

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Mohamed Khider-BISKRA.

Faculté des sciences exactes et des
sciences de la nature et de la vie



Département d'informatique

Calcul à l'ADN et Bio-informatique

Réalisé par :

Dr. TBERMACINE Ahmed

CHAPITRE 3: ALIGNEMENT MULTIPLE (MULTIPLE ALIGNMENT)

CH3: ALIGNEMENT MULTIPLE

INTRODUCTION

- Après avoir vu le cas particulier de l'alignement de deux séquences, nous généralisons le problème à un nombre quelconque de séquences.
- Dans un premier temps nous allons présenter le problème, et nous verrons que la forte complexité ne permet pas en pratique d'utiliser des méthodes exactes (programmation dynamique).

CH3: ALIGNEMENT MULTIPLE

INTRODUCTION

- Le problème d'alignement multiple de séquences étant très important,
- de nombreux algorithmes ont été développés.
- Nous présenterons donc une classification des méthodes, ainsi que les plus représentatives dans chacune des catégories.

CH3: ALIGNEMENT MULTIPLE

DÉFINITION

- Problème précédent généralisé à k séquences ($k > 2$) = alignement multiple de séquences.
- Généralisation des définitions des propriétés

CH3: ALIGNEMENT MULTIPLE

DÉFINITION

- Soit $S = \{S_1, S_2, \dots, S_k\}$ un ensemble de k séquences définies sur un alphabet Σ , et soit $\Sigma' = \Sigma \cup \{-\}$, où « - » est le symbole pour représenter une brèche dans une séquence. Un alignement S est un ensemble $S' = \{S'_1, S'_2, \dots, S'_k\}$ de séquences sur Σ' , satisfaisant les propriétés suivantes :
 - ❖ les séquences de S' sont toutes de la même longueur n ,
 - ❖ aucune colonne n'est constituée uniquement de gaps.

Le problème de l'alignement multiple consiste donc à déterminer si pour une fonction d'évaluation, S' est le meilleur alignement multiple possible des séquences de S

CH3: ALIGNEMENT MULTIPLE

UTILISATION EN BIO-INFORMATIQUE

- ❖ Déterminer leur degré d'homologie ou de divergence.
- ❖ Construire les arbres phylogénétiques
- ❖ Identifier les motifs dans des familles de protéines
- ❖ Prédiction de l'aspect structurel et fonctionnel de protéines

CH3: ALIGNEMENT MULTIPLE

ÉVALUATION DE L'ALIGNEMENT MULTIPLE

- Le critère d'évaluation utilisé est souvent une fonction de score.
- Fonctions les plus utilisées:
 1. **Somme des paires (Sum of Pairs : SP)**
 2. **Fonction consensus**

CH3: ALIGNEMENT MULTIPLE

ÉVALUATION DE L'ALIGNEMENT MULTIPLE

- Somme des paires (Sum of Pairs : SP)

- C'est la fonction la plus répandue, elle consiste à sommer les scores de paires de séquences alignée dans un alignement multiple A_i .
- Soit A_i un alignement de K séquences $S_1, \dots, S_k$:

$$SP(A_i) = \sum_{i=1}^{k-1} \sum_{j=i+1}^k Sc(S_i S_j)$$

- $Sc(S_i, S_j)$: le score de l'alignement de la paire des séquences S_i et S_j (calculé par une mesure de distance ou de similitude).

CH3: ALIGNEMENT MULTIPLE

ÉVALUATION DE L'ALIGNEMENT MULTIPLE

- Somme des paires (Sum of Pairs : SP)
- Exemple: soit l'alignement A suivant :

$S_1 :$	a	c	-	c	d	b	-
$S_2 :$	-	c	-	a	d	b	d
$S_3 :$	a	-	b	c	d	a	d

- Si on considère que la fonction des distances est :
- $d(x,x)=0$, $d(x,y)=1$ pour $x \neq y$
- Le **SP** de l'alignement A = $Sc(S_1, S_2) + Sc(S_1, S_3) + Sc(S_2, S_3)$
 $= 3+4+5 = 12$.

CH3: ALIGNEMENT MULTIPLE

ÉVALUATION DE L'ALIGNEMENT MULTIPLE

- **Fonction consensus**

- L'idée consiste donc à trouver une séquence consensus C : $C_1, C_2, C_3, \dots, C_L$ où L est la longueur de l'alignement, de telle sorte que chaque symbole C_i de consensus C représente le symbole le plus fréquent à la même position dans l'alignement alignement de N séquences $S_1, S_2, S_3, \dots, S_N$.

CH3: ALIGNEMENT MULTIPLE

ÉVALUATION DE L'ALIGNEMENT MULTIPLE

○ Fonction consensus

- Cette séquence Consensus sert à déterminer le score de l'alignement multiple, en sommant les scores des alignements de paires entre chaque séquence de l'alignement multiple et la séquence Consensus.

$$d(i) = \sum_{j=1}^N d(S'_{j,i}, C_i)$$

- Tels que :

S'_j : est la séquence S_j alignée, i : le $i^{\text{ème}}$ symbole dans la séquence S_j

La séquence consensus C : $C_1 C_2 C_3 \dots C_L$

Les erreurs d'alignement sont définis par $\sum_{i=1}^N d(i)$.

CH3: ALIGNEMENT MULTIPLE

ÉVALUATION DE L'ALIGNEMENT MULTIPLE

- **Fonction consensus**

- **Exemple** soit l'alignement suivant:

$S_1 :$	a	c	-	c	d	b	-
$S_2 :$	-	c	-	a	d	b	d
$S_3 :$	a	-	b	c	d	a	d
Consensus :	a	c	-	c	d	b	d

- Si on considère la fonction des distances est:

- $d(x,x)=0$, $d(x,y)=1$ pour $x \neq y$

CH3: ALIGNEMENT MULTIPLE

ÉVALUATION DE L'ALIGNEMENT MULTIPLE

◦ Fonction consensus

◦ Solution

$$d(1) = d(S_{1,1}, C_1) + d(S_{2,1}, C_1) + d(S_{3,1}, C_1) = 0 + 1 + 0 = 1$$

$$d(2) = d(S_{1,2}, C_2) + d(S_{2,2}, C_2) + d(S_{3,2}, C_2) = 0 + 0 + 1 = 1$$

$$d(3) = d(S_{1,3}, C_3) + d(S_{2,3}, C_3) + d(S_{3,3}, C_3) = 0 + 0 + 1 = 1$$
$$d(4) = d(S_{1,4}, C_4) + d(S_{2,4}, C_4) + d(S_{3,4}, C_4) = 0 + 1 + 0 = 1$$

$$d(5) = d(S_{1,5}, C_5) + d(S_{2,5}, C_5) + d(S_{3,5}, C_5) = 0 + 0 + 0 = 0$$

$$d(6) = d(S_{1,6}, C_6) + d(S_{2,6}, C_6) + d(S_{3,6}, C_6) = 0 + 0 + 1 = 1$$

$$d(7) = d(S_{1,7}, C_7) + d(S_{2,7}, C_7) + d(S_{3,7}, C_7) = 1 + 0 + 0 = 1$$

Alors →

$$\sum_{i=1}^L d(i) = 6$$

CH3: ALIGNEMENT MULTIPLE

LES APPROCHES D'ALIGNEMENTS MULTIPLE

- Dans la littérature, on rencontre trois catégories essentielles :
 - A. méthodes exactes
 - B. Méthodes progressives
 - c. Méthode itératives

CH3: ALIGNEMENT MULTIPLE

LES APPROCHES D'ALIGNEMENTS MULTIPLE

○ Méthodes exactes

- L'algorithme de PD peut être étendu dans le cadre de l'alignement de plus de 2 séquences.
- Cependant sa complexité spatiale est en $O(n^k)$ et sa complexité temporelle est en $O(2^k \times n^k)$, où k est le nombre de séquences à aligner, ce qui rend cet algorithme particulièrement **inadapté** à l'alignement de plus de 6 à 8 séquences.
- Par exemple pour un alignement global de 3 séquences avec gaps linéaires on utilise la formule suivante :

CH3: ALIGNEMENT MULTIPLE

LES APPROCHES D'ALIGNEMENTS MULTIPLE

- initialisation :

$$\begin{aligned} M[0,0,0] &= 0 \\ M[i,0,0] &= i*2g \\ M[0,j,0] &= j*2g \\ M[0,0,k] &= k*2g \end{aligned}$$

- calcul du score optimal:

$$M[i,j,k] = \text{optimise} \left\{ \begin{array}{l} M[i-1,j-1,k-1] + w(a_i, b_j, c_k) \\ M[i,j-1,k-1] + w(-, b_j, c_k) \\ M[i-1,j,k-1] + w(a_i, -, c_k) \\ M[i-1,j-1,k] + w(a_i, b_j, -) \\ M[i-1,j,k] + w(a_i, -, -) \\ M[i,j-1,k] + w(-, b_j, -) \\ M[i,j,k-1] + w(-, -, c_k) \end{array} \right.$$

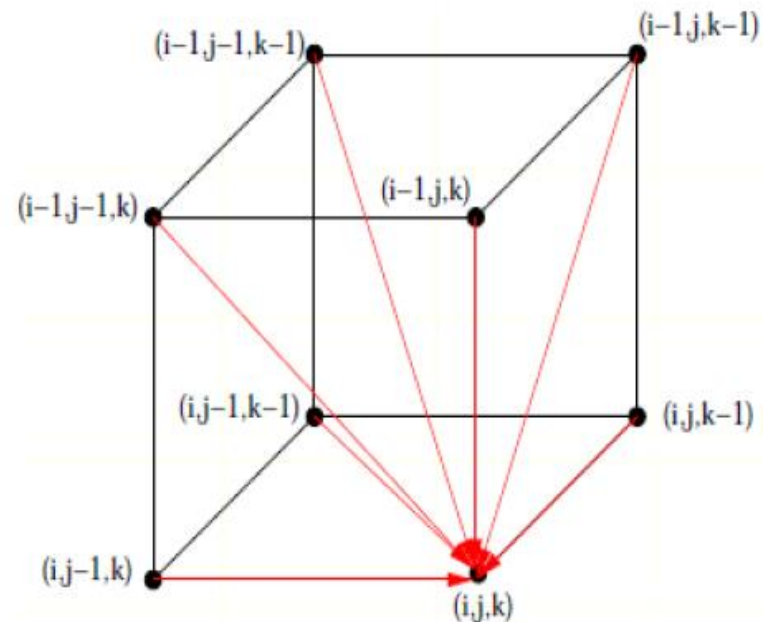
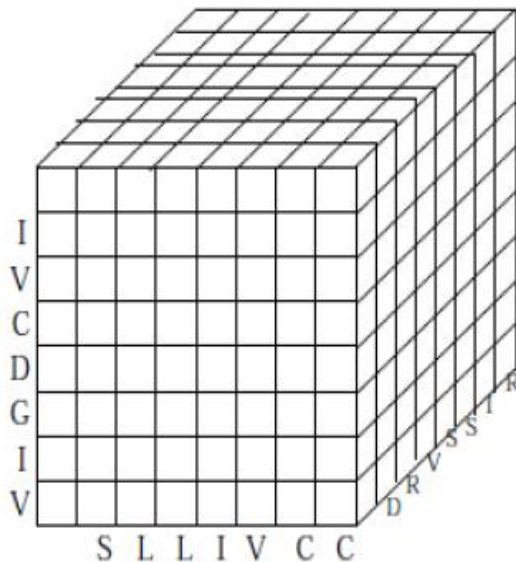
- où $M[i,j,k]$ représente le score de l'alignement de $a[1..i]$ avec $b[1..j]$ avec $c[1..k]$, w matrice de score, g pénalité de gap et optimise est la fonction :
 - maximum* si w est une matrice de score maximisante.
 - minimum* si w est une matrice de score minimisante.

CH3: ALIGNEMENT MULTIPLE

LES APPROCHES D'ALIGNEMENTS MULTIPLE

○ Méthodes exactes

- Avec cette approche en employant une table de score ***n-dimensionnelle*** pour l'alignement multiples de **n** séquences où la valeur d'une cellule, dans le cas de **trois séquence**, dépend de **sept voisins** au lieu des trois voisins dans le cas d'une table de score bidimensionnelle



CH3: ALIGNEMENT MULTIPLE

LES APPROCHES D'ALIGNEMENTS MULTIPLE

○ Méthodes exactes

- Si on doit aligner 10 séquences de longueur 100, il est nécessaire d'effectuer de l'ordre de: $2^{10} \times 100^{10}$ calculs soit environ 10^{23} calculs.
- Si on était capable d'effectuer l'un des calculs en 1 an il faudrait: $10^{23} \times 10^{-9} = 10^{14}$ s, soit 3 millions d'années

CH3: ALIGNEMENT MULTIPLE

LES APPROCHES D'ALIGNEMENTS MULTIPLE

○ Méthodes exactes

- Des alternatives ont été développées afin de s'attaquer à ce problème difficile.
- Citons en particulier:
 - l'alignement progressif (Clustal W et PileUp).
 - l'utilisation de méthodes itératives comme les Algorithmes Génétiques (SAGA).
 - Hidden Markov Model (HMMER)
 - Gibbs Sampling

CH3: ALIGNEMENT MULTIPLE

LES APPROCHES D'ALIGNEMENTS MULTIPLE

◦ Présentation

- l'*alignement progressif* se décompose en plusieurs étapes:
- Calcul d'une *matrice de distance* entre toutes les paires de séquences.
- Construction d'un *arbre-guide* à partir de cette matrice de distances
- Construction de l'*alignement* sur base de l'arbre-guide

CH3: ALIGNEMENT MULTIPLE

LES APPROCHES D'ALIGNEMENTS MULTIPLE

- Méthodes Progressives
- (Matrice de distances entre paires de séquences)
- La première étape d'un alignement progressif consiste à aligner chaque paire de séquences et à calculer leur distance. On regroupe les résultats dans une *matrice de distances*, où:

	seq 1	seq 2	...	seq n
seq 1	d1,1	d1,2	...	d1,n
seq 2	d2,1	d2,2	...	d2,n
...	...	...	...	...
seq n	dn,1	dn,2	...	dn,n

CH3: ALIGNEMENT MULTIPLE

LES APPROCHES D'ALIGNEMENTS MULTIPLE

- Méthodes Progressives
- (Matrice de distances entre paires de séquences)
- Les alignements par paires peuvent être effectués en utilisant: algorithme de Needleman-Wunsch ou une heuristique plus rapide (fasta, blast).
- A partir de chaque alignement par paire, calculer la distance entre les deux séquences :
- $d_{i,j} = s_{i,j} / L_{j,j}$ ou
 - $d_{j,j}$ distance entre les séquences i and j
 - $L_{j,j}$ longueur de l'alignement
 - $s_{j,j}$ nombre de substitutions

CH3: ALIGNEMENT MULTIPLE

LES APPROCHES D'ALIGNEMENTS MULTIPLE

- Méthodes Progressives
- Matrice de distances entre paires de séquences
- Remarques
 - Les gaps ne sont pas pris en compte dans la métrique de distance
 - La matrice est symétrique: $d_{i,j} = d_{j,i}$
 - Les éléments diagonaux sont nuls: $d_{i,i} = 0$

CH3: ALIGNEMENT MULTIPLE

LES APPROCHES D'ALIGNEMENTS MULTIPLE

- Méthodes Progressives
- Construction de l'arbre Guide
- A partir de la matrice de distance, on peut construire un *arbre-guide* par regroupement hiérarchique (par exemple la méthode UPGMA).
- Le principe est d'établir en premier lieu un branchement qui relie les deux séquences les plus proches, puis les séquences un peu moins proches, et ainsi de suite jusqu'à avoir branché toutes les séquences.

CH3: ALIGNEMENT MULTIPLE

LES APPROCHES D'ALIGNEMENTS MULTIPLE

- Méthodes Progressives
- Construction de l'arbre Guide
- **Attention:** l'arbre-guide ne doit en aucun cas être considéré comme inférence des relations phylogénétiques entre séquences.

CH3: ALIGNEMENT MULTIPLE

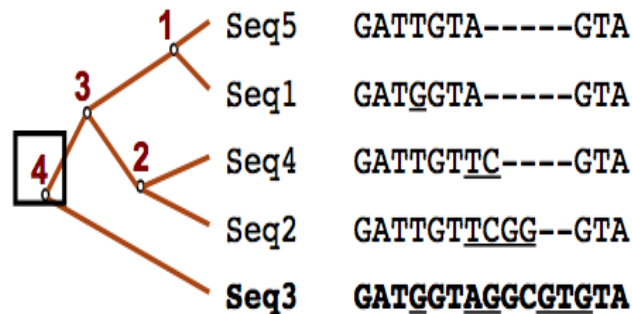
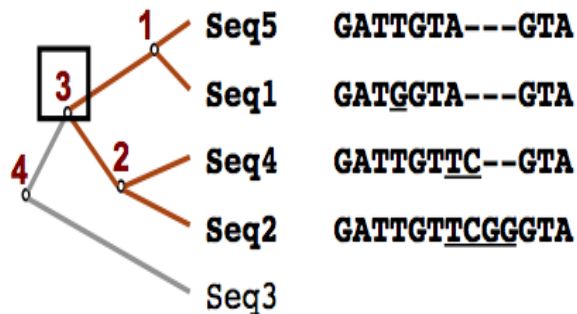
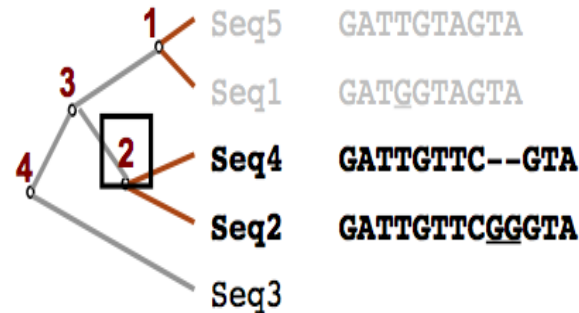
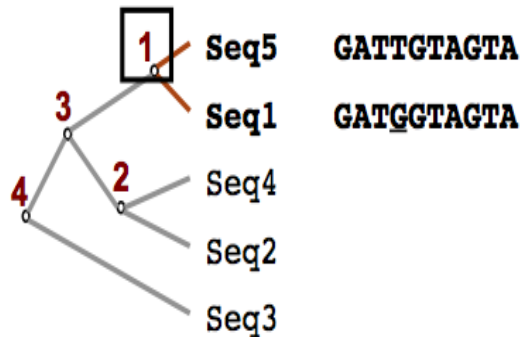
LES APPROCHES D'ALIGNEMENTS MULTIPLE

- Méthodes Progressives
- Construction de l'arbre Guide
- **Algorithme** (création d'un arbre)
- Déterminer dans la matrice des distances les deux séquences S et T devant être reliées.
- Créer un nœud U joignant les deux séquences S et T, de telle sorte que $d(S, U) + d(U, T) = d(S, T)$. Dans l'arbre, le nœud U représente le profil résultant de l'alignement de S et T.
- Mettre à jour la matrice des distances en supprimant S et T, et en les remplaçant par U.
- Si la matrice contient deux séquences ou plus, reprendre en 1.

CH3: ALIGNEMENT MULTIPLE

LES APPROCHES D'ALIGNEMENTS MULTIPLE

- o Méthodes Progressives
- o Construction de l'alignement multiple



CH3: ALIGNEMENT MULTIPLE

LES APPROCHES D'ALIGNEMENTS MULTIPLE

◦ Méthodes Progressives

◦ UPGMA

UPGMA (Unweighted pair group method with arithmetic mean) est le nom d'un algorithme destiné à la construction d'un arbre phylogénétique.

Cette méthode permet la transformation d'une matrice de distances (entre différents organismes, populations, ou séquences de nucléotides) en un arbre enraciné.

CH3: ALIGNEMENT MULTIPLE

LES APPROCHES D'ALIGNEMENTS MULTIPLE

● L'ALGORITHME UPGMA

- Trouver les deux taxons i et j pour lesquels la distance D_{ij} est la plus petite;
- Créer le nouveau groupe (ij) contenant n_{ij} membres avec $n_{ij}=n_i+n_j$;
- Connecter i et j dans l'arbre à un nouveau nœud (ij) qui correspond au nouveau groupe. Attribuer aux branches connectant i à (ij) et j à (ij) la longueur $D_{ij}/2$;
- Calculer la distance entre le nouveau groupe (ij) et tous les autres groupes en utilisant

$$D_{ij,k} = \frac{D_{ik} + D_{jk}}{n_i + n_j}$$

- Éliminer les colonnes et les lignes correspondant aux groupes i et j et ajouter celles correspondant au nouveau groupe (ij);
- Si il reste un seul élément dans la matrice, arrêter, sinon retourner en 1.

CH3: ALIGNEMENT MULTIPLE

LES APPROCHES D'ALIGNEMENTS MULTIPLE

- UPGMA (Exemple)
- La matrice de distance:

	A	B	C	D
A	0	5	2	8
B		0	5	7
C			0	8
D				0

- A, B, C, D sont les quatre taxons que l'on souhaite classer.

CH3: ALIGNEMENT MULTIPLE

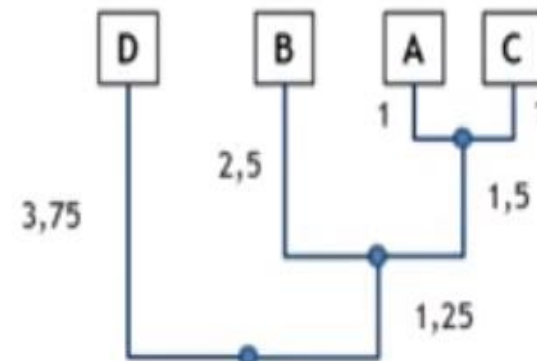
LES APPROCHES D'ALIGNEMENTS MULTIPLE

◦ UPGMA (solution)

Matrice d'origine

	A	B	C	D
A	0	5	2	8
B		0	5	7
C			0	8
D				0

Arbre



CH3: ALIGNEMENT MULTIPLE

LES APPROCHES D'ALIGNEMENTS MULTIPLE

- UPGMA (Exemple)

- Si on considère la matrice de distances associées à un groupe de 4 éléments et que l'on veuille obtenir l'arbre Guide associé::

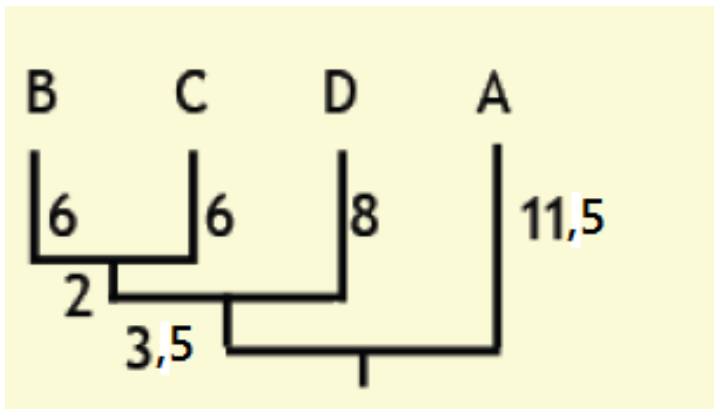
	A	B	C	D
A	0	17	21	27
B		0	12	18
C			0	14
D				0

CH3: ALIGNEMENT MULTIPLE

LES APPROCHES D'ALIGNEMENTS MULTIPLE

○ UPGMA (Exemple)

- Si on considère la matrice de distances associées à un groupe de 4 éléments et que l'on veuille obtenir l'arbre l'arbre guide associé::



	A	B	C	D
A	0	17	21	27
B		0	12	18
C			0	14
D				0

CH3: ALIGNEMENT MULTIPLE

LES APPROCHES D'ALIGNEMENTS MULTIPLE

- Méthodes Progressives (LES PROFILS)
- C'est une technique essentielle pour l'alignement de séquences multiples.
- Un **profil** est une séquence virtuelle formée à partir d'un alignement.
- Il s'agit d'une séquence consensus qui est la plus proche possible de l'ensemble des séquences alignées.
- **L'alignement de profil** aligne une séquence T avec à un ensemble de séquences déjà alignées sous la forme d'un profil S , ou aligne deux profils $S1$ et $S2$.

CH3: ALIGNEMENT MULTIPLE

LES APPROCHES D'ALIGNEMENTS MULTIPLE

- Méthodes Progressives (Alignement des PROFILS)
- Il existe différentes approches pour l'alignement de profils.
- Le plus simple est d'obtenir une séquence consensus de S (CS) et d'aligner T et CS en utilisant la méthode d'alignement par paires que nous avons apprise dans les sections précédentes.
- Chaque fois que nous insérons un espace dans CS , nous insérons un espace correspondant dans toutes les séquences de S .

CH3: ALIGNEMENT MULTIPLE

LES APPROCHES D'ALIGNEMENTS MULTIPLE

- Méthodes Progressives (LES PROFILS)
- **Exemple**
- Supposons que nous voulions aligner la séquence T = "ACG" avec le profile S contenant les trois séquences alignées suivantes:

A	C	-	G	T
A	C	-	G	T
G	C	C	A	T

CH3: ALIGNEMENT MULTIPLE

LES APPROCHES D'ALIGNEMENTS MULTIPLE

- Méthodes Progressives (LES PROFILS)

- Exemple

- La première étape de l'alignement du profil consiste à représenter S avec un profil de fréquence de chaque symbole à chaque colonne.
- L'ensemble des trois séquences alignées comporte cinq symboles (A, C, G, T et le symbole d'intervalle «-»).
- Cet ensemble peut être représenté par le profil illustré dans le tableau suivant.

CH3: ALIGNEMENT MULTIPLE

LES APPROCHES D'ALIGNEMENTS MULTIPLE

o Méthodes Progressives (LES PROFILS)

Consensus CS)	C ₁	C ₂	C ₃	C ₄	C ₅
Symbole					
A	2/3	0	0	1/3	0
C	0	1	1/3	0	0
G	1/3	0	0	2/3	0
T	0	0	0	0	1
-	0	0	2/3	0	0

CH3: ALIGNEMENT MULTIPLE

LES APPROCHES D'ALIGNEMENTS MULTIPLE

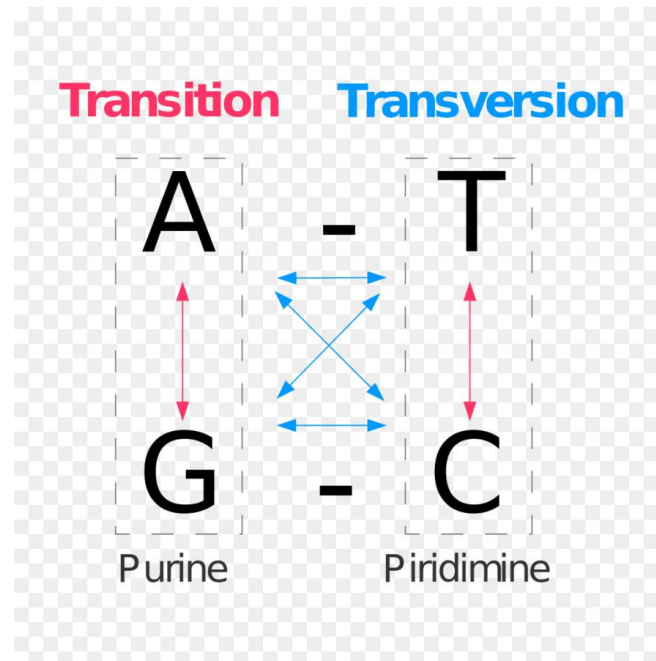
❖ Méthodes Progressives (LES PROFILS)

- 2eme étape: exécuter une version spéciale de la PD.
- Il existe 2 types de correspondance:
 - deux nucléotides identiques $M_{\text{Nuc}} = 2$.
 - deux gaps $M_{\text{Gap}} = 1$.

CH3: ALIGNEMENT MULTIPLE

LES APPROCHES D'ALIGNEMENTS MULTIPLE

- Méthodes Progressives (LES PROFILS)
- Deux types de substitutions :
 - Différence de transition $MM_s = 1$
 - différence transversale $MM_v = -1$
- Pénalité de gap constante $g = -3$.



CH3: ALIGNEMENT MULTIPLE

LES APPROCHES D'ALIGNEMENTS MULTIPLE

o Méthodes Progressives (LES PROFILS)

- initialisation :

$$M[0,0] = 0$$

$$M[i,0] = M[i-1,0] + w(x_i, -) \text{ pour tout } i \text{ de } 1 \text{ à } N$$

$$M[0,j] = M[0,j-1] + w(-, y_j) \text{ pour tout } j \text{ de } 1 \text{ à } P$$

- calcul du score optimal:

$$M[i,j]_{\text{optimise}} = \begin{cases} \nwarrow \text{DIAG: } M[i-1,j-1] & + \sum_{k \in \text{Car}} \text{prob}(cs_{j,k}) \times w(x_i, y_j) \\ \uparrow \text{UP: } M[i-1,j] & + \sum_{k \in \text{Car}} \text{prob}(cs_{j,k}) \times w(x_i, -) \\ \leftarrow \text{LEFT: } M[i,j-1] & + \sum_{k \in \text{Car}} \text{prob}(cs_{j,k}) \times w(-, y_j) \end{cases}$$

CH3: ALIGNEMENT MULTIPLE

- Méthodes Progressives (LES PROFILS)
- Solution =

T/CS	-	C ₁	C ₂	C ₃	C ₄	C ₅
-	0	-3	-6	-9	-12	-15
A	-3					
C	-6					
G	-9					

CH3: ALIGNEMENT MULTIPLE

- Méthodes Progressives (LES PROFILS)
- Solution =

T/CS	-	C ₁	C ₂	C ₃	C ₄	C ₅	i
-	0	-3	-6	-9	-12	-15	0
A	-3	↖ 5/3	← -4/3	← -5/3	← -14/3	← -23/3	1
C	-6	↑ -4/3	↖ 11/3	← 10/3	← 1/3	← -8/3	2
G	-9	↑ -13/3	↑ 2/3	↑ 3	↖ 5	← 2	3
j	0	1	2	3	4	5	

CH3: ALIGNEMENT MULTIPLE

LES APPROCHES D'ALIGNEMENTS MULTIPLE

- Méthodes Progressives (LES PROFILS)

- L'alignement, de la séquence T avec le profile S, obtenu est:

A	C	-	G	T
A	C	-	G	T
G	C	C	A	T
A	C	-	G	-

CH3: ALIGNEMENT MULTIPLE

LES APPROCHES D'ALIGNEMENTS MULTIPLE

- Méthodes Progressives (CLUSTAL W)
- ALGORITHME

1. Alignement de toutes les paires de séquences deux à deux par l'algorithme de Needleman et Wunsh.
2. Construction d'une matrice de distances d'après la divergence mesurée entre chaque paire de séquences
3. Calcul d'un arbre guide à partir de la matrice de distances
4. Alignement progressif des séquences suivant l'ordre de branchement donné par l'arbre

CH3: ALIGNEMENT MULTIPLE

LES APPROCHES D'ALIGNEMENTS MULTIPLE

○ Méthodes Itératives

- les algorithmes itératifs sont beaucoup plus variés dans les méthodes qu'ils utilisent pour réaliser l'alignement des séquences.
- Leur point commun: réaliser l'alignement multiple en prenant simultanément toutes les séquences dès le début.
- Les stratégies utilisées sont très différentes, et nous citons principalement SAGA, car cet algorithme reste une référence importante.

CH3: ALIGNEMENT MULTIPLE

LES APPROCHES D'ALIGNEMENTS MULTIPLE

❖ Méthodes Itératives (SAGA)

- SAGA: Sequence Alignment by Genetic Algorithm
- SAGA est basée sur un algorithme de type génétique.
- Une population G_0 d'alignements est initialement générée, et le programme permet de contrôler son évolution.
- La population initiale est créée en générant cent individus par insertion aléatoire de brèches.

CH3: ALIGNEMENT MULTIPLE

LES APPROCHES D'ALIGNEMENTS MULTIPLE

o Méthodes Itératives (SAGA)

1. sélection de la partie de la population devant être remplacée; par défaut cette valeur est de 50%,
2. utilisation de l'un des nombreux opérateurs les individus sélectionnés,
3. mise à jour de la population en ne conservant que les cent meilleurs individus.