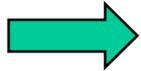


Alignement de deux séquences protéiques

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, il faut prendre en compte en plus de l'identité et de la différence, la similarité qui peut exister entre deux acides aminés.

Comment quantifier la similarité 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 | | |
| 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)

M. Dayhoff *et al* (1978) A model of evolutionary change in proteins. In Atlas of protein sequence and structure Vol 5, No suppl 3, p.345-351

22 A Model of Evolutionary Change in Proteins

M.O. Dayhoff, R.M. Schwartz, and B.C. Orcutt

In the eight years since we last examined the amino acid exchanges seen in closely related proteins,¹ the information has doubled in quantity and comes from a much wider variety of protein types. The matrices derived from these data that describe the amino acid replacement probabilities between two sequences at various evolutionary distances are more accurate and the scoring matrix that is derived is more sensitive in detecting distant relationships than the one that we previously derived.^{2,3} The method used in this chapter is essentially the same as that described in the *Atlas*, Volume 3⁴ and Volume 5.¹

Accepted Point Mutations

The matrix of accepted point mutations calculated from this tree is shown in Figure 79. We have assumed that the likelihood of amino acid X replacing Y is the same as that of Y replacing X, and hence 1 is entered in box YX as well as in box XY. This assumption is reasonable, because this likelihood should depend on the product of the frequencies of occurrence of the two amino acids and on their chemical and physical similarity. As a consequence of this assumption, no change in amino acid frequencies over evolutionary distance will be detected.

By comparing observed sequences with inferred ancestral sequences, rather than with each other, a sharper

Elle rend compte de deux processus :

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

Etant donné ces mutations ponctuelles acceptées, on pourrait calculer la probabilité qu'un acide aminé dans une position reste le même ou la probabilité qu'il change vers un autre acide aminé.

Matrices PAM

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

Le modèle utilisé est dérivé d'un modèle général pour l'évolution des protéines.

Deux hypothèses :

- Les événements de mutation sont indépendants du contexte, c'est-à-dire de la position et de la nature des acides aminés adjacents. Les positions de la protéine sont considérées comme indépendantes.
- La probabilité de mutation d'un acide aminé est indépendante de ce qui s'est produit à cette position dans le passé. Par exemple, la probabilité qu'un acide aminé valine soit remplacé par l'acide aminé leucine est la même pour tous les acides aminés valine rencontrés dans la protéine et ceci indépendamment des acides aminés qui occupaient cette position dans le passé (donc avant l'acide aminé valine)

Matrices PAM

Construction :

- 71 familles de protéines (environ 1300 séquences)
- Une famille de protéines est constituée de protéines homologues
- Choix des séquences : très proches minimum 85% d'identité entre chaque paire de séquences de manière à éviter la présence de substitutions multiples).

Comment à partir d'un ensemble de séquences protéiques homologues (d'une famille) on détermine les mutations acceptées ?

- alignements multiples (global)
- reconstruction des arbres phylogénétiques et inférences des séquences ancêtres

Exemple simple : alignement multiple des 4 séquences suivantes

ACGH

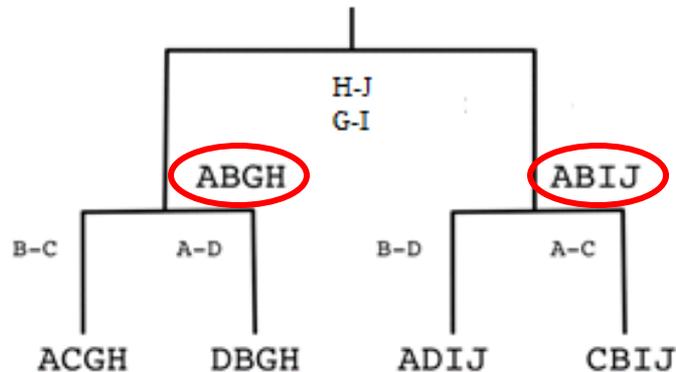
DBGH

ADIJ

CBIJ

Matrices PAM

Cet alignement multiple est utilisé pour reconstruire un arbre phylogénétique ainsi que les séquences ancêtres



Alignement de chaque séquence avec sa séquence ancêtre :

-> détermination des mutations acceptées

ACGH
ABGH -> Mutation B-C

ADIJ
ABIJ -> Mutation B-D

DBGH
ABGH -> Mutation A-D

CBIJ
ABIJ -> Mutation A-C

Les deux séquences ancêtres ABGH et ABIJ (entourées) avec l'ancêtre commun :
H-J et G-I

Matrices PAM

Extrait des alignements multiples du jeu de données de Dayhoff

| | | 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... | | | | | | | | | | |
| 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 |
| ... | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |

reconstruction des arbres phylogénétiques et inférences des séquences ancêtres
 -> 1 572 mutations

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) issue du comptage des n_{ij}

| | 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 | 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 | |

Nombre de fois où C (cys) est muté : 280
 Nombre de fois où V (val) est muté : 2003

Pourquoi cette différence ?
 V plus fréquent que C dans l'échantillon ?
 V mute plus fréquemment que C ?

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 (pourcentage d'acide aminé i muté et non muté) :

$$m_i = \frac{\text{nombre de changements de } i}{\text{nombre d'occurrences de } i} = \frac{\sum_{i \neq j} n_{ij}}{\sum_j n_{ij}}$$

➤ Calcul de la probabilité de mutation de chaque paire d'acides aminés

Le calcul de la mutabilité nous indique si un acide aminé i est plus souvent (ou moins souvent) muté qu'un autre acide aminé k . Par contre, ceci ne nous indique pas quel est le pourcentage, par exemple, de i muté en j . Ce pourcentage correspond à la probabilité de mutation de la paire d'acides aminés ij . Cette valeur est donnée par :

$$p_{ij} = m_i \frac{n_{ij}}{\sum_{i \neq j} n_{ij}} = \text{mutabilité de } i \frac{\text{nombre d'acides aminés } i \text{ muté en } j}{\text{nombre d'acides aminés } i \text{ muté}}$$

Exemple : on à 100 acides aminés i et 10 sont mutés $\rightarrow m_i = 0,1$

5 i sont mutés en A $\rightarrow p_{iA} = 0,1 * 5/10 = 0,05$ (5%)

3 i sont mutés en V $\rightarrow p_{iV} = 0,1 * 3/10 = 0,03$ (3%)

2 i sont mutés en S $\rightarrow p_{iS} = 0,1 * 2/10 = 0,02$ (2%)

Principes généraux de la construction de la PAM

Pour le moment, dans les calculs le temps n'a pas été pris en compte. Comment ramener les valeurs calculées à un même intervalle de temps ?

➤ Prise en compte du temps : calcul de la matrice de probabilité 1-PAM

La matrice « de base » construite par Dayhoff correspond à la 1-PAM dans laquelle on a 1 mutation acceptée pour 100 acides aminés. 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%. Ceci définit une unité de changement évolutif. Les termes de la 1-PAM sont donnés par :

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

Et pour les termes diagonaux

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

q_{ij} donne la probabilité de mutation de la paire d'acides aminés ij pour un intervalle d'évolution dans lequel 1% des substitutions sont acceptées

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

Matrice 1-PAM

probabilités de mutation 1-PAM (une mutation pour 100 aa) (ajustement obtenu grâce à ρ)

Les éléments sont multipliés par 10000

| | Ala | Arg | Asn | Asp | Cys | Gln | Glu | Gly | His | Ile | Leu | Lys | Met | Phe | Pro | Ser | Thr | Trp | Tyr | Val |
|-------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| Ala A | 9867 | 2 | 9 | 10 | 3 | 8 | 17 | 21 | 2 | 6 | 4 | 2 | 6 | 2 | 22 | 35 | 32 | 0 | 2 | 18 |
| Arg R | 1 | 9913 | 1 | 0 | 1 | 10 | 0 | 0 | 10 | 3 | 1 | 19 | 4 | 1 | 4 | 6 | 1 | 8 | 0 | 1 |
| Asn N | 4 | 1 | 9822 | 36 | 0 | 4 | 6 | 6 | 21 | 3 | 1 | 13 | 0 | 1 | 2 | 20 | 9 | 1 | 4 | 1 |
| Asp D | 6 | 0 | 42 | 9859 | 0 | 6 | 53 | 6 | 4 | 1 | 0 | 3 | 0 | 0 | 1 | 5 | 3 | 0 | 0 | 1 |
| Cys C | 1 | 1 | 0 | 0 | 9973 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 5 | 1 | 0 | 3 | 2 |
| Gln Q | 3 | 9 | 4 | 5 | 0 | 9876 | 27 | 1 | 23 | 1 | 3 | 6 | 4 | 0 | 6 | 2 | 2 | 0 | 0 | 1 |
| Glu E | 10 | 0 | 7 | 56 | 0 | 35 | 9865 | 4 | 2 | 3 | 1 | 4 | 1 | 0 | 3 | 4 | 2 | 0 | 1 | 2 |
| Gly G | 21 | 1 | 12 | 11 | 1 | 3 | 7 | 9935 | 1 | 0 | 1 | 2 | 1 | 1 | 3 | 21 | 3 | 0 | 0 | 5 |
| His H | 1 | 8 | 18 | 3 | 1 | 20 | 1 | 0 | 9912 | 0 | 1 | 1 | 0 | 2 | 3 | 1 | 1 | 1 | 4 | 1 |
| Ile I | 2 | 2 | 3 | 1 | 2 | 1 | 2 | 0 | 0 | 9872 | 9 | 2 | 12 | 7 | 0 | 1 | 7 | 0 | 1 | 33 |
| Leu L | 3 | 1 | 3 | 0 | 0 | 6 | 1 | 1 | 4 | 22 | 9947 | 2 | 45 | 13 | 3 | 1 | 3 | 4 | 2 | 15 |
| Lys K | 2 | 37 | 25 | 6 | 0 | 12 | 7 | 2 | 2 | 4 | 1 | 9926 | 20 | 0 | 3 | 8 | 11 | 0 | 1 | 1 |
| Met M | 1 | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 5 | 8 | 4 | 9874 | 1 | 0 | 1 | 2 | 0 | 0 | 4 |
| Phe F | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 2 | 8 | 6 | 0 | 4 | 9946 | 0 | 2 | 1 | 3 | 28 | 0 |
| Pro P | 13 | 5 | 2 | 1 | 1 | 8 | 3 | 2 | 5 | 1 | 2 | 2 | 1 | 1 | 9926 | 12 | 4 | 0 | 0 | 2 |
| Ser S | 28 | 11 | 34 | 7 | 11 | 4 | 6 | 16 | 2 | 2 | 1 | 7 | 4 | 3 | 17 | 9840 | 38 | 5 | 2 | 2 |
| Thr T | 22 | 2 | 13 | 4 | 1 | 3 | 2 | 2 | 1 | 11 | 2 | 8 | 6 | 1 | 5 | 32 | 9871 | 0 | 2 | 9 |
| Trp W | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 9976 | 1 | 0 |
| Tyr Y | 1 | 0 | 3 | 0 | 3 | 0 | 1 | 0 | 4 | 1 | 1 | 0 | 0 | 21 | 0 | 1 | 1 | 2 | 9945 | 1 |
| Val V | 13 | 2 | 1 | 1 | 3 | 2 | 2 | 3 | 3 | 57 | 11 | 1 | 17 | 1 | 3 | 2 | 10 | 0 | 2 | 9901 |

Matrices PAM

Calcul de la Log odd matrice : la matrice PAM1

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

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

Ceci va permettre de comparer deux modèles. Les fréquences observées correspondent à l'hypothèse d'un modèle de substitution avec contrainte et les fréquences théoriques à un modèle de substitution aléatoire (hypothèse nulle).

$w_{i,j} > 0$: la probabilité de substitution observée de i vers j est plus élevée qu'attendue

$w_{i,j} < 0$: la probabilité de substitution observée de i vers j est moins élevée qu'attendue

$w_{i,j} = 0$: la probabilité de substitution observée de i vers j n'est pas différente de l'attendue (modèle aléatoire)

Matrices PAM

Comment estimer les probabilités de mutation des paires d'acides aminés pour des distances évolutives plus grande ?

Comme on a fait l'hypothèse que la probabilité de mutation d'un acide aminé est indépendante de ce qui s'est produit à cette position dans le passé, on va pouvoir obtenir les probabilités de mutation pour des intervalles d'évolution plus grands par la multiplication de la PAM1 avec elle-même. Une PAM k sera obtenue en multipliant la PAM1 k fois par elle-même (k mutations acceptées pour 100 sites)

$$\text{PAM2} = \text{PAM1} \times \text{PAM1} = \text{PAM1}^2$$

intervalle d'évolution : 2 mutations acceptées pour chaque 100 résidus

$$\text{PAM40} = \text{PAM1}^{40}$$

intervalle d'évolution : 40 mutations acceptées pour chaque 100 résidus

$$\text{PAM120} = \text{PAM1}^{120}$$

intervalle d'évolution : 120 mutations acceptées pour chaque 100 résidus

$$\text{PAM250} = \text{PAM1}^{250}$$

intervalle d'évolution : 250 mutations acceptées pour chaque 100 résidus

divergence



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

S. Henikoff et J.G. Henikoff (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 89:10915-10919

Proc. Natl. Acad. Sci. USA
Vol. 89, pp. 10915-10919, November 1992
Biochemistry

Amino acid substitution matrices from protein blocks

(amino acid sequence/alignment algorithms/data base searching)

STEVEN HENIKOFF* AND JORJA G. HENIKOFF

Howard Hughes Medical Institute, Basic Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98104

Communicated by Walter Gilbert, August 28, 1992 (received for review July 13, 1992)

ABSTRACT Methods for alignment of protein sequences typically measure similarity by using a substitution matrix with scores for all possible exchanges of one amino acid with another. The most widely used matrices are based on the Dayhoff model of evolutionary rates. Using a different approach, we have derived substitution matrices from about 2000 blocks of aligned sequence segments characterizing more than 500 groups of related proteins. This led to marked improvements in alignments and in searches using queries from each of the groups.

Among the most useful computer-based tools in modern biology are those that involve sequence alignments of proteins, since these alignments often provide important insights into gene and protein function. There are several different

new sequence and every other sequence in the block. For example, if the residue of the new sequence that aligns with the first column of the first block is A and the column has 9 A residues and 1 S residue, then there are 9 AA matches and 1 AS mismatch. This procedure is repeated for all columns of all blocks with the summed results stored in a table. The new sequence is added to the group. For another new sequence, the same procedure is followed, summing these numbers with those already in the table. Notice that successive addition of each sequence to the group leads to a table consisting of counts of all possible amino acid pairs in a column. For example, in the column consisting of 9 A residues and 1 S residue, there are $8 + 7 + \dots + 1 = 36$ possible AA pairs, 9 AS or SA pairs, and no SS pairs. Counts of all possible pairs in each column of each block in the data base are summed.

Matrices BLOSUM (Henikoff et Henikoff, 1992)

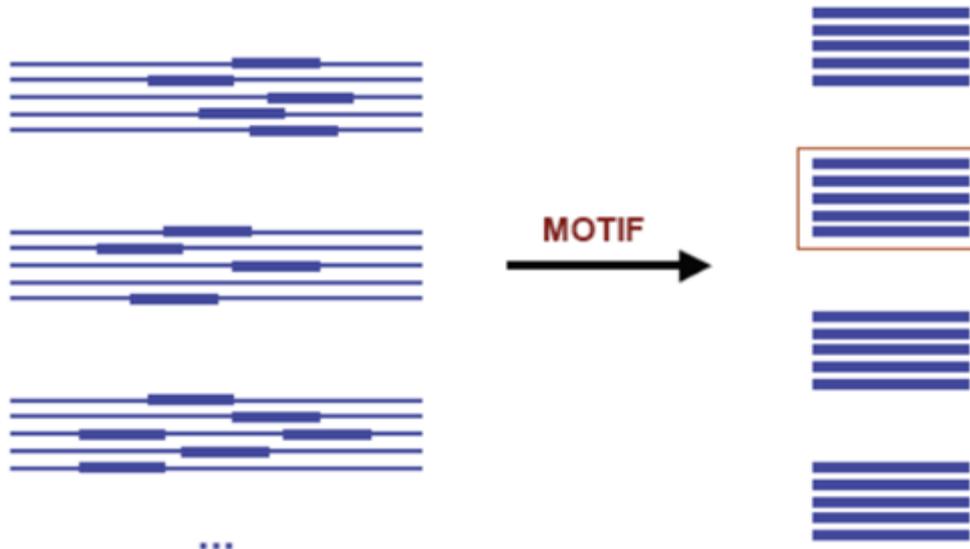
BLOSUM : BLOcks SUBstituion Matrix

Principe :

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

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)



504 groupes de protéines apparentées

2205 blocs (domaines alignement sans gap)

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

Calcul de la fréquence théorique de chaque paire i,j

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 :

$$p_N = ((6+(12/2))/21) = 12/21$$

De manière générale :

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

ATCKQ
ATCRN
ASCKN
SSCRN
RDCEQ
RECEQ
TECRQ

(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 :

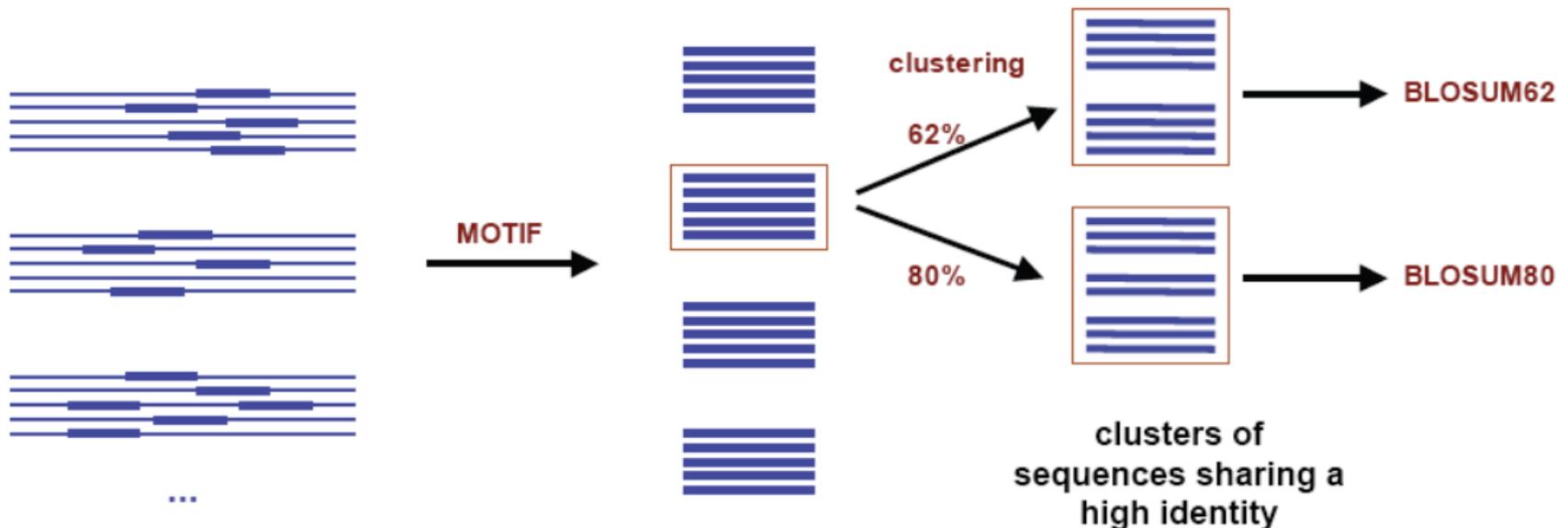
$$q_{QN} = 12/21 = 0,571$$

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 dû 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 pondérée par un poids basé sur leur nombre.

Par exemple, ci-dessous, les séquences dont l'identité est $\geq 62\%$ et les séquences dont l'identité est $\geq 80\%$.

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 \geq au numéro.



Matrices BLOSUM

Exemple : construction BLOSUM 50

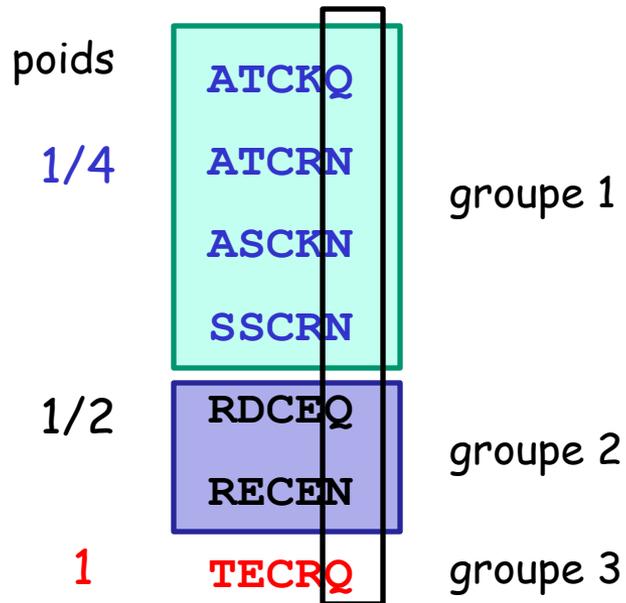
On regroupe les séquences dont l'identité est ≥ 50

| |
|-------|
| 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 :

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

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

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

Matrices BLOSUM

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 |

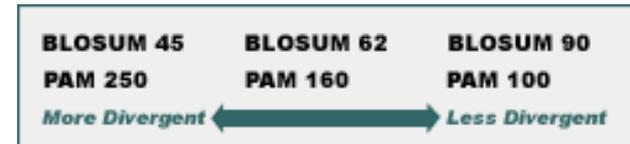
Choix des Matrices de substitution

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.

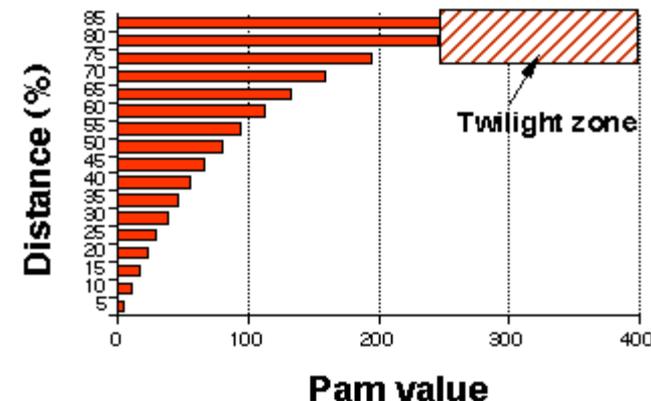


Source figure : ebi.ac.uk

| 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 |



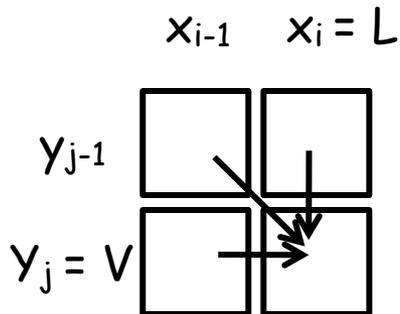
Source figure : Infobiogen.fr

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

Alignements de deux séquences : même principe que pour les séquences d'acides nucléiques (programmation dynamique).



la cellule (i,j) atteinte par la diagonale en venant de la cellule $(i-1,j-1)$

$$s(i, j) = s(i-1, j-1) + w(x_i, y_j)$$

Dans l'exemple $x_i=L$ et $y_j = V$, $w(L,V)$ correspondra à la valeur présente dans la matrice de substitution utilisée

Déplacement horizontal ou vertical on ajoute la pondération de l'indel

Alignement de deux séquences protéiques

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 accepté au cours de l'évolution (acides 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 similarité correspond au pourcentage d'acides aminés identiques et similaires. Dans certains programmes, le pourcentage de similarité est appelé Positives Percentage (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 MAAATTTTTTSSSISFSTKPSFSSSKSPLPISRFLPFSLNPKNKSSSSSR
      10      20      30      40      50
      20      30      40      50
PDC1_M CNGVGALPVANSHAIATPAAAAATLAPAGAT----LGRH-----
      :. . . . :. : : : : : : : : :
ILVB_A RRGIKSSSPSSISAVLNTTNTVTTTTPSPTKPKPKPETFISRFAPDQPRKGA
      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 AEGYARSSGKPGICIIATSGPGATNLVSLGLADALLDSVPLVAITGQVPRRM
      160     170     180     190     200
      160     170     180     190
PDC1_M YGTNRILHHTIGLPDFSQELRCFQTITCYQAIINLDDAHEQIDTA--IA
      :. . : . . . . . . . . . : : : :
ILVB_A IGTDAFQETPI-----VEVTRSITKHNLYLVMDEDIPRIIEEAFFLA
      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 MAAATTTTTTSSSISFSTKPSFSSSKSPLPISRFLPFSLNPKNKSSSSSR
      10      20      30      40      50
      40      50
PDC1_M -----AIIATPAAAAATLAPAGAT----LGRHLA----RR-
      :. : : : : : : : : : : : : : :
ILVB_A RRGIKSSSPSSISAVLNTTNTVTTTTPSPTKPKPKPETFISR-FAPDQPRKG
      60      70      80      90
      60      70      80      90      100
PDC1_M ---LVQI---GASDVFAVPGDFNLTLDDYLIAEPLTLVGCCNELNAGY
      :. : : : : : : : : : : : : : :
ILVB_A ADILVEALERQGVETVFAYPGGASMEIHQALTRSSSIRNVLPRHEQGGVFA
      100     110     120     130     140
      110     120     130     140
PDC1_M AADGYARSRG-VGACAVTFTVGGLSVLNAIAGAYSENLPVVCIVGGPNSN
      : : : : : : : : : : : : : : :
ILVB_A AAEGYARSSGKPGICIIATSGPGATNLVSLGLADALLDSVPLVAI-----
      150     160     170     180     190
      150     160     170     180
PDC1_M DYGTNRILHHTIGLPDFSQELRCFQT----ITCYQAII--NNL----DDA
      : . : : : : : : : : : : : : : :
ILVB_A ---TGQVPRRMIGTDAF-QE-----TFIVEVT--RSITKHNLYLVMDEDI
      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 MAAATTTTTTSSSISFSTKPSFSSSKSPLPISRFLPFSFLNPNKSSSSSR
      10          20          30          40          50

      20          30          40          50
PDC1_M CNGVGALPVANSHAIATPAAAAATLAPAGAT---LGRH-----
      :. . . . . : : : : : : : : : :
ILVB_A RRGIKSSSPSSISAVLNTTTNVTTPSPKPKPETFISRFPDQPRKGA
      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 AEGYARSSGKPGICIAITSGPGATNLVSLGLADALLDVSPLVAITGQVPRRM
      160         170         180         190         200

      160         170         180         190
PDC1_M YGTNRILHHTIGLPDFSQELRCFQTITCYQAIINNLLDAHEQIDTA--IA
      :. . : : : . . . . . : : : : :
ILVB_A IGTDAFQETPI-----VEVTRSITKHNLYLVMDEDIPRIIEEAFFLA
      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 MAAATTTTTTSSSISFSTKPSFSSSKSPLPISRFLPFSFLNPNKSSSSSR
      10          20          30          40          50

      30          40          50
PDC1_M -----SHAIATPAAAAATLAPAGAT---LGRH-----
      :. . . . . : : : . . . . .
ILVB_A RRGIKSSSPSSISAVLNTTTNVTTPSPKPKPETFISRFPDQPRKGA
      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 AEGYARSSGKPGICIAITSGPGATNLVSLGLADALLDVSPLVAITG-----
      160         170         180         190

      160         170         180         190
PDC1_M YGTNRILHHTIGLPDFSQE--LRCFQTITCYQAIINNLLDAHEQIDTA--
      .. . : : . . . . . : : . . . . . : : :
ILVB_A ---QVPRRMIGTDAFQETPIVEVTRSITKHNLYLVMDEDIPRIIEEAFF
      200         210         220         230         240

      200         210         220         230         240
PDC1_M IATALRESKPVYISVSCNLAG-LSHPTFSRD-PVPMFISPRLSNKANLEY
      :. : : : : . . . . . : : : : : : : : :
ILVB_A LATSGRPG-PVLVDVPKDIQQQLAIPNWEQAMRLPGYMS-RMPKPE-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-KTLPSSSTSSTLLP--RSTF--PPF-HHPHK
      . . . . . : : : : : : : : : : : : : : : : : : : : :
ILVB_A MAAATTTTTTSSSISFSTKP-SPSSSKSP-L-PISR--FSLPFLN-PNK
      10      20      30      40

      50      60      70
ILV1_T TTPPLHLTHTHIHIHSQRRR-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--ETFVSRFAPDEPRKGSVDLVEALEREGV-TDVFAYPGGASMEIH
      : : : : : : : : : : : : : : : : : : : : : : : : :
ILVB_A -KPTKPETFISRFPDQPRKGADILVEALERQGVET-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 VSGLADALLDSVPIVAITGQVPRRMIGTDAFQETPIVEVTRSITKHNLYLV
      : : : : : : : : : : : : : : : : : : : : : : : : :
ILVB_A VSGLADALLDSVPLVAITGQVPRRMIGTDAFQETPIVEVTRSITKHNLYLV
      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---PSSSAFSKTLSPSSSTSSTLLPRSTFFPFPHPHKTTPPPL
      . . . . . : : : : : : : : : : : : : : : : : : : :
ILVB_A MAAATTTTTTSSSISFSTKPSPPSSKSP-LPISRFLPFLNPNKSSS---
      10      20      30      40

      50      60      70      80
ILV1_T HLTHTHIHIHSQRRR-----FTISNVISTNQKVSQTE-----KTETF
      : : : : : : : : : : : : : : : : : : : : : : : : :
ILVB_A -----SSRRRGIKSSSPSSISAVLNTTTNVTTTPSPTKPTKPTETF
      50      60      70      80

      90      100     110     120     130
ILV1_T VSRFAPDEPRKGSVDLVEALEREGVTDVFAYPGGASMEIHQALTRSSIIIR
      : : : : : : : : : : : : : : : : : : : : : : : : :
ILVB_A ISRFAPDQPRKGADILVEALERQGVETVFAYPGGASMEIHQALTRSSSIR
      90      100     110     120     130

      140     150     160     170     180
ILV1_T NVLPRHEQGGVFAAEGYARATGFPGVCIAATSGPGATNLVSGLADALLDSV
      : : : : : : : : : : : : : : : : : : : : : : : : :
ILVB_A NVLPRHEQGGVFAAEGYARSSGKPGICIAATSGPGATNLVSGLADALLDSV
      140     150     160     170     180

      190     200     210     220     230
ILV1_T PIVAITGQVPRRMIGTDAFQETPIVEVTRSITKHNLYLVMVDVEDIPRVVRE
      : : : : : : : : : : : : : : : : : : : : : : : : :
ILVB_A PLVAITGQVPRRMIGTDAFQETPIVEVTRSITKHNLYLVMVDVEDIPRIIEE
      190     200     210     220     230

      240     250     260     270     280
ILV1_T AFFLARSGRPGPILIDVPKDIQQQLVIPDWDQPMRLPGYMSRRLPKLPNEM
      : : : : . . . . : : : : : : : : : : : : : : : : : :
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 MAAAAPSPSSSAFSKTLPSSSTSSLLPRSTFFPHHPKHTPPPLHLT
10 20 30 40 50

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

frag_n -----
ILV1_T VEALEREGVTDVFAYPGGASMEIHQALTRSSIIRNVLPREHGGVFAAEG
110 120 130 140 150

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

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

frag_n -----WS-----
:
ILV1_T VPKDIQQQLVIPDWDQPMRLPGYMSRLPKLPNEMLLEQIVRLISESKKPV
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 TVYANYAVDSSDLLLAFGVRFDDRVTGKLEAFASRAKIVHIDIDSAEIGK
360 370 380 390 400

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

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

70
frag_n W-----HITSITT---
: : : :
ILV1_T WLTSGGLGAMGFLPAAIGA AVGRPDEVVVDIDGDGSFIMNVQELATIKV
510 520 530 540 550

80
frag_n -----NDYAHV-----EAB-----CK
. : : : :
ILV1_T ENLPVKIMLLNNQHLMVVQWEDRFYKANRAHTYLGNPSNEAEIIFNMLK
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  GVRFDDDRVTKLEAFASRAKIVHIDIDSAEIGKNKQPHVSICADIKL
      370      380      390      400      410
```