**2. Programmation dynamique : Alignememt des séquences nucléiques (L'algorithme de Needleman et Wunsch)**

**2.1. Les Méthodes d’Alignement de Deux Séquences**

Il existe deux types d’alignements de séquences : global et local. Le premier prend en considération l'ensemble des résidus de chacune des séquences. Si les longueurs des séquences sont différentes, alors la plus courtes va subir des insertions de gaps afin d’arriver à aligner les deux séquences d'une extrémité à l'autre. Cependant dans un alignement global, si uniquement des segments courts sont très similaires entre deux séquences, les autres parties des séquences risquent de diminuer le poids de ces régions. C'est pourquoi d'autres algorithmes d'alignements, dits locaux, basés sur la localisation des zones de similarité sont nés. Le but de ces alignements locaux est de trouver sans prédétermination de longueur les zones les plus similaires entre deux séquences. L'alignement local comporte donc une partie de chacune des séquences et non la totalité des séquences comme dans la plupart des alignements globaux.

 **2.2. Traitement des séquences nucléiques (ADN OU ARN)**

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

 **Notion de score** :

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

 **Les principes de la détermination d'un score**

           Pour qualifier et quantifier la similitude entre séquences, un score est calculé. Celui-ci peut mesurer soit le rapprochement, soit l'éloignement des séquences pour refléter ce qui les sépare. Ce score repose sur un système qui permet d'attribuer un score élémentaire pour chaque position. pour les acides nucléiques, la matrice d'identité ou unitaire est principalement employée. Elle rend compte de l'identité des résidus pour chacune des positions de la comparaison, on parle ainsi de bon ou de mauvais appariement ou bien de bonne ou mauvaise association. Ce critère qui permet déjà d'établir des ressemblances ne suffit pas toujours pour révéler au mieux les similitudes entre séquences. Très rapidement, on s'est aperçu qu'une insertion ou une délétion d'une ou plusieurs bases pouvait améliorer le score d'une comparaison et ainsi faire davantage ressortir les zones identiques ou très proches.

 Pourquoi avons-nous affecté la valeur de 1 dans le cas d’identité et zéro dans le cas contraires.

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

**Etape 1 : calcule de la matrice initiale**,

 Du fait de la pauvreté de l'alphabet que représentent les bases qui composent les molécules d'ADN, Il existe peu de matrices nucléiques. La plus utilisée est certainement la matrice d'identité.

EX : S1 : ATTGAGCTAC S2 : AAGCATACTT

 Il s’agit d’insérer les deux séquences S1 et S2 dans une matrice de sorte que S1 soit à l’horizontal et S2 à la verticale du tableau, puis remplir les cases par 1 : identité des deux nucléotides de S1et de S2 ou 0 sinon.

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

**Etape 2 : calcule de la matrice transformée :** initialisation de la matrice :construction d’une nouvelle matrice à m+2 colonne et n+2 ligne  dans la quelle la 1er ligne et la 1er colonne seront initialisées à zéro, puis l’application de **L'algorithme de Needleman et Wunsch :**

 S e(i,j)+S(i-1,j-1)

S(i,j)= Max S S(i-1,j)

 S(i,j-1)

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

**Etape 3 : parcours de la matrice transformée**

 **2 INDEL** après le C

**INDEL** après le C

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | A | T | T | G | A | G | C | T | A | C |
|  **3 INDEL** après le A | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0Le remplissage de la matrice transformée de haut vers le basEt du gauche vers la droite  |
| A | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| A | 0 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 |
| G | 0 | 1 | 1 | 1 | 2 | 2 | 3 | 3 | 3 | 3 | 3 |
| C | 0 | 1 | 1 | 1 | 2 | 2 | 3 | 4 | 4 | 4 | 4 |
| A | 0 | 1 | 1 | 1 | 2 | 3 | 3 | 4 | 4 | 5 | 5 |
| T | 0 | 1 | 2 | 2 | 2 | 3 | 3 | 4 | 5 | 5 | 5 |
| A | 0 | 1 | 2 | 2 | 2 | 3 | 3 | 4 | 5 | 6 | 6 |
| C | 0 | 1 | 2 | 2 | 2 | 3 | 3 | 4 | 5 | 6 | 7 |
| T | 0 | 1 | 2 | 3 | 3 | 3 | 3 | 4 | 5 | 6 | 7Score =7 |
| T | 0 | 1 | 2 | 3 | 3 | 3 | 3 | 4 | 5 | 6 | 7 |

**Alignement**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Seq1A** | **T** | **T** | **G** | **A** | **G**  | **C** | **-** | **T** | **A** | **C** | **-** | **-** |
| **Seq2A** | **-** | **-** | **-** | **A** | **G** | **C** | **A** | **T** | **A** | **C** | **T** | **T** |

**Résultats**

Identité =7

Gaps = 6

Le pourcentage d’identité=7 /13 x100 =

**Interprétation**

 On peut dire que soit la séquence 1 qu’a subit une mutation par délétion dans les positions
8, 12 et 13 , soit la séquence 2 qu’a subit une mutation par insertion dans les mêmes positions

 Et la séquence 1 qu’a subit une mutation par insetion dans les positions 2, 3 et 4 ou bien la séquence 2 qu’a subit une mutation par délétion dans les mêmes positions.

1. **Programmation dynamique : Alignememt des séquences protéiques (L'algorithme de Needleman et Wunsch)**

Plus difficile à modéliser que celui des nucléotides:

–Un acide aminé peut être remplacé par un autre de différentes façons (code génétique).

–Le nombre de substitutions requises pour passer d’un acide aminé à un autre diffère.

–La probabilité des substitutions au niveau nucléotidique diffère:

–Certaines substitutions peuvent avoir plus ou moins d’effet sur la fonction des protéines.

•Acidité, hydrophobicité, structure des protéines, etc.

Val et Ile :Substitutions conservatrices

**3.1. Les matrices protéiques**

           Si un système basé uniquement sur l'identité donne une sensibilité satisfaisante pour les acides nucléiques, celui-ci devient moins approprié pour les séquences protéiques. Si l'on considère qu'un acide aminé peut être substitué à un autre en fonction de certaines propriétés sans que la structure ou la fonctionnalité d'une protéine soit grandement altérée, on peut classer les acides aminés en familles et obtenir ainsi un système de scores qui rende compte de l'affinité des résidus protéiques entre eux. Les matrices de scores qui en découlent permettront d'augmenter la fiabilité des recherches de similitudes protéiques.. Les scores élémentaires ont été alors déterminés en fonction du nombre commun de nucléotides présents dans les codons des acides aminés, ce qui revient à considérer le minimum de changements nécessaires en bases pour convertir un acide aminé en un autre. Depuis de nombreuses matrices ont été créées et l'on peut classer celles-ci en deux catégories. La première est celle qui regroupe plutôt les matrices issues d'études montrant le caractère de substitution des acides aminées au cours de l'évolution et la deuxième est basée plus particulièrement sur les caractéristiques physico-chimiques des acides aminés.

 **Les matrices protéiques liées à l'évolution**

           **Les matrices de type PAM, la matrice de mutation de Dayhoff**

           Elles sont sans aucun doute celles qui ont été les plus utilisées dans les programmes de comparaison de séquences protéiques. Elles représentent les échanges possibles ou acceptables d'un acide aminé par un autre lors de l'évolution des protéines. Elles ont été déduites de l'étude de 71 familles de protéines (de l'ordre de 1300 séquences) très semblables (moins de 15% de différence) que l'on pouvait facilement aligner. De ces alignements, une matrice de probabilité a été calculée où chaque élément de la matrice donne la probabilité qu'un acide aminé A soit remplacé par un acide aminé B durant une étape d'évolution. Cette matrice de probabilité de mutation correspond en fait à une substitution acceptée pour 100 sites durant un temps d'évolution particulier, c'est à dire une substitution qui ne détruise pas l'activité de la protéine. On parle ainsi d'une 1PAM (Percent Accepted Mutations) matrice. Si l'on multiplie la matrice par elle-même un certain nombre de fois, on obtient une matrice XPAM qui donne des probabilités de substitution pour des distances d'évolution plus grande. Pour être plus facilement utilisable dans les programmes de comparaison de séquences, chaque matrice XPAM est transformée en une matrice de similitudes PAM-X que l'on appelle matrice de mutation de Dayhoff. Cette transformation est effectuée en considérant les fréquences relatives de mutation des acides aminés.

**Les matrices de type BLOSUM (BLOcks SUbstitution Matrix)**

 le degré de substitution des acides aminés a été mesuré en observant des blocs d'acides aminés issus de protéines plus éloignées. Chaque bloc est obtenu par l'alignement multiple sans insertion-délétion de courtes régions très conservées (cf. la base BLOCK). Ces blocs sont utilisés pour regrouper tous les segments de séquences ayant un pourcentage d'identité minimum au sein de leur bloc. On en déduit des fréquences de substitution pour chaque paire d'acides aminés et l'on calcule ensuite une matrice logarithmique de probabilité dénommée BLOSUM (BLOcks SUbstitution Matrix). A chaque pourcentage d'identité correspond une matrice particulière. Ainsi la matrice BLOSUM60 est obtenue en utilisant un seuil d'identité de 60%. Henikoff et Henikoff, ([1992](http://www.dsi.univ-paris5.fr/bio2/biocours/biblio.html#55)) ont réalisé un tel traitement à partir d'une base contenant plus de 2000 blocs.

**3.2 Alignement des séquences protéiques**

**Seq1 : VTEERDAF seq 2 :LTSHEAL**

**Etape 1 :** calcule de la matrice initiale à partir de PAM 250

Il s’agit d’insérer les deux séquences S1 et S2 dans une matrice de sorte que S1 soit à l’horizontal et S2 à la verticale du tableau ,

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

**Etape 2 :** calcule de la matrice transformée. On garde les valeurs de la dernière ligne et la dernière colonne, puis on applique l’algorithme de **Needleman et Wunsch** pour le cas des protéines :

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

Le remplissage de la matrice transformée de bas vers le haut

Et de droite vers la gauche

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

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| V | T | - | E | E | R | D | A | F |
| L | T | S | H | E | - | - | A | L |
| 14 | 12 |  | 9 | 8 |  |  | 4 | 2 |

Indel (Gaps) : 3 égalité : 3 substitution : 3 le score = 49 ( 14 + 12 + 9+ 8 +4+2).