       190
PDC1_M YGTNRILHHTIGLPDFSQELRCFQTITCYQAIINNLDDAHEQIDTA--IA
        ::.      :        .   ..:: .  .. ...:    :. :  .:
ILVB_A IGTDAFQETPI--------VEVTRSITKHNYLVMDVEDIPRIIEEAFFLA
               210               220       230       240

      200       210        220          230       240
PDC1_M TALRESKPVYISVSCNLAG-LSHPTFS---RDPVPMFISPRLSNKANLEY
       :. :   :: . :  ..   :. : .    : :  :   :.    ..::
ILVB_A TSGRPG-PVLVDVPKDIQQQLAIPNWEQAMRLPGYMSRMPKPPEDSHLEQ
              250       260       270       280       290
```

```
# Aligned_sequences: 2
# 1: PDC1_MAIZE
# 2: ILVB_ARATH
# Matrix: EPAM30
# Gap_penalty: 12
# Extend_penalty: 2
#
# Length: 797
# Identity:     173/797 (21.7%)
# Similarity:   216/797 (27.1%)
# Gaps:        314/797 (39.4%)
# Score: -977

                   10        20                30
PDC1_M ME----TLLAGNPANGVAKPT-CNGVGALPVA--------NSH-------
       :     :           ::        ::.         :
ILVB_A MAAATTTTTTSSSISFSTKPSPSSSKSPLPISRFSLPFSLNPNKSSSSSR
                10        20        30        40        50

                             40             50
PDC1_M -------------AIIATPAAAAATLAPAGAT-----LGRHLA----RR-
                    :.  :      :  :   :       :  :    :
ILVB_A RRGIKSSSPSSISAVLNTTTNVTTTPSPTKPTKPETFISR-FAPDQPRKG
                60        70        80        90

           60            70        80        90       100
PDC1_M ---LVQI----GASDVFAVPGDFNLTLLDYLIAEPGLTLVGCCNELNAGY
          ::.     :   ::: ::   .     :        :    :    .
ILVB_A ADILVEALERQGVETVFAYPGGASMEIHQALTRSSSIRNVLPRHEQGGVF
     100       110       120       130       140

              110       120       130       140
PDC1_M AADGYARSRG-VGACAVTFTVGGLSVLNAIAGAYSENLPVVCIVGGPNSN
       ::.::::: :  : :  :   :        : :  .  : : :
ILVB_A AAEGYARSSGKPGICIATSGPGATNLVSGLADALLDSVPLVAI-------
     150       160       170       180       190

     150       160       170           180
PDC1_M DYGTNRILHHTIGLPDFSQELRCFQT----ITCYQAII--NNL----DDA
          :  .    ::   : ::     :    .:    :   : :    .:
ILVB_A ---TGQVPRRMIGTDAF-QE-----TPIVEVT--RSITKHNYLVMDVEDI
                200             210         220       230

     190         200       210                 220
PDC1_M HEQIDTA--IATALRESKPVYISVSCN----LA------GLSHPTF-SRD
          :. :   ::  :   :: . :  .    ::        .  : . ::
ILVB_A PRIIEEAFFLATSGRPG-PVLVDVPKDIQQQLAIPNWEQAMRLPGYMSR-
             240        250       260       270
```

Effet du choix de la matrice de substitution

```
# Aligned_sequences: 2
# 1: PDC1_MAIZE
# 2: ILVB_ARATH
# Matrix: EBLOSUM62
# Gap_penalty: 12
# Extend_penalty: 2
#
# Length: 692
# Identity:     133/692 (19.2%)
# Similarity:   244/692 (35.3%)
# Gaps:         104/692 (15.0%)
# Score: -14
                                                 10
PDC1_M METLLAG--------------------------------NPANGVAKPT
       :                                       ::    .
ILVB_A MAAATTTTTTSSSISFSTKPSPSSSKSPLPISRFSLPFSLNPNKSSSSSR
                10        20        30        40        50

        20        30        40             50
PDC1_M CNGVGALPVANSHAIIATPAAAAATLAPAGAT-----LGRH---------
         :. .   ..  :.. :      : .:   :     . :
ILVB_A RRGIKSSSPSSISAVLNTTTNVTTTPSPTKPTKPETFISRFAPDQPRKGA
                60        70        80        90       100

               60        70        80        90       100
PDC1_M --LARRLVQIGASDVFAVPGDFNLTLLDYLIAEPGLTLVGCCNELNAGYA
         :   : . :   ::: ::  .. .   :     .  :   .:    .:
ILVB_A DILVEALERQGVETVFAYPGGASMEIHQALTRSSSIRNVLPRHEQGGVFA
               110       120       130       140       150

              110       120       130       140       150
PDC1_M ADGYARSRGV-GACAVTFTVGGLSVLNAIAGAYSENLPVVCIVGGPNSND
       :.::::: :  : :  :   :  .... .: :  ...:.: : :
ILVB_A AEGYARSSGKPGICIATSGPGATNLVSGLADALLDSVPLVAITGQVPRRM
               160       170       180       190       200

               160       170       180       190
PDC1_M YGTNRILHHTIGLPDFSQELRCFQTITCYQAIINNLDDAHEQIDTA--IA
        ::.     :        .   ..:: .  .. ...:    :. :  .:
ILVB_A IGTDAFQETPI--------VEVTRSITKHNYLVMDVEDIPRIIEEAFFLA
               210               220       230       240

      200       210       220          230       240
PDC1_M TALRESKPVYISVSCNLAG-LSHPTFS---RDPVPMFISPRLSNKANLEY
       :. :   :: . :  ..   :. : .    : :  :   :.    ..::
ILVB_A TSGRPG-PVLVDVPKDIQQQLAIPNWEQAMRLPGYMSRMPKPPEDSHLEQ
              250       260       270       280       290
```

```
# Aligned_sequences: 2
# 1: PDC1_MAIZE
# 2: ILVB_ARATH
# Matrix: EPAM350
# Gap_penalty: 12
# Extend_penalty: 2
#
# Length: 700
# Identity:     133/700 (19.0%)
# Similarity:   360/700 (51.4%)
# Gaps:         120/700 (17.1%)
# Score: 396
                10             20
PDC1_M METLLAGNPANGV----AKPT-CNGVGALPVAN----------------
       :..  .......     .::.  .. ..::...
ILVB_A MAAATTTTTTSSSISFSTKPSPSSSKSPLPISRFSLPFSLNPNKSSSSSR
                10        20        30        40        50

                   30        40             50
PDC1_M -----------SHAIIATPAAAAATLAPAGAT-----LGRH---------
                    :.. :..  ..: .:. .:     ..:
ILVB_A RRGIKSSSPSSISAVLNTTTNVTTTPSPTKPTKPETFISRFAPDQPRKGA
                60        70        80        90       100

               60        70        80        90       100
PDC1_M --LARRLVQIGASDVFAVPGDFNLTLLDYLIAEPGLTLVGCCNELNAGYA
         :   : . :   ::: ::. .. . . :   ...  :   .: .. .:
ILVB_A DILVEALERQGVETVFAYPGGASMEIHQALTRSSSIRNVLPRHEQGGVFA
               110       120       130       140       150

              110       120       130       140       150
PDC1_M ADGYARSRG-VGACAVTFTVGGLSVLNAIAGAYSENLPVVCIVGGPNSND
       :.::::: :  : :  : . :. ......:.:  ...:.: : :
ILVB_A AEGYARSSGKPGICIATSGPGATNLVSGLADALLDSVPLVAITG------
               160       170       180       190

               160         170       180       190
PDC1_M YGTNRILHHTIGLPDFSQE--LRCFQTITCYQAIINNLDDAHEQIDTA--
           .. .. ::  .: .   .   ..::  . .. ...:    :. :
ILVB_A ----QVPRRMIGTDAFQETPIVEVTRSITKHNYLVMDVEDIPRIIEEAFF
               200       210       220       230       240

        200       210       220       230       240
PDC1_M IATALRESKPVYISVSCNLAG-LSHPTFSRD-PVPMFISPRLSNKANLEY
       .::. : . :: ..:. ..   :. :.. ..  .: ..: :... .. .
ILVB_A LATSGRPG-PVLVDVPKDIQQQLAIPNWEQAMRLPGYMS-RMPKPPE-DS
                250       260       270       280
```

Effet de la pénalité des indels

```
# 1: ILV1_TOBAC
# 2: ILVB_ARATH
# Matrix: EPAM60
# Gap_penalty: 2
# Extend_penalty: 2
#
# Length: 715
# Identity:     531/715 (74.3%)
# Similarity:   586/715 (82.0%)
# Gaps:          93/715 (13.0%)
# Score: 3415

                 10          20          30          40
ILV1_T MAAAAPSP--SSS-AFS-KTLSPSSSTSSTLLP--RSTF--PFP-HHPHK
       :::::. .   ::: .:: :  ::::: :  : :  :  :  ::  . :.:
ILVB_A MAAATTTTTTSSSISFSTKP-SPSSSKSP-L-PISR--FSLPFSLN-PNK
              10        20          30          40

              50        60              70
ILV1_T TTPPPLHLTHTHIHIHSQRRR-F-T-----ISNVIST--NQKV----SQT
       ..      . .     : :::   .     :: : .:  :  :    : :
ILVB_A SS------S-S-----S-RRRGIKSSSPSSISAVLNTTTN--VTTTPSPT
                       50        60        70

       80          90       100        110       120
ILV1_T EK-T--ETFVSRFAPDEPRKGSDVLVEALEREGV-TDVFAYPGGASMEIH
        : :  :::.::::::.::::.:.::::::::.:: : :::::::::::::
ILVB_A -KPTKPETFISRFAPDQPRKGADILVEALERQGVET-VFAYPGGASMEIH
       80        90       100       110        120

          130        140       150        160       170
ILV1_T QALTRSS-IIRNVLPRHEQGGVFAAEGYARATG-FPGVCIATSGPGATNL
       ::::::: : :::::::::::::::::::::..:  ::.::::::::::::
ILVB_A QALTRSSSI-RNVLPRHEQGGVFAAEGYARSSGK-PGICIATSGPGATNL
       130        140       150       160        170

            180       190       200       210       220
ILV1_T VSGLADALLDSVPIVAITGQVPRRMIGTDAFQETPIVEVTRSITKHNYLV
       :::::::::::: :::::::::::::::::::::::::::::::::::::
ILVB_A VSGLADALLDSVPLVAITGQVPRRMIGTDAFQETPIVEVTRSITKHNYLV
         180       190       200       210       220

            230         240       250       260       270
ILV1_T MDVEDIPRVVRE-AFFLA-RSGRPGPILIDVPKDIQQQLVIPDWDQPMRL
       :::::::::.. : :::::  :::::::.:.::::::::::: ::.:.: :::
ILVB_A MDVEDIPRII-EEAFFLAT-SGRPGPVLVDVPKDIQQQLAIPNWEQAMRL
         230        240        250       260       270
```

```
# 1: ILV1_TOBAC
# 2: ILVB_ARATH
# Matrix: EPAM60
# Gap_penalty: 12
# Extend_penalty: 2
#
# Length: 683
# Identity:     520/683 (76.1%)
# Similarity:   575/683 (84.2%)
# Gaps:          29/683 ( 4.2%)
# Score: 3275

                  10        20        30        40
ILV1_T MAAAAPS---PSSSAFSKTLSPSSSTSSTLLPRSTFPFPHHPHKTTPPPL
       ::::. .    :: .::   ::::: :     : . ::  .:.:..
ILVB_A MAAATTTTTTSSSISFSTKPSPSSSKSPLPISRFSLPFSLNPNKSSS---
               10        20        30        40

        50        60              70              80
ILV1_T HLTHTHIHIHSQRRR-------FTISNVISTNQKVSQTE------KTETF
                : :::        .:: : .:   :. :       : :::
ILVB_A ----------SSRRRGIKSSSPSSISAVLNTTTNVTTTPSPTKPTKPETF
                 50        60        70        80

           90       100       110       120       130
ILV1_T VSRFAPDEPRKGSDVLVEALEREGVTDVFAYPGGASMEIHQALTRSSIIR
       .::::::.::::.:.::::::::.::  :::::::::::::::::::: ::
ILVB_A ISRFAPDQPRKGADILVEALERQGVETVFAYPGGASMEIHQALTRSSSIR
        90       100       110       120       130

          140       150       160       170       180
ILV1_T NVLPRHEQGGVFAAEGYARATGFPGVCIATSGPGATNLVSGLADALLDSV
       ::::::::::::::::::..: ::.::::::::::::::::::::::::
ILVB_A NVLPRHEQGGVFAAEGYARSSGKPGICIATSGPGATNLVSGLADALLDSV
       140       150       160       170       180

          190       200       210       220       230
ILV1_T PIVAITGQVPRRMIGTDAFQETPIVEVTRSITKHNYLVMDVEDIPRVVRE
       : ::::::::::::::::::::::::::::::::::::::.. :
ILVB_A PLVAITGQVPRRMIGTDAFQETPIVEVTRSITKHNYLVMDVEDIPRIIEE
       190       200       210       220       230

          240       250       260       270       280
ILV1_T AFFLARSGRPGPILIDVPKDIQQQLVIPDWDQPMRLPGYMSRLPKLPNEM
       ::::: ::::::.:.::::::::::: ::.:.: ::::::::::.:: : .
ILVB_A AFFLATSGRPGPVLVDVPKDIQQQLAIPNWEQAMRLPGYMSRMPKPPEDS
       240       250       260       270       280
```

Alignement global versus Alignement local

```
# Aligned_sequences: 2
# 1: frag_new
# 2: ILV1_TOBAC
# Matrix: EBLOSUM45
# Gap_penalty: 12
# Extend_penalty: 2
#
# Length: 667
# Identity:      40/667 ( 6.0%)
# Similarity:    56/667 ( 8.4%)
# Gaps:         576/667 (86.4%)
# Score: -1062
```

Frag_n : 83 aa
ILV1_T : 667 aa

```
frag_n M-----------------------ETLL----------------------
       :                        :::
ILV1_T MAAAAPSPSSSAFSKTLSPSSSTSSTLLPRSTFPFPHHPHKTTPPPLHLT
               10        20        30        40        50


frag_n --------------------------------------------------

ILV1_T HTHIHIHSQRRRFTISNVISTNQKVSQTEKTETFVSRFAPDEPRKGSDVL
               60        70        80        90       100


frag_n --------------------------------------------------

ILV1_T VEALEREGVTDVFAYPGGASMEIHQALTRSSIIRNVLPRHEQGGVFAAEG
              110       120       130       140       150

              10
frag_n ----AGNPA-----------NGVS-----------------------IG-
            : :            : ::                        ::
ILV1_T YARATGFPGVCIATSGPGATNLVSGLADALLDSVPIVAITGQVPRRMIGT
              160       170       180       190       200


frag_n --------------------------------------------------

ILV1_T DAFQETPIVEVTRSITKHNYLVMDVEDIPRVVREAFFLARSGRPGPILID
              210       220       230       240       250


frag_n -------------WS-----------------------------------
                    :
ILV1_T VPKDIQQQLVIPDWDQPMRLPGYMSRLPKLPNEMLLEQIVRLISESKKPV
              260       270       280       290       300
```

```
                                20                      30
frag_n ------------------------VGATLGYAGAV--------------S
                                : .::   ::
ILV1_T LYVGGGCSQSSEDLRRFVELTGIPVASTLMGLGAFPTGDELSLSMLGMHG
              310       320       330       340       350

               40        50
frag_n TTFCAEIVESADAYLFAGPIFND---------------------------
       : .    :.:.:  :  :  :.:
ILV1_T TVYANYAVDSSDLLLAFGVRFDDRVTGKLEAFASRAKIVHIDIDSAEIGK
              360       370       380       390       400

frag_n ---------------------------------YSSWQEN----------
                                        .:.:..
ILV1_T NKQPHVSICADIKLALQGLNSILESKEGKLKLDFSAWRQELTEQKVKHPL
              410       420       430       440       450


frag_n ------DQCP--Y----------------------------------RT
             :  :  :                                   :
ILV1_T NFKTFGDAIPPQYAIQVLDELTNGNAIISTGVGQHQMWAAQYYKYRKPRQ
              460       470       480       490       500

                                                 70
frag_n W--------------------------------------HITSITT---
       :                                      ..  . :
ILV1_T WLTSGGLGAMGFGLPAAIGAAVGRPDEVVVDIDGDGSFIMNVQELATIKV
              510       520       530       540       550

                                     80
frag_n ---------------------------NDYAHV-------EAB-----CK
                                 . ::        ::.      :
ILV1_T ENLPVKIMLLNNQHLGMVVQWEDRFYKANRAHTYLGNPSNEAEIFPNMLK
              560       570       580       590       600

                              90
frag_n F---------------------ERME------------------------
       :                     ..:
ILV1_T FAEACGVPAARVTHRDDLRAAIQKMLDTPGPYLLDVIVPHQEHVLPMIPS
              610       620       630       640       650


frag_n -----------------

ILV1_T GGAFKDVITEGDGRSSY
              660
```

Alignement global versus Alignement local

Frag_n : 83 aa
ILV1_T : 667 aa

```
# Aligned_sequences: 2
# 1: frag_new
# 2: ILV1_TOBAC
# Matrix: EBLOSUM45
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 97
# Identity:      25/97 (25.8%)
# Similarity:    37/97 (38.1%)
# Gaps:          16/97 (16.5%)
# Score: 72.5
#
#
#=======================================

           10        20                30        40
frag_n LAGNPANGVSIGWSVGA-------TLGYAGAVSTTFCAEIVESADAYLFA
       : : :     .:  .::       .:   :   : .    :.:.:  :
ILV1_T LTGIPVASTLMG--LGAFPTGDELSLSMLGMHGTVYANYAVDSSDLLLAF
     320       330         340       350       360

        50         60        70            80
frag_n GPIFNDYSSWQ-ENDQCPYRTWHI----TSITTNDYAHVE--ABCKF
       :  :.:  . . :      .  ::      :  :   ::   :. :.
ILV1_T GVRFDDRVTGKLEAFASRAKIVHIDIDSAEIGKNKQPHVSICADIKL
       370       380       390       400       410
```