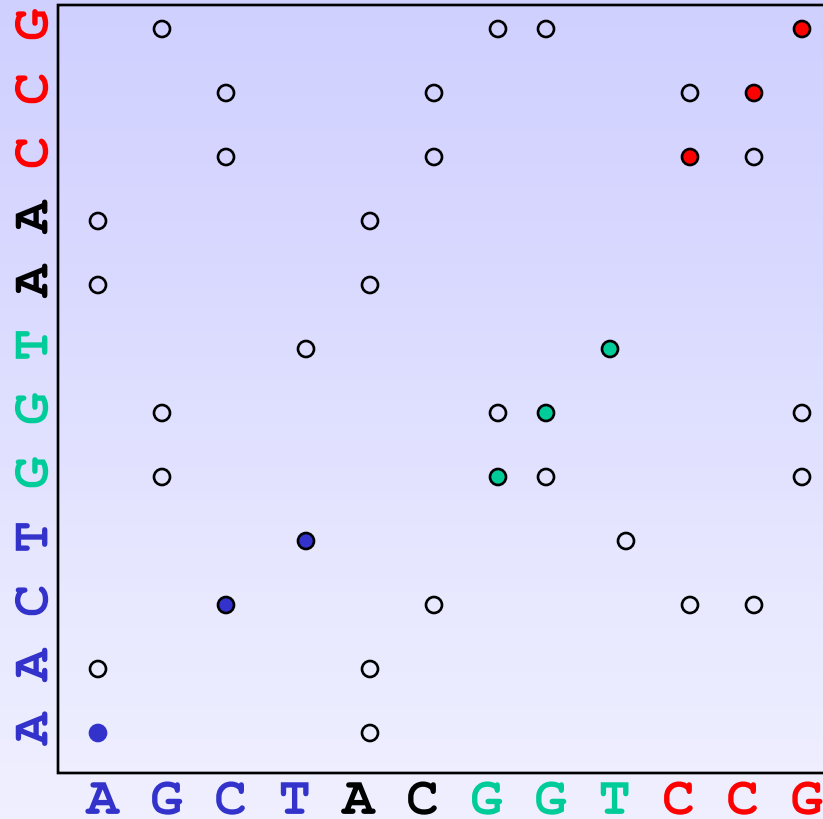
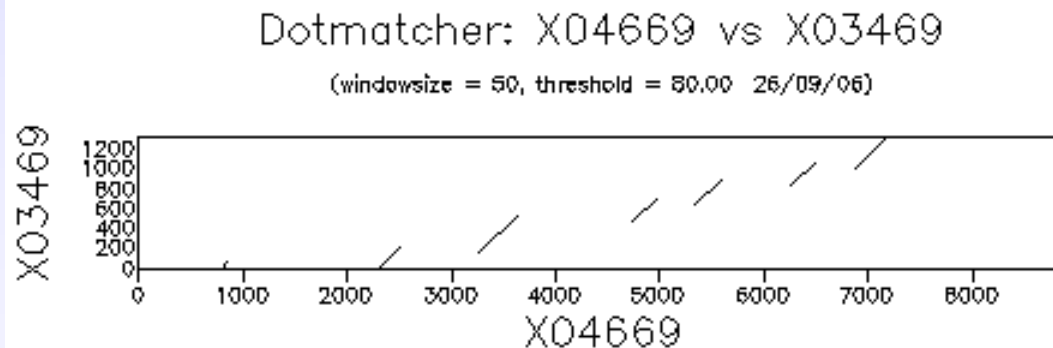
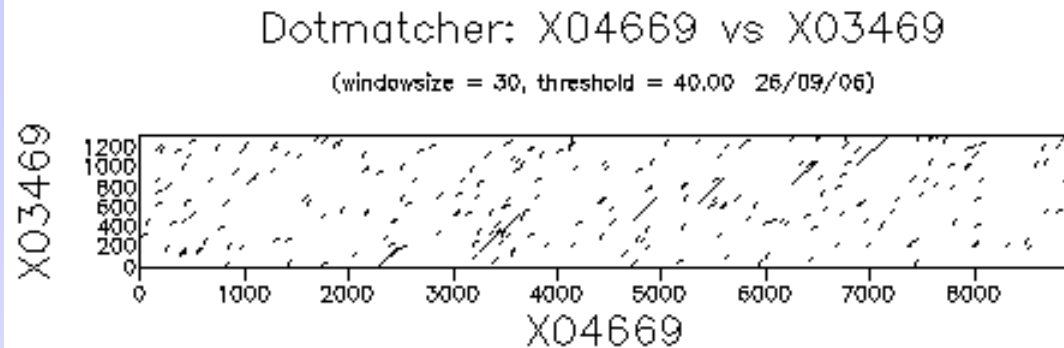


# Comparaison de deux séquences

# Matrice de points (Dotplot)

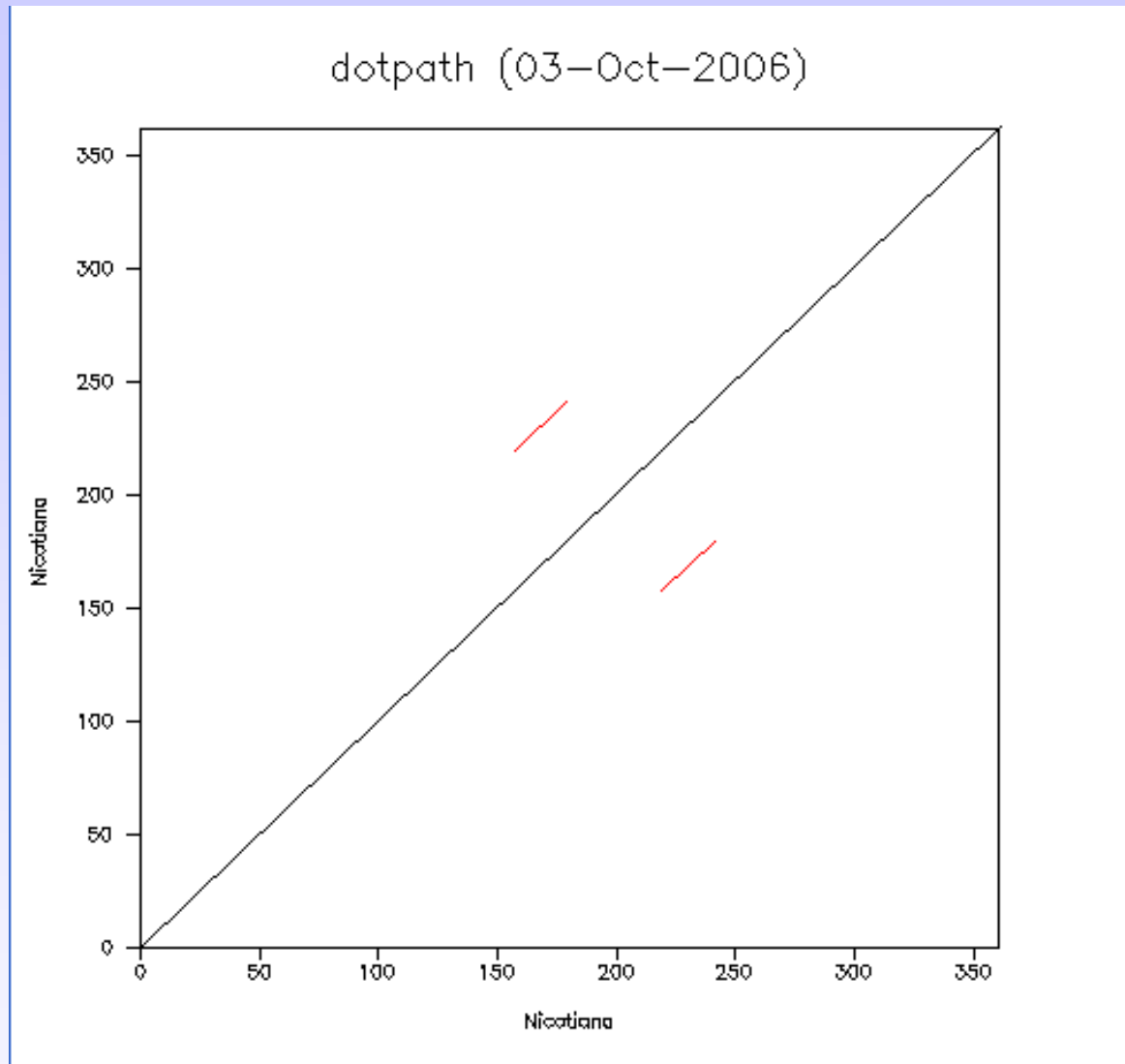


# Exemple d'utilisation d'un dotplot : mise en évidence des exons

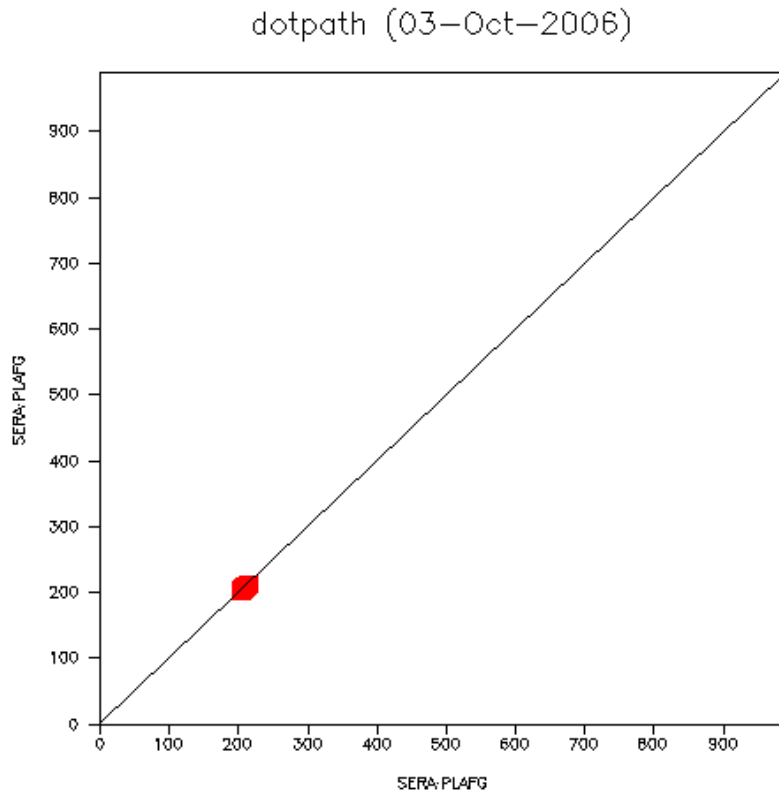


Effet des paramètres  
pour filtrer le bruit de  
fond

# Exemple d'utilisation d'un dotplot : mise en évidence de répétitions



# Exemple d'utilisation d'un dotplot : mise en évidence de « structure »



```
>SERA_PLAFG (P13823): Plasmodium falciparum serine-
repeat antigen protein precursor.
MKSYSISLFFILCVIFNKNVIKCTGESQTGNTGGGQAGNTVGDQAGSTGGSPQG
STGASQPGSSEPSNPVSSGHSVSTVSVSQTSTSSSEKQDTIQVKSALLKDYMGL
KVTGPCNENFIMFLVPHIYIDVDTEDTNIELRTTLKETNNAISFESNSGSLEK
KKYVLPSTNGTTGEQGSSTGTVRGDTETPISDSSSSSSSSSSSSSSSSSSSSSS
SSSSSSSSSSSESLPANGPDSPTVKPPRNLQNICETGKNFKLVVYIKENTLI
IKWKVYGETKDDTENNKVDVRKYLINEKETPFTSILIHAYKEHNGTNLIESKN
YALGSDIPEKCDTLASNCFLSGNFNIEKCFQCALLVEKENKNDVCYKYLSEDI
VSNFKEIKAETEDDDDDDYTEYKLTESIDNILVKMFKTNNNDKSELIKLEEV
DDSLKLELMNYCSLLKDVDTTGTLDNYGMGNEMDIFNNLKRLLIYHSEENINT
LKNKFRNAAVCLKNVDDWIVNKRGLVLPENLYDLEYFNEHLYNDKNSPEDKDN
KGKGVVHVDTTLEKEDTLSYDNSDNMFCNKEYCNRLKDENNCISNLQVEDQGN
CDTSWIFASKYHLETIRCMKGYEPTKISALYVANCYKGEHKDRCDGSSPMEF
LQIIEDYGFPLPAESNYPYNYVKVGEQCPKVEDHWMNLWDNGKILHNKNEPNSL
DGKGYTAYESERFHDNMDAFVKIIEKTEVMNKGSVIAYIKAENVMGYEFSGKKV
QNLGDDTADHAVNIVGYGNYVNSEGEKKS YWIVRNSWGPYWDEGYFKVDMY
GPTHCHFNFVIFSVVIFNVLDLPMNNKTTKESKIYDYLLKASPEFYHNLYFKNF
NVGKKNLFSEKEDNENNNKLGNNYIIFGQDTAGSGQSGKESNTALESAGTSNE
VSERVHVYHILKHIKDGKIRMGMRKYIDTQDVNKKHSCTRSYAFNPENYEKCV
NLCNVNWKTCCEKTSPGLCLSKLDTNNECYFCYV
```

# Alignement de deux séquences

La matrice de point est un bon outil visuel permettant de détecter les régions conservées entre deux séquences mais elle est insuffisante car on ne peut pas quantifier les similitudes observées

➡ Calcul d'un score

Alignement déduit du premier dotplot:

```
AACT--GGTAACCG
AGCTACGGT--CCG
```

Le score de l'alignement doit prendre en compte toutes les positions alignées : identités, substitutions et indels. Chacun de ces événements va recevoir un poids, appelé score élémentaire  $s_e$ . Le score de l'alignement correspondra à la somme des scores élémentaires correspondant aux positions alignées.

$$S = \sum_{i=1}^l s_e(i)$$

Où  $l$  est le nombre de positions alignées

exemple:  $l = 14$

$s_e$  identité = +2

$s_e$  substitution = -1

$s_e$  indels = -2



$S = 9$

# Algorithme de programmation dynamique

Etant donné un système de score, garantit l'obtention de l'alignement optimal

Hypothèse : l'évolution est parcimonieuse

Signification: pour trouver l'alignement optimal, l'algorithme va rechercher le chemin permettant de passer d'une séquence à l'autre avec le minimum de changements

Deux types de score en fonction des algorithmes :

- score d'homologie: la valeur du score diminue avec le nombre de différences observées entre les deux séquences
- score de distance: la valeur du score augmente avec le nombre de différences observées entre les deux séquences

Exemples de systèmes de scores

|          | Score d'homologie | Score de distance |
|----------|-------------------|-------------------|
| identité | +1                | 0                 |
| mismatch | -1                | +1                |
| indel    | -2                | +2                |

# Algorithme de programmation dynamique

Trois types d'algorithmes d'alignement de deux séquences:

- alignement global (proposé en premier par Needleman and Wunsch). Les séquences vont être alignées sur toutes leurs longueurs (du premier au dernier résidus). Utilisé quand les séquences ont à peu près la même longueur
- alignement semi-global (pas de pénalités des gaps aux extrémités). Utilisé quand une séquence est plus courte que l'autre ou quand on recherche des chevauchements aux extrémités.



ou



- alignement local (connu comme l'algorithme de Smith and Waterman). L'algorithme recherche les deux sous-régions les plus conservées entre les deux séquences. Seulement ces deux régions seront alignées.



# Algorithme de programmation dynamique

Comment ça marche ?

Prenons comme exemple deux séquences X et Y de longueur M et N :

X = AGTCCATC M=8

Y = TCCGC N=5

Matrice de programmation dynamique :

|     |   |     |   |   |   |   |   |   |   |
|-----|---|-----|---|---|---|---|---|---|---|
|     |   | → i |   |   |   |   |   |   |   |
|     |   | A   | G | T | C | C | A | T | C |
| ↓ j |   |     |   |   |   |   |   |   |   |
|     | T |     |   |   |   |   |   |   |   |
|     | C |     |   |   |   |   |   |   |   |
|     | C |     |   |   |   |   |   |   |   |
|     | G |     |   |   |   |   |   |   |   |
|     | C |     |   |   |   |   |   |   |   |

Le score optimal sera calculé récursivement. Le score calculé pour la cellule (i,j) correspondra au meilleur alignement des résidus  $x_1.....x_i$  avec les résidus  $y_1.....y_j$

# Algorithme de programmation dynamique

Comment calcule-t-on le score d'une cellule ?

Il y a seulement trois façons d'aligner  $x_i$  avec  $y_j$  :

- les deux résidus s'alignent (identité ou substitution)
- le résidu  $x_i$  est aligné avec un gap (insertion dans la séquence X)
- le résidu  $y_j$  est aligné avec un gap (insertion dans la séquence Y)

Cela correspond à trois chemins différents pour atteindre la cellule  $(i,j)$

- on atteint la cellule par la diagonale en venant de la cellule  $(i-1,j-1)$
- on atteint la cellule par l'horizontale en venant de la cellule  $(i-1,j)$
- on atteint la cellule par la verticale en venant de la cellule  $(i, j-1)$

L'algorithme doit choisir entre ces trois chemins. S'il utilise un score d'homologie, il choisira le chemin qui maximise la valeur du score  $s(i,j)$ . S'il utilise un score de distance, il choisira le chemin qui minimise la valeur du score  $s(i,j)$ .

# Algorithme de programmation dynamique

Score d'homologie:

$$s(i, j) = \max \begin{cases} s(i-1, j-1) + w(x_i, y_j) \\ s(i-1, j) + \delta \\ s(i, j-1) + \delta \end{cases}$$

Score de distance:

$$s(i, j) = \min \begin{cases} s(i-1, j-1) + w(x_i, y_j) \\ s(i-1, j) + \delta \\ s(i, j-1) + \delta \end{cases}$$

Où :

- $w(x_i, y_j)$  est la valeur dans le système de score correspondant soit à l'identité soit à la substitution (mismatch)
- $\delta$  est le poids de l'insertion/délétion (indel)

# Algorithme de programmation dynamique

Une fois la matrice remplie, le score de l'alignement optimal est le dernier calculé  $s(M,N)$ . Mais on ne connaît pas encore l'alignement proprement dit. Il va être construit par une procédure de « retour en arrière » réursive. En partant de la dernière cellule  $(M,N)$ , on détermine le chemin utilisé pour l'atteindre et on le traduit en terme d'alignement. On continue le processus, une cellule du chemin optimal à la fois, jusqu'à atteindre la première cellule  $(1,1)$ . A ce point, l'alignement optimal est complètement reconstruit.

## Alignement local:

Deux différences majeures:

- l'alignement peut commencer à n'importe quelles positions, pas forcément les premières
- l'alignement peut se terminer à n'importe quelles positions, pas forcément les dernières

L'algorithme va utiliser un score d'homologie et seule l'identité recevra un poids positif (score élémentaire).

Quand la valeur du score d'une cellule devient négatif, il est remplacé par zéro. Il vaut mieux recommencer un nouvel alignement que de le prolonger. Donc une cellule contenant un zéro indique le début d'un alignement.

Quand on reconstruit l'alignement par la procédure de « retour en arrière », au lieu de partir de la dernière cellule, on choisira celle qui a le score le plus élevé.

# Alphabet étendu pour les nucléotides

- Problème de séquençage
- Polymorphisme

L'alphabet étendu permet de modéliser l'incertitude sur une séquence : le nucléotide à une position n'est pas clairement identifié ou peut varier.

| Symbol | Meaning          | Nucleic Acid |
|--------|------------------|--------------|
| -----  |                  |              |
| A      | A                | Adenine      |
| C      | C                | Cytosine     |
| G      | G                | Guanine      |
| T      | T                | Thymine      |
| U      | U                | Uracil       |
| M      | A or C           | purine       |
| R      | A or G           |              |
| W      | A or T           |              |
| S      | C or G           |              |
| Y      | C or T           | pyrimidine   |
| K      | G or T           |              |
| V      | A or C or G      | not T        |
| H      | A or C or T      | not G        |
| D      | A or G or T      | not C        |
| B      | C or G or T      | not A        |
| X      | G or A or T or C | any          |
| N      | G or A or T or C | any          |
| .      | G or A or T or C | any          |

## Problème :

un mismatch entre A et C  
n'a pas le même coût qu'un  
mismatch entre A et M !

# Système de score : matrices de substitution

|          | Score d'homologie | Score de distance |
|----------|-------------------|-------------------|
| identité | +1                | 0                 |
| mismatch | -1                | +1                |
| indel    | -2                | +2                |



|   | A  | T  | C  | G  | -  |
|---|----|----|----|----|----|
| A | +1 | -1 | -1 | -1 | -2 |
| T | -1 | +1 | -1 | -1 | -2 |
| C | -1 | -1 | +1 | -1 | -2 |
| G | -1 | -1 | -1 | +1 | -2 |
| - | -2 | -2 | -2 | -2 |    |

|   | A  | T  | C  | G  | -  |
|---|----|----|----|----|----|
| A | 0  | +1 | +1 | +1 | +2 |
| T | +1 | 0  | +1 | +1 | +2 |
| C | +1 | +1 | 0  | +1 | +2 |
| G | +1 | +1 | +1 | 0  | +2 |
| - | +2 | +2 | +2 | +2 |    |

Les matrices de substitution permettent de spécifier le coût/score de chaque substitution possible (A avec C, A avec T, ...) de manière indépendante

# Matrice pour les nucléotides (alphabet étendu)

## NUC4.4 pour BLAST ou EDNAFULL pour EMBOSS

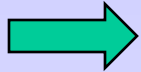
```
# This matrix was created by Todd Lowe    12/10/92
#
# Uses ambiguous nucleotide codes, probabilities rounded to
# nearest integer
#
# Lowest score = -4, Highest score = 5
#
```

|   | A  | T  | G  | C  | S  | W  | R  | Y  | K  | M  | B  | V  | H  | D  | N  |
|---|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| A | 5  | -4 | -4 | -4 | -4 | 1  | 1  | -4 | -4 | 1  | -4 | -1 | -1 | -1 | -2 |
| T | -4 | 5  | -4 | -4 | -4 | 1  | -4 | 1  | 1  | -4 | -1 | -4 | -1 | -1 | -2 |
| G | -4 | -4 | 5  | -4 | 1  | -4 | 1  | -4 | 1  | -4 | -  | -  | -  | -  | -  |
| C | -4 | -4 | -4 | 5  | 1  | -4 | -4 | 1  | -4 | 1  | -  | -  | -  | -  | -  |
| S | -4 | -4 | 1  | 1  | -1 | -4 | -2 | -2 | -2 | -2 | -  | -  | -  | -  | -  |
| W | 1  | 1  | -4 | -4 | -4 | -1 | -2 | -2 | -2 | -2 | -  | -  | -  | -  | -  |
| R | 1  | -4 | 1  | -4 | -2 | -2 | -1 | -4 | -2 | -2 | -  | -  | -  | -  | -  |
| Y | -4 | 1  | -4 | 1  | -2 | -2 | -4 | -1 | -2 | -2 | -  | -  | -  | -  | -  |
| K | -4 | 1  | 1  | -4 | -2 | -2 | -2 | -2 | -1 | -4 | -  | -  | -  | -  | -  |
| M | 1  | -4 | -4 | 1  | -2 | -2 | -2 | -2 | -4 | -1 | -  | -  | -  | -  | -  |
| B | -4 | -1 | -1 | -1 | -1 | -3 | -3 | -1 | -1 | -3 | -  | -  | -  | -  | -  |
| V | -1 | -4 | -1 | -1 | -1 | -3 | -1 | -3 | -3 | -1 | -  | -  | -  | -  | -  |
| H | -1 | -1 | -4 | -1 | -3 | -1 | -3 | -1 | -3 | -1 | -  | -  | -  | -  | -  |
| D | -1 | -1 | -1 | -4 | -3 | -1 | -1 | -3 | -1 | -3 | -  | -  | -  | -  | -  |
| N | -2 | -2 | -2 | -2 | -1 | -1 | -1 | -1 | -1 | -1 | -  | -  | -  | -  | -  |

| Symbol | Meaning          | Nucleic Acid |
|--------|------------------|--------------|
| A      | A                | Adenine      |
| C      | C                | Cytosine     |
| G      | G                | Guanine      |
| T      | T                | Thymine      |
| U      | U                | Uracil       |
| M      | A or C           |              |
| R      | A or G           | purine       |
| W      | A or T           |              |
| S      | C or G           |              |
| Y      | C or T           | pyrimidine   |
| K      | G or T           |              |
| V      | A or C or G      | not T        |
| H      | A or C or T      | not G        |
| D      | A or G or T      | not C        |
| B      | C or G or T      | not A        |
| X      | G or A or T or C |              |
| N      | G or A or T or C |              |
| .      | G or A or T or C |              |

# 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

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

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

| le code génétique         |   |                    |     |     |     |                      |      |     |      |   |
|---------------------------|---|--------------------|-----|-----|-----|----------------------|------|-----|------|---|
|                           |   | Deuxième lettre    |     |     |     |                      |      |     |      |  |
|                           |   | U                  |     | C   |     | A                    |      | G   |      |   |
| Première lettre (côté 5') | 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 |      |     |      |   |

# 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)**

## **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 *Phe* 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,  $f(\text{Phe} \rightarrow \text{X})$ )
- Calcul de la matrice de probabilités de mutation à partir des 2 étapes précédentes
  - Pour chaque a.a., ex:  $\text{Phe} \rightarrow \text{Ala} = \text{mutabilité}(\text{Phe}) * \text{cumul}(\text{Phe} \rightarrow \text{Ala}) / \text{nb}(\text{Phe})$
- Calcul de la matrice Lods (Log odd ratios) : PAM1



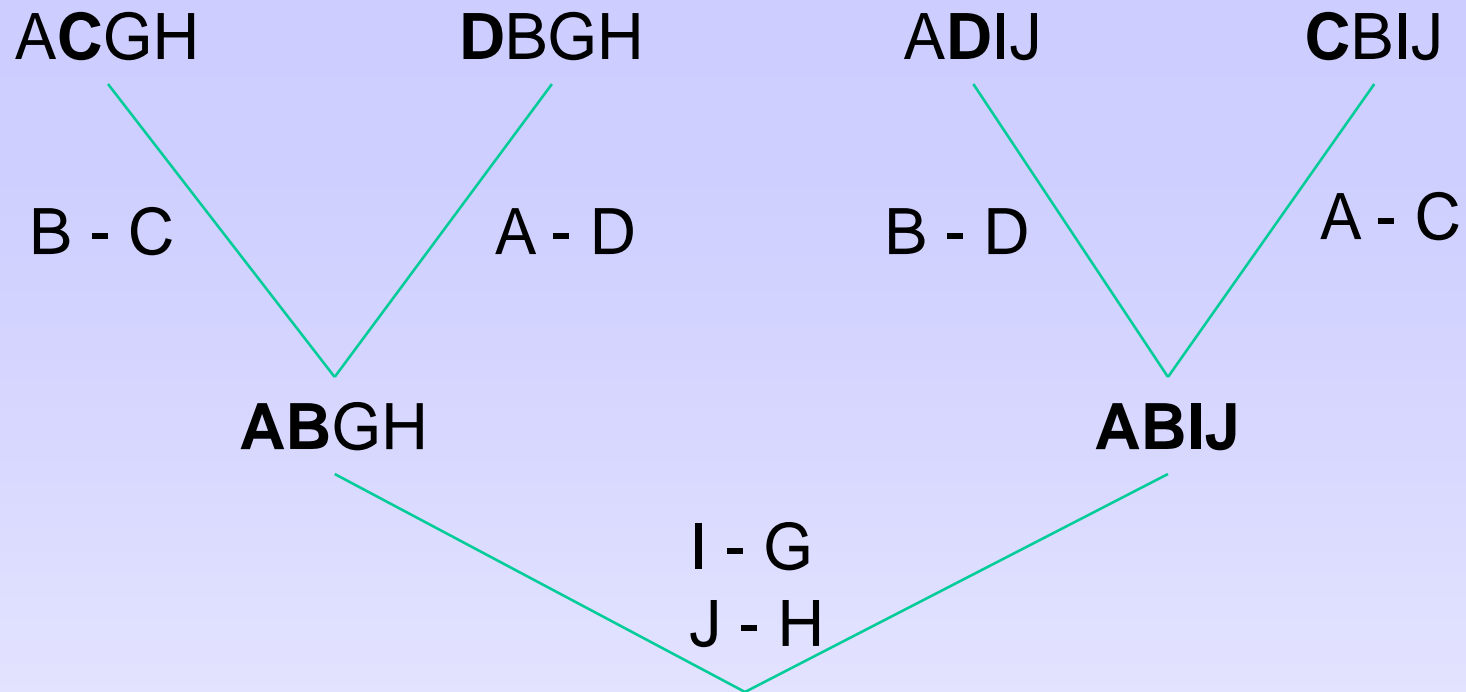
# Matrices PAM

**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 *Phe* 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,  $f(\text{Phe} \rightarrow \text{X})$ )
- Calcul de la matrice de probabilités de mutation à partir des 2 étapes précédentes
  - Pour chaque a.a., ex:  $\text{Phe} \rightarrow \text{Ala} = \text{mutabilité}(\text{Phe}) * \text{cumul}(\text{Phe} \rightarrow \text{Ala}) / \text{nb}(\text{Phe})$
- Calcul de la matrice Lods (Log odd ratios) : PAM1

## Arbre phylogénétique



## Matrice des mutations acceptées

|   | A | B | C | D | G | H | I | J |
|---|---|---|---|---|---|---|---|---|
| A |   |   | 1 | 1 |   |   |   |   |
| B |   |   | 1 | 1 |   |   |   |   |
| C | 1 | 1 |   |   |   |   |   |   |
| D | 1 | 1 |   |   |   |   |   |   |
| G |   |   |   |   |   |   | 1 |   |
| H |   |   |   |   |   |   |   | 1 |
| I |   |   |   |   | 1 |   |   |   |
| J |   |   |   |   |   | 1 |   |   |

# 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 *Phe* 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,  $f(\text{Phe} \rightarrow \text{X})$ )
- Calcul de la matrice de probabilités de mutation à partir des 2 étapes précédentes
  - Pour chaque a.a., ex:  $\text{Phe} \rightarrow \text{Ala} = \text{mutabilité}(\text{Phe}) * \text{cumul}(\text{Phe} \rightarrow \text{Ala}) / \text{nb}(\text{Phe})$
- Calcul de la matrice Lods (Log odd ratios) : PAM1

# 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 *Phe* 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,  $f(\text{Phe} \rightarrow \text{X})$ )
- Calcul de la matrice de probabilités de mutation à partir des 2 étapes précédentes
  - Pour chaque a.a., ex:  $\text{Phe} \rightarrow \text{Ala} = \text{mutabilité}(\text{Phe}) * \text{cumul}(\text{Phe} \rightarrow \text{Ala}) / \text{nb}(\text{Phe})$
- Calcul de la matrice Lods (Log odd ratios) : PAM1

Séquences alignées

A D A

A D B



| Acides aminés        | A   | B | D |
|----------------------|-----|---|---|
| Changements observés | 1   | 1 | 0 |
| Occurrences          | 3   | 1 | 2 |
| Mutabilité           | .33 | 1 | 0 |

Mutabilité (Dayhoff, 1978)

Ser 149

Met 122

Asn 111

Ile 110

Glu 102

Ala 100

Gln 98

Asp 90

Thr 90

Gap 84

Val 80

Lys 57

Pro 56

His 50

Gly 48

Phe 45

Arg 44

Leu 38

Tyr 34

Cys 27

Trp 22



Positionnée à 100 arbitrairement

# 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 *Phe* 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,  $f(\text{Phe} \rightarrow \text{X})$ )
- Calcul de la matrice de probabilités de mutation à partir des 2 étapes précédentes

- Pour chaque a.a., ex:  $\text{Phe} \rightarrow \text{Ala} = \text{mutabilité}(\text{Phe}) * \text{cumul}(\text{Phe} \rightarrow \text{Ala}) / \text{nb}(\text{Phe})$

$m_j$ : mutabilité de l'a.a.  $j$

$A_{ij}$ : nombre de fois que l'a.a.  $j$  a été remplacé par l'a.a.  $i$

$\lambda$ : paramètre d'ajustement pour avoir 1 mutation acceptée pour 100 résidus

$$M_{i,j} = \lambda \frac{m_j A_{ij}}{\sum_i A_{ij}}$$

- Calcul de la matrice Lods (Log odd ratios) :  $PAM 1_{i,j} = \log \frac{M_{ij}}{f_i}$

# Matrices PAM

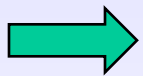
## Calcul de la matrice Lods (Log odd ratios) :

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

$$PAM2 = PAM1 \times PAM1$$

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 | 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  |
|   | C  | S  | T  | P  | A  | G  | N  | D  | E  | Q  | H  | R  | K  | M  | I  | L  | V  | F | Y  |
|   |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |   | 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)

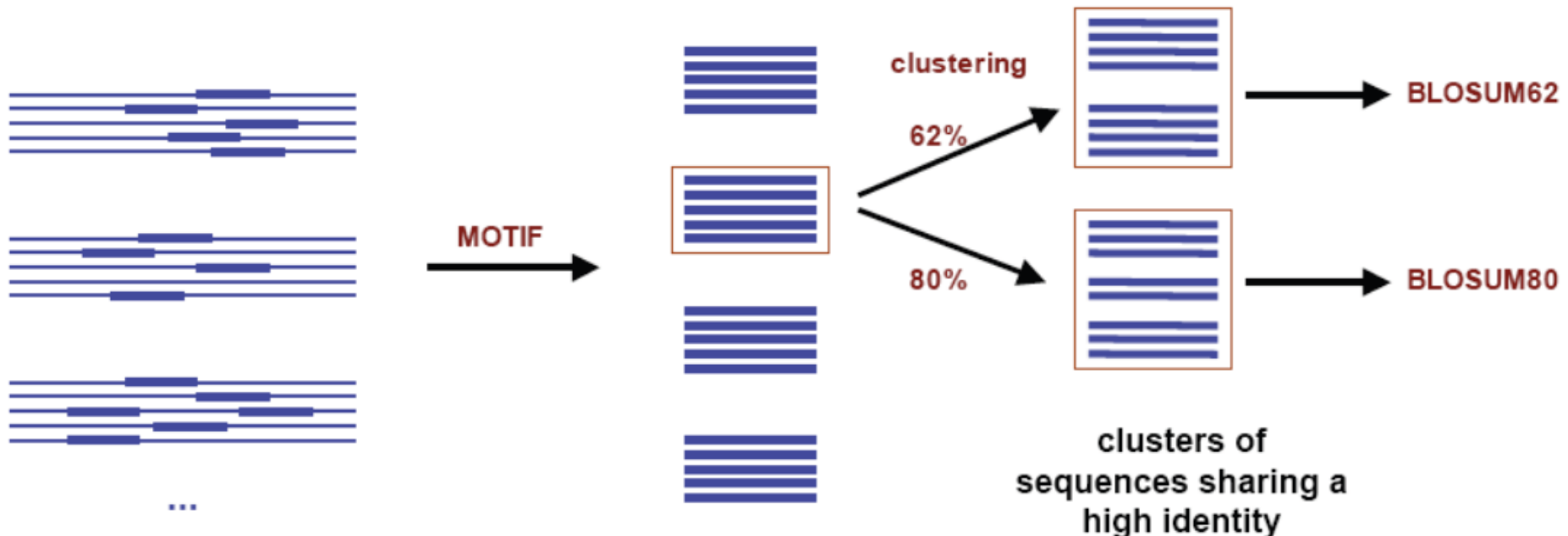
BLOSUM : BLOcks SUBstitution Matrix

Principe :

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

Avantages par rapport aux matrices PAM :

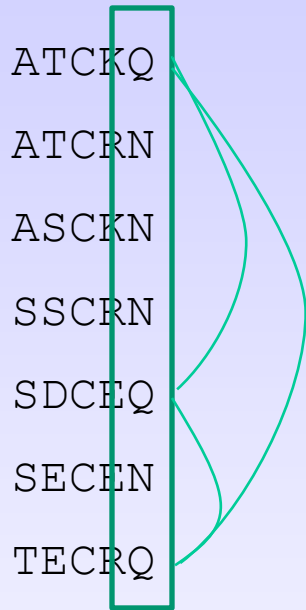
- contrairement aux matrices PAM, les matrices BLOSUM pour différentes distances évolutives sont obtenues directement avec des séquences plus ou moins divergentes
- l'utilisation de blocs plutôt que de séquences complètes : modélise les contraintes uniquement sur les régions conservées
- obtenues à partir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



# Matrices BLOSUM

## Calcul de la fréquence de chaque paire

Identité à 62% -> chaque séquence a le même poids



$$f_{Q,N} = 12/21$$

$$f_{N,N} = 6/21$$

$$f_{Q,Q} = 3/21$$

[illegible]



# Matrices BLOSUM

Calcul de la fréquence de chaque paire

Identité à 50%

$1/4$  ATCKQ  
 ATCRN  
 ASCKN  
 SSCRN  
  
 $1/2$  SDCEQ  
 SECEN  
  
 $1$  TECRQ

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

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

$$f_{N,N} = 3/4 * 1/2 = 3/8$$

$$f_{Q,Q} = 1/4 * 1/2 + 1/4 * 1 + 1/2 * 1 = 7/8$$

# Matrices BLOSUM

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  $j$

Calcul de la fréquence théorique

Ex: 12 QN, 6 NN et 3 QQ

$$P(Q) = 12/2 + 3 = 9$$

$$P(N) = 12/2 + 6 = 12$$

$$P(QN) = 2 * p(Q) * p(N) = 2 * 9/21 * 12/21$$

De manière générale :

$$p_{ii} = p_i^2$$

$$p_{ij} = 2 * p_i * p_j$$

$$p_i = q_{ii} + \frac{1}{2} \sum q_{ij}$$

ATCKQ

ATCRN

ASCKN

SSCRN

SDCEQ

SECEN

TECRQ

# Alignement de deux séquences protéiques

## Matrices de substitution

### La matrice BLOSUM62

|   |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |   |   |    |
|---|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|---|---|----|
| 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 |
|   | C  | S  | T  | P  | A  | G  | N  | D  | E  | Q  | H  | R  | K  | M  | I  | L  | V  | F | Y | W  |

# Alignement de deux séquences protéiques

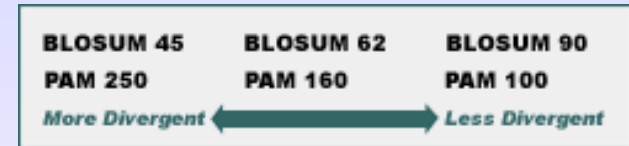
## Matrices de substitution

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.

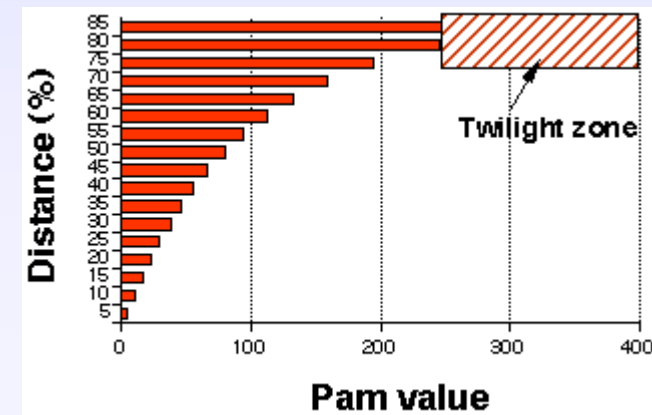


Source figure : ebi.ac.uk

| longueur | 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 BLOSUM30 donnera de meilleurs résultats pour des séquences éloignées dans l'évolution. Elle trouvera de plus longs alignements locaux de faible conservation.

# 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-KTLPSSSTSTLLP--RSTF--PFP-HHPHK
      . . . . . : : : : : : : : : : : : : : : : : :
ILVB_A MAAATTTTTTSSSISFSTKP-SPSSSKSP-I-PISR--FSLPFSLN-PNK
      10      20      30      40

      50      60      70
ILV1_T TTPPPLHLTHTHIHSQRRR-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--ETFVSRFAPDEPRKGSDDLVEALEREGV-TDVFAYPGGASMEIH
      : : : : : : : : : : : : : : : : : :
ILVB_A -KPTKPETFISRFPADQPRKGADILVEALERQGVET-VFAYPGGASMEIH
      80      90     100     110     120

      130     140     150     160     170
ILV1_T QALTRSS-IIRNVLPHEQGGVFAAEGYARATG-FPGVCIATSGPGATNL
      : : : : : : : : : : : : : : : : : :
ILVB_A QALTRSSSI-RNVLPHEQGGVFAAEGYARSSGK-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-RSGRPGPVLVDVPKDIQQQLAIPNWEQAMRL
      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 MAAAPSP--PSSSAFSKTLPSSSTSTLLPSTFFPFPHPHKTTPPPL
      . . . . . : : : : : : : : : : : : : : : : :
ILVB_A MAAATTTTTTSSSISFSTKPSPPSSSKSPISRFSLPFSLNPNKSSS--
      10      20      30      40

      50      60      70      80
ILV1_T HLTHTHIHSQRRR-----FTISNVISTNQKVSQTE-----KTETF
      : : : : : : : : : : : : : : : : : :
ILVB_A -----SSRRRGIKSSSPSSISAVLNTTTNVTTTPSPTKPTKPTF
      50      60      70      80

      90     100     110     120     130
ILV1_T VSRFAPDEPRKGSDDLVEALEREGVTDVFAYPGGASMEIHQALTRSSIIR
      : : : : : : : : : : : : : : : : : :
ILVB_A ISRFAPDQPRKGADILVEALERQGVETVFAYPGGASMEIHQALTRSSSIIR
      90     100     110     120     130

      140     150     160     170     180
ILV1_T NVLPHEQGGVFAAEGYARATGFPVCIATSGPGATNLVSGLADALLDSV
      : : : : : : : : : : : : : : : : : :
ILVB_A NVLPHEQGGVFAAEGYARSSGKPGICIATSGPGATNLVSGLADALLDSV
      140     150     160     170     180

      190     200     210     220     230
ILV1_T PIVAITGQVPRRMIGTDAFQETPIVEVTRSITKHNYLVMDVEDIPRVVRE
      : : : : : : : : : : : : : : : : : :
ILVB_A PIVAITGQVPRRMIGTDAFQETPIVEVTRSITKHNYLVMDVEDIPRIIEE
      190     200     210     220     230

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

# 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 MAAATTTTTTSSSISFSTKPSPPSSSKSPLPISRFLPFSLNPKNKSSSSSR
      10      20      30      40      50

      20      30      40      50
PDC1_M CNGVGALPVA NSHAIATPAAAAATLAPAGAT----LGRH-----
      :. . . . : : : : : :
ILVB_A RRGIKSSSPSSISAVLNTTNTVTTTSPPTKPTKPKETFISRFAPDQPRKGA
      60      70      80      90     100

      60      70      80      90     100
PDC1_M --LARRLVQIGASDVFAVPGDFNLTLDDYLIAEPLGLTVGCCNELNAGYA
      : : : : : : : : : : : :
ILVB_A DILVEALERQGVETVFAYPGGASMEIHQALTRSSSIRNVLP RHEQGGVFA
      110     120     130     140     150

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

      160     170     180     190
PDC1_M YGTNRILHHTIGLPDFSQELRCFQTITCYQAIINNLLDAHEQIDTA--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 MAAATTTTTTSSSISFSTKPSPPSSSKSPLPISRFLPFSLNPKNKSSSSSR
      10      20      30      40      50

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

      60      70      80      90     100
PDC1_M ---LVQI---GASDVFAVPGDFNLTLDDYLIAEPLGLTVGCCNELNAGY
      :. . : : : : : : : : :
ILVB_A ADILVEALERQGVETVFAYPGGASMEIHQALTRSSSIRNVLP RHEQGGVF
      100     110     120     130     140

      110     120     130     140
PDC1_M AADGYARSRG-VGACAVTFTVGGLSVLNAIAGAYSENLPVVCIVGGPNSN
      : : : : : : : : : : :
ILVB_A AAEGYARSSGKPGICIA TSGPGATNLVSGLADALLDSVPLVAI-----
      150     160     170     180     190

      150     160     170     180
PDC1_M DYGTNRILHHTIGLPDFSQELRCFQT---ITCYQAI--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 MAAATTTTTTSSSISFSTKPSPPSSSKSPLPISRFLPFSLNPNKSSSSSR
      10          20          30          40          50

      20          30          40          50
PDC1_M CNGVGALPVANSHAIATPAAAAATLAPAGAT----LGRH-----
      : . . . . . : : : : :
ILVB_A RRGIKSSSPSSISAVLNTTTNVTTPSPPTKPTKPTFISRFAPDQPRKGA
      60          70          80          90         100

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

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

      160         170         180         190
PDC1_M YGTNRILHHTIGLPDFSQELRCFQTITCYQAIINNLLDAHEQIDTA--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 MAAATTTTTTSSSISFSTKPSPPSSSKSPLPISRFLPFSLNPNKSSSSSR
      10          20          30          40          50

      30          40          50
PDC1_M -----SHAIATPAAAAATLAPAGAT----LGRH-----
      : . . . . . : : : . . . . .
ILVB_A RRGIKSSSPSSISAVLNTTTNVTTPSPPTKPTKPTFISRFAPDQPRKGA
      60          70          80          90         100

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

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

      160         170         180         190
PDC1_M YGTNRILHHTIGLPDFSQE--LRCFQTITCYQAIINNLLDAHEQIDTA--
      : . . : : : : : : : :
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
```

# 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 MAAAAPSPSSSAFSKTLPSSSSTSSTLLPRSTFFPHHPHKTTPPPLHLT
      10          20          30          40          50

frag_n -----
ILV1_T HTHIHHSQRRRFTISNVISTNQKVSQTEKTETFVSRFAPDEPRKGSVDL
      60          70          80          90         100

frag_n -----
ILV1_T VEALEREGVTDVFAYPGGASMEIHQALTRSSIIRNVLP RHEQGGVFAAEG
      110         120         130         140         150

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

frag_n -----
ILV1_T DAFQETPIVEVTRSITKHNYLVMDVEDIPRVVREAFFLARSGRPGPILID
      210         220         230         240         250

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

```

      20          30
frag_n -----VGATLGYAGAV-----S
      : : : :
ILV1_T LYVGGGCSQSSDLRRFVELTGIPVASTLMGLGAFPTGDELSLSMLGMHG
      310         320         330         340         350

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

frag_n -----YSSWQEN-----
      : : : :
ILV1_T NKQPHVSICADIKLALQGLNSILESKEGKCLKLDFSAWRQELTEQVKVHPL
      410         420         430         440         450

frag_n -----DQCP--Y-----RT
      : : : :
ILV1_T NEKTFGDAIPPPQYAIQVLDEL TNGNAIISTGVGQHQMWA AQYYKYRKPRQ
      460         470         480         490         500

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

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

      90
frag_n F-----ERME-----
      :          . :
ILV1_T FAEACGVPAARVTHRDDLRAAIQKMLDTPGPYLLDVIVPHQEHVLP MIP S
      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
```