

## **M2 Mycologie**

Blast et alignement multiple

# Recherche de motifs

---

- recherche par similarité d'une séquence déjà connue ou proche de séquences connues
- trouver des alignements significatifs entre une séquence (query) et une banque de séquences (subjects)
- algorithme de Smith-Waterman trop lent donc utilisation de l'algorithme BLAST (Basic Local Alignment Search Tool)

## Principe général (2)

Blast utilise des heuristiques pour accélérer la production de résultats et il permet de calculer la significativité statistique des résultats d'alignement. Attention, ce sont les parties conservées de l'alignement qui sont généralement favorisées.

Matrice de substitution :

- permet d'associer un score à chaque paire de résidus que l'on trouve dans un alignement, sauf cas des nucléotides où la pénalisation est identique quelque soit la substitution
- scores positifs indiquent des substitutions fréquentes ou acceptables ( $>$  hasard), scores négatifs indiquent des mutations rares (contre sélection)

1. définir à partir de la séquence requête  $q$  (longueur  $L$ ) une liste de mots (graines,  $\max L - w + 1$ ) de taille définie  $w$  (taille par défaut de 11 pour l'ADN et de 3 pour les protéines) ; dans le cas des protéines ( $w = 3$ ), utilisation d'une matrice de substitution (PAM ou BLOSUM62)
2. rechercher des alignements exacts (hits, de taille  $w$ ) entre les mots de la liste (ADN) ou de la liste étendue (protéines) et les séquences de la base de donnée
3. chaque hit est étendu à gauche et à droite : L'extension est stoppée lorsque de le score du hit décroît de plus de  $X$  (X-drop)

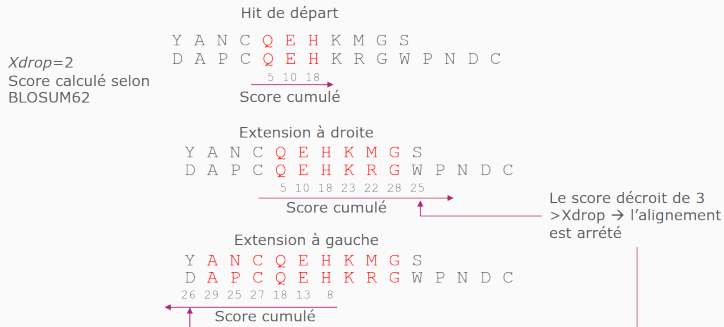


- LMSP = Locally Maximal scoring Segment Pair
- HSP = High scoring Segment Pair
- MSP = Maximum scoring Segment Pair

# Illustration<sup>1</sup>

Query *q* : Y A N C Q E H K M G S

Subject *ti* : D A P C Q E H K R G W P N D C



<sup>1</sup>Source : [https://wikis.univ-lille.fr/bilille/\\_media/cours\\_blast.pdf](https://wikis.univ-lille.fr/bilille/_media/cours_blast.pdf)

On peut toujours aligner deux séquences et définir le meilleur MSP : comment fixer un seuil de significativité statistique (i.e., pour démontrer une réelle homologie) ?

Soit le  $S$  le score d'alignement des deux séquences. On définit la E-value comme l'espérance du nombre  $n$  de MSP de score  $\geq S$  dans 2 séquences aléatoires de même longueur et de même composition. Par exemple, si  $E=10$ , on dira que 10 HSP  $\geq S$  peuvent être trouvés par chance.



## **Alignement multiple**

---

# Principe général

On dispose de  $k$  séquences (NN) de taille variable et on cherche à aligner "au mieux" l'ensemble des séquences, en introduisant des indels. On parle également de résidus équivalents alignés par site.

## Entrée

CATGCGAGTAGTAG  
CATGGTAGTAG  
CCTGGAGTACGTAG  
CATGAGCGTAG

## Sortie

CATGCGAGTA-GTAG  
CATG---GTA-GTAG  
CCTG-GAGTACGTAG  
CATG--AG--CGTAG

## **Progressive method (guide trees were built 2 times.) (FFT-NS-2)**

```
$ mafft test.fasta
```

```
---catgcgagtagtag  
---catggtagtag---  
cctggagtacgtag---  
---catgagcgtag---
```

## **Iterative refinement method (<16) with LOCAL pairwise alignment information (L-INS-i)**

```
$ mafft --auto test.fasta
```

```
catgcgagtag-tag  
catggtagtag----  
cctg-gagtacgtag  
catgagcgtag----
```

## Score d'un alignement multiple

On pénalise plus largement les indels :

$$s(x, x) = 1, s(x, y) = -1, s(x, -) = s(-, x) = -2, s(-, -) = 0$$

Les sites sont considérés comme indépendants, et le score résultat est la somme des scores individuels.

Distance entre deux séquences :  $1 - \underbrace{\frac{\text{\# résidus identiques}}{\text{\# résidus comparés}}}_{\% \text{ identité}}$

1. Alignement optimal : approche exact (par paires) par programmation dynamique + heuristiques, impossible en pratique
2. Alignement progressif : utilisation d'un arbre de guidage à partir d'une matrice de distance et alignement des paires les plus proches (ClustalW – utilisation de profils) [AA]
3. Alignement itératif : tri des séquences à partir d'une matrice de similarités (scores) et alignement par ordre croissant (sans arbre de guidage) (DIALIGN)
4. Alignement hybride (progressif + itératif) : arbre guide à partir de la matrice de distance (UPGMA ou NJ) puis alignement progressif ; amélioration de l'arbre guide, nouvel alignement progressif... ; raffinement (MUSCLE)

- ClustalW/Clustal Omega : efficace mais lent, moins performant si les séquences sont de longueurs très différentes ou présentent peu de similarités
- Muscle : méthode rapide mais approximative pour le calcul des distances (k-mers partagés par paire de séquences) et donc arbre guide moins fiable

- T-coffee : approche hybride combinant les resultats d'alignement global (ClustalW) et local (Lalign)
- MAFFT : utile dans le cas où le nombre de séquence est grand ; bonnes performances en règle générale

1. identification de segments de similarité entre chaque paire de séquences par FFT (méthode FFT-NS-1), puis alignement des paires de séquence (10x plus rapide que Clustal)
2. arbre de guidage basé sur ces alignements et calcul des distances simplifié à l'aide de k-mer
3. alignement progressif, éventuellement répété (comme MUSCLE – FFT-NS-2)

Possibilité de raffiner l'arbre de guidage (séparation de l'arbre en deux puis réaligement des deux moitiés) : méthode FFT-NS-i.

Sources : NCBI, Mafft, <https://wikis.univ-lille.fr/bilille>