

Alignements multiples, Recherche et Extraction de motifs

scp -r **monLogin**@sftpcampus.pasteur.fr:/pasteur/entites/CoursGeno/Bioinfo/AlignMotifs .

Exercice I: Alignements Multiples

Utilisez **clustalX** pour générer les alignements multiples de deux jeux de séquences protéiques : **exoMotif.faa** et **seqs8Lionel.faa**.

Vous devez télécharger les données (**File; Load Sequences**) avant de lancer le calcul (**Alignment; Do Complete Alignment**)

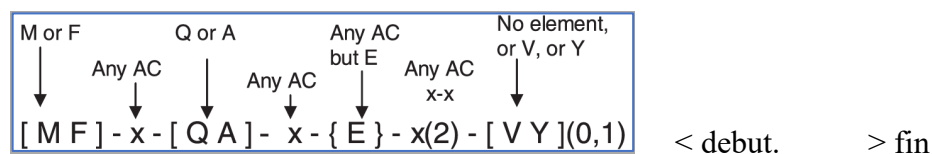
1. Faites les deux alignements avec les paramètres par défaut. Qu'observez-vous?
2. Puis, modifiez les pénalités d'ouverture et d'extension de gap, respectivement, à 2 et 0. Qu'observez-vous?
3. Enfin, refaite l'alignement multiples de **seqs8Lionel.faa** avec l'un des programmes suivant:
 - **clustalO** : <https://www.ebi.ac.uk/Tools/msa/clustalo/>
 - **Muscle** : <https://www.ebi.ac.uk/Tools/msa/muscle/>
 - **mafft** : <https://www.ebi.ac.uk/Tools/msa/mafft/>

Choisissez le format clustal en output.

Exercice II: Définition de motifs

En observant les alignements obtenus précédemment, vous devez voir apparaître des zones conservées/motifs. Modélisez ces motifs sous forme d'une expression rationnelle simple. Pour vous aider, consultez la syntaxe **Prosite** :

http://prosite.expasy.org/scanprosite/scanprosite_doc.html



1. Déterminez les régions bien conservées de l'alignement multiple obtenu avec les séquences **exoMotif.faa**.

2. Maintenant, vous allez modéliser celui des [seqsXLionel](#). Mais utilisez l'alignement précalculé contenu dans le fichier [seqs41Lionel.aln](#) obtenu avec 41 séquences protéiques.

Exercice III: Visualisation du motif

Pour identifier plus facilement la conservation des colonnes, il est possible d'utiliser la représentation [WebLogo](http://weblogo.berkeley.edu/logo.cgi) (<http://weblogo.berkeley.edu/logo.cgi>).

1. Collez les alignements multiples donnés précédemment : [exoMotif.aln](#) et [seqs41Lionel.aln](#). Pour une meilleure lisibilité des résultats, limitez l'affichage à la région qui contient le motif à l'aide de l'option "Logo Range".
2. Est-ce que vous retrouvez le motif que vous avez défini ? Est-ce que des colonnes bien conservées sont visibles ?

Exercice IV: Séquence consensus

Le programme [Cons](http://www.bioinformatics.nl/cgi-bin/emboss/cons) (<http://www.bioinformatics.nl/cgi-bin/emboss/cons>), de la suite de programmes [Emboss](#), calcule la séquence consensus d'un alignement multiple.

Calculez la séquence consensus des séquences protéiques [exoMotif](#) et [seqs41Lionel](#) à partir des alignements. Retrouvez-vous les motifs modélisés précédemment ?

Exercice V: Recherche de motifs dans une banque de séquences

Le programme [ScanProsite](http://prosite.expasy.org/scanprosite/) (<http://prosite.expasy.org/scanprosite/>) permet de localiser dans la banque [Swissprot](#) toutes les séquences contenant un motif spécifié par une expression rationnelle.

Utilisez [ScanProsite: Motif\(s\) to scan for](#) pour rechercher dans Swissprot les séquences contenant les motifs modélisés précédemment pour [exoMotif](#) et [seqXLionel](#).

1. Est-ce que vos motifs correspondent à des motifs connus? Lesquels?
2. Quelles sont les fonctions associées à ces motifs?
3. Combien de séquences ont ces motifs?

Exercice VI: Recherche de motifs contenus dans des séquences

En fait, le travail que nous venons de faire manuellement aurait pu être facilité par l'utilisation de la banque [Prosité](#). Cette banque permet d'analyser une séquence pour détecter la présence de motifs.

Rechercher avec **ScanProsité: Sequence(s) to be scanned** le motif contenu dans une des séquences du fichier: [exoMotif.faa](#) puis [seqs8Lionel.faa](#).

1. Quelle est la taille du motif et le nombre d'occurrences ?

En suivant le lien documentation, vous obtenez la description des motifs utilisés par [Prosité](#). Comparez avec vos motifs.

Exercice VII: En option : Recherche de motifs contenus dans des séquences

L'analyse d'une séquence peut également se faire avec le programme [MotifScan](#) (http://myhits.isb-sib.ch/cgi-bin/motif_scan).

1. Contre quelles banques de motifs, pouvez-vous effectuer la recherche?
2. Faites la recherche contre les profils et pattern Prosité et les HMM Pfam. Que trouvez-vous?

Exercice VIII: Extraction de motifs communs à plusieurs séquences

Le programme [MEME](#) (<http://meme-suite.org/>) découvre des motifs inconnus dans un groupe de séquences.

- Il représente les motifs par une matrice décrivant la probabilité de chaque lettre à chaque position du motif.
- Les motifs déterminés ne contiennent pas de gap (Les motifs contenant des gaps sont divisés en différents motifs).
- Il utilise un modèle statistique pour déterminer la meilleure longueur du motif, le nombre d'occurrence et la description du motif.

1. Utilisez [MEME](#) avec les séquences protéiques [exoMotif.faa](#) et [seq41Lionel.faa](#).
2. Retrouvez-vous le motif précédant ? Enregistrez le motif MEME au format text.

[MEME](#) est associé au programme [MAST](#) qui recherche, dans une banque, les séquences contenant un ou plusieurs motifs connus. [MAST](#) prend en entrée un fichier

contenant la description d'un ou plusieurs motifs et recherche les séquences matchant ce ou ces motifs. Le fichier d'entrée contient les motifs déterminés par [MEME](#).

3. A l'aide de [MAST](#), vous allez rechercher le motif [MEME.txt](#) dans les séquences de la banque [swissprot](#) (Other Databases / Swissprot)

```
meme exoMotif.faa -protein -oc . -nostatus -time 18000 -mod zoops -nmotifs
3 -minw 6 -maxw 50 -objfun classic -markov_order
mast meme.xml exoMotif.faa -oc . -nostatus
```

```
mast -oc . -nostatus -bfile db/swissprot_PROTEIN_1559743200000.faa.bfile -
minseqs 560292 -remcorr -ev 10.0 -df swissprot_PROTEIN_1559743200000.faa -
dl http://www.uniprot.org/uniprot/?query=SEQUENCEID&sort=score meme.txt
db/swissprot_PROTEIN_1559743200000.faa
```