# Formation "Initiation à la bioinformatique": Module 2 Alignement de séquences

## Listes des exercices



Les logiciels que nous allons utiliser dans la première partie sont majoritairement issus de la suite EMBOSS. Ils sont disponibles à plusieurs endroits, par exemple:

- http://bioinf.ibun.unal.edu.co/cgi-bin/emboss
- http://emboss.toulouse.inra.fr/cgi-bin/emboss/
- https://www.bioinformatics.nl/cgi-bin/emboss

## Exercice 1: Logiciels de dotplot

### Les paramètres

Sur le site de Emboss Explorer dans le menu `ALIGNMENT DOT PLOTS` vous disposez de la liste des logiciels de dotplot. Répondez aux questions suivantes sur l'utilisation des paramètres de ces logiciels (cliquez sur le lien **read the manual**) :

- dottup :

  > Quel paramètre modifier pour changer la taille de la fenêtre glissante ? Quelle est la valeur par défaut ? Quelle est la taille minimale ?

- dotmatcher:

  > Quelle est la taille de la fenêtre par défaut ? Quelle est la taille minimale ? Quel est le score par défaut ?

### Exemple de dotplot et interprétation

Voici une séquence d'ADN courte et singulière:

```
tcgcgcgcgcgctgtaagagtgtgtgtggct
```

Tracez avec dottup le dotplot de la séquence sur elle-même, en utilisant la *fenêtre* de **taille minimale** (paramètre *Word size*).

Qu'observez vous sur la diagonale principale ? Pourquoi ?

Que representent les deux carrés hachurés ? A quoi correspondent-ils respectivement sur la séquence ?

Qu'observez vous d'autre ? Expliquez en référence à la séquence.

Réalisez un nouveau dotplot (même paramètres) en comparant la séquence originale contre son **complémentaire inversé** (utilisez le programme `revseq` disponible également sur Emboss Explorer).

Que n'observez-vous plus sur la diagonale principale ? Pourquoi ?

Quel régularité observez vous à nouveau ? A quoi cela correspond t-il sur la séquence ?

En particulier, expliquez désormais pourquoi on n'observe pas deux carrés hachurés, mais un seul.

## Exercice 2: Analyse monoséquence

Le dotplot peut également être utilisé pour étudier les régularités structurelles d'une séquence. Vous allez tester cette approche sur les deux exemples suivants.

### Régularité structurelle

Expliquez les résultats des trois programmes de dotplot (dotpath, dottup et dotmatcher) sur cette séquence.

### Région de faible complexité

Observez la séquence contenue dans le fichier falciparum.fasta. En utilisant dotmatcher et en faisant varier le score de la fenêtre vous devez observer quatre tâches. A quoi correspondent-elles ?

## Exercice 3: Découverte des logiciels d'alignement

Deux logiciels d'alignement de séquences, `matcher` et `stretcher`, sont disponibles également sur Emboss Explorer.

Trouvez lequel des deux fait des alignements locaux? L'autre effectue-t-il des alignements globaux ou semi-globaux? Comment vérifier celà à l'aide d'un exemple?

Quelle(s) différences y'a-t-il avec le logiciel `needle`? Vérifiez celà à l'aide d'un exemple.

## Exercice 4: Comparaison d'un gène et de son ARNm

Pour trouver la structure d'un gène, c'est-a-dire la position des introns, la manière la plus efficace est de comparer la séquence génomique à l'ARNm mature qui lui correspond. Cela veut dire que l'on cherche à construire un alignement composé de régions identiques à 100% (pas de substitution) ou presque (aux erreurs de séquençage près), séparées par des régions d'indel plutôt longues (les introns). Nous allons étudier la séquence génomique du gène `MAKORIN1` chez le poisson *Seriola quinqueradiata*.

Allez rechercher les séquences complètes du géne `MAKORIN1` et de son ARNm sur GQuery.

### Dotplots

Avec le logiciel `dotpath` (Emboss Explorer) réalisez un dotplot.
Que voyez-vous? Combien comptez-vous de diagonales? A quoi correspondent-elles? Quelle est la taille **approximative** du plus petit exon?

### Alignement

Nous allons maintenant essayer de retrouver ces résultats en réalisant un alignement entre les deux séquences.

Utilisez `strecher` pour aligner les deux séquences. Pourquoi choisir un alignement global? Retrouvez-vous ce à quoi vous vous attendiez? Quel est le score de cet alignement?

## Exercice 5: Comparaison de deux séquences protéiques

Nous allons étudier deux séquences protéiques:

```
>prot1
MVMEYLVLEKRLKRLREVLEKRQKDLIVFADNVKNEHNFSAIVRTCDAVGVLYLYYYHAEGKKAKINEGI
TQGSHKWVFIEKVDNPVQKLLEFKNRGFQIVATWLSKESVNFREVDYTKPTVLVVGNELQGVSPEIVEIA
DKKIVIPMYGMAQSLNVSVATGIILYEAQRQREEKGMYSRPSLSEEEIQKILKKWAYEDVIKERKRTLST
S
>prot2
MVMEYLVLEKRLKRLREVLEKRQKDLIVFADNVKNEHNFSAIVRTCDAVATWLSKESVNFREVDYTKPTV
LVVGNELQGVSPEIVEIAVGVLYLYYYHAEGKKAKINEGIS
```

Faire un dotplot de ces 2 séquences avec dotmatcher (Emboss Explorer). **Attention les séquences sont inversées sur les axes du dotplot.**

> Qu'observez-vous?
> Ces protéines partagent des domaines communs. Combien de domaines communs pouvez-vous identifier? Quel est l'ordre des domaines pour chaque protéine? Quelles sont les coordonnées approximatives de ces domaines pour chaque protéine?

Réalisez un alignement semi-global avec needle.

> Qu'observez-vous ? Est-ce que les domaines identifiés précédemment sont retrouvés? Pourquoi?
> Combien y a t-il de gaps? A quoi correspond le pourcentage d'identité? A quoi correspond le pourcentage de similarité?
> Quels sont les paramètres de calcul du score ? Modifiez-les et regardez en quoi l'alignement change.

Réalisez maintenant un alignement local avec matcher.

> Qu'observez-vous ? Comparez et expliquez les différences obtenues entre une méthode d'alignement global (needle) et une méthode d'alignement local (matcher).

Relancez matcher en demandant plus d'alignements en sortie (*Number of alternative matches*)

> Quel nombre d'alignements alternatifs semble le plus approprié? Retrouvez-vous des conclusions similaires à celles obtenues avec dotmatcher?
> A quoi correspond le quatrième alignement retourné par matcher?

## Exercice 6: Conservation de domaine

Vous allez maintenant comparer deux autres séquences: ce sont deux facteurs de transcription *krox 24* et *sp1*, contenus dans les fichiers krox24.fasta et sp1.fasta.

> Construisez un dotplot avec *dotmatcher* (Emboss Explorer) de ces deux séquences. Qu'observez-vous?

Vous devez observer une similitude locale : c'est un motif doigt de zinc, impliqué dans la liaison à l'ADN.

> Comparez ensuite les deux séquences avec un logiciel d'alignement. Utilisez-vous un logiciel d'alignement local ou global? Retrouvez-vous le résultat précédent?

Afin de vérifier le résultat, nous allons interroger une banque de domaines protéiques : Prosite.

> Copiez-coller la partie d'une des deux séquences obtenue dans l'alignement local et faite une recherche de domaine. Cela confirme-t-il les résultats ?

## Exercice 7: Recherche de domaines

Nous allons comparer 3 enzymes de la famille des TPP (Thiamine Pyrophosphate dependent enzymes). Le but de l'exercice est de détecter s'il existe un ou plusieurs domaines communs à ces trois enzymes.

> Obtenez les séquences des trois protéines `PDC1_MAIZE`, `ILVB1_TOBAC` et `ILVB_ARATH` à partir de la banque protein au NCBI.

> Grâce à des dotplots (à vous de choisir le logiciel le mieux approprié sur Emboss Explorer), faites-vous une idée de la similarité entre les trois séquences.

**Impact des pénalités associées aux gaps :**
Réalisez deux alignements entre ILVB_ARATH et ILVB1_TOBAC (dans 2 pages), avec stretcher, avec 2 jeux de paramètres pour les gaps :

- ouverture 12, extension 2, matrice EPAM60
- ouverture 2, extension 2, matrice EPAM60

> Quelles différences remarquez-vous ? Quel est celui des deux alignements qui vous paraît le plus pertinent ?

**Impact de la matrice de score**
Effectuez un alignement global entre ILVB_ARATH et PDC1_MAIZE avec stretcher en utilisant les paramètres par défaut (ouverture de gap=12, extension de gap=2, matrice=EBLOSUM62).
**Note:** le score de cet alignement et le pourcentage d'identité sont faibles.

> Pensez-vous que ce soit un bon alignement ? Pensez-vous que la matrice BLOSUM62 est appropriée dans ce cadre. Quelle matrice pourrait être meilleure ? Pourquoi ?

Essayez avec les matrices PAM. Construisez les alignements avec EPAM30 et EPAM350.

> Quel est le meilleur alignement ? Etait-ce prévisible?

**Identification de domaines conservés**
Il existe au moins un domaine conservé entre les trois séquences.

> En utilisant les outils à votre disposition (dotplots, alignements) identifiez s'il existe des domaines conservés entre les 3 séquences. Si oui, identifiez-les (position et longueur dans les séquences).

> Grâce à la banque de données de domaines Interpro, identifiez les domaines.
> Que peut-on dire sur les domaines protéiques présents (complets, identiques, ...)?

## Exercice 8: Significativité des scores

Nous allons comparer les séquences ADN et peptidiques de la thiorédoxine provenant des organismes *Helicobacter pylori* et *Staphylococcus aureus*.

```
>H.pylori, trxA
atgagtcactatattgaattaactgaagaaaattttgaaagcaccattaa
aaaaggggttgcgttagtggatttttgggcaccatggtgtggtccttgta
agatgctatcccctgtgattgatgaattagctagcgaatatgaaggtaag
gctaagatttgtaaagttaataccgatgagcaagaagaattgagcgcgaa
atttggtattaggagcattcctacgcttttattcacaaaagatggcgaag
ttgtccatcagttggtgggcgtgcaaactaaagtcgctttaaaagagcaa
ttgaacaagcttttaggctag
>S.aureus, trxA
atggcaatcgtaaaagtaacagatgcagattttgattcaaaagtagaatc
tggtgtacaactagtagatttttgggcaacatggtgtggtccatgtaaaa
tgatcgctccggtattagaagaattagcagctgactatgaaggtaaagct
gacattttaaaattagatgttgatgaaaatccatcaactgcagctaaata
tgaagtgatgagtattccaacattaatcgtctttaaagacggtcaaccag
ttgataaagttgttggtttccaaccaaaagaaaacttagctgaagtttta
gataaacatttataa


>H.pylori, TRX
MSHYIELTEENFESTIKKGVALVDFWAPWCGPCKMLSPVIDELASEYEGKAKICKVNTDE
QEELSAKFGIRSIPTLLFTKDGEVVHQLVGVQTKVALKEQLNKLLG
>S.aureus, TRX
MAIVKVTDADFDSKVESGVQLVDFWATWCGPCKMIAPVLEELAADYEGKADILKLDVDEN
PSTAAKYEVMSIPTLIVFKDGQPVDKVVGFQPKENLAEVLDKHL
```

> Réalisez un alignement local entre les séquences d'ADN sur Emboss Explorer. Est-ce que ces séquences se ressemblent? Quel est le pourcentage d'identité entre les séquences?

> Quel est le pourcentage de similarité entre les séquences ? Quel est le score de l'alignement? L'alignement est-il selon vous significatif?

Pour que vous puissiez répondre plus facilement à cette dernière question, nous allons faire une évaluation de la significativité des alignements à l'aide du programme PRSS proposé à l'Université de Virginie (Etats-Unis).

Vous veillerez à selectionner le **bon programme** et à changer le format d'entrée des séquences pour le format fasta.

> Quel est le score de l'alignement obtenu entre les 2 séquences? Combien de fois un score meilleur est-il attendu? Est-ce que l'alignement est significatif?

Réalisez un alignement local entre les séquences protéiques.

> Est-ce que ces séquences se ressemblent ? Quel est le pourcentage d'identité entre les séquences? Quel est le pourcentage de similarité entre les séquences? Quel est le score de l'alignement?

De la même manière que pour l'ADN, estimez la significativité de l'alignement des séquences protéiques.

**Attention:** changez la matrice de score pour **BlastP62**.

> Est-ce que l'alignement est significatif?

> Comparez les valeurs de significativité obtenues pour l'ADN et les protéines.
> Quel alignement est le plus significatif? Est-ce en accord avec ce à quoi l'on s'attend? Que dire de la comparaison des pourcentages d'identité obtenus pour les deux alignements? Que dire de leur pourcentage de similarité?

## Exercice 9: BLAST et comparaison de séquences

Voici deux petites séquences que l'on souhaite comparer:

```
>Sequence_1
ATTGATTCATTCATTCATTCATTCATTC

>Sequence_2
ATTGATTGATTGATTGATTGATTGATTG
```

> Observez ces séquences. Quelles sont leurs particularités? Quel est le pourcentage d'identité attendu?

Réalisez un alignement global avec needle sur Emboss Explorer.

> Quel est le pourcentage d'identité? Est-ce que needle retourne l'alignement attendu?

Il est possible d'utiliser BLAST pour comparer la séquence requête avec une ou plusieurs séquences de votre choix et non contre une banque de données. Pour ce faire, cochez la case *Align two or more sequences*. Entrez les deux séquences que l'on souhaite comparer et lancez *Somewhat similar sequences (blastn)*.

> Est-ce que BLAST parvient à identifier une homologie entre ces deux séquences? Pourquoi?

En cliquant sur *edit search* relancez la requête en ajustant les paramètres.

**Note:** il y a 2 paramètres à changer.

> Est-ce que cette fois BLAST parvient à identifier une homologie entre ces deux séquences? Quel est le pourcentage d'identité obtenu? Expliquez ce résultat.

## Exercice 10: Y a-t-il des insertions de séquences de HIV dans le génome de SARS-CoV-2 ?

Dans cet exercice nous allons utiliser BLAST pour repérer les régions similaires entre le génome de SARS-CoV-2 (MT019529.1) et la séquence référence du génome de virus de SIDA (HIV-1 taxid:11676). Cet exercice provient d'un TP de Jacques van Helden.

Lancez BLAST avec la séquence génomique du SARS-CoV-2 MT019529.1 contre la banque RefSeq Genome Database *(refseq_genomes)* pour se concentrer uniquement sur les génomes de référence. Dans le champs *Organism*, pour limiter la recherche aux séquences de HIV-1, entrez le taxid **11676**. Choisissez le programme Somewhat *similar sequences (blastn)* pour augmenter la sensibilité de la recherche. Cochez la case Show results in a new window pour que les résultats apparaissent dans une nouvelle fenêtre. Dans *Algorithm parameters*, chagez *Expect threshold* à 10 pour afficher les hits avec une E-valeur jusqu'à 10. Gardez la page de résultats ouverte pour pouvoir répondre aux questions.

> Quel est le nombre de fragments alignés entre les génomes de SARS-CoV-2 et de HIV-1?
> Quelle est la meilleure et la pire E-valeur parmi ces alignements?
> Quelle est la longueur de l'alignement le plus long?
> Quel est le pourcentage d'identité le plus élevé et le plus bas parmi ces alignements?

Le génome de SARS-CoV-2 a 29899 nucléotides. Nous allons générer une séquence aléatoire (suite des nucléotides aléatoires) de même longueur et comparer cette séquence avec le génome de HIV-1 à l'aide de BLAST.
Sur le site web Regulatory Sequence Analysis Tools [RSAT], utilisez l'outil random sequence pour générer une séquence aléatoire de 29899 nt :

- Sequence length : 29899
- Number of sequences : 1
- Background model : Independent and equiprobable nucleotides

Réalisez un BLAST avec cette séquence aléatoire comme séquence requête contre le génome de HIV-1 (mêmes paramètres que précédemment).

> Quel est le nombre de fragments alignés entre les génomes de SARS-CoV-2 et de HIV-1?
> Quelle est la meilleure et la pire E-valeur parmi ces alignements?
> Quelle est la longueur de l'alignement le plus long?
> Quel est le pourcentage d'identité le plus élevé et le plus bas parmi ces alignements?
>
> Répondez par VRAI ou FAUX sur base de la comparaison entre les génomes de SARS-CoV-2 et de HIV-1.
>
> - Le très haut pourcentage d'identité des alignements est suffisant pour conclure sur l'homologie entre des fragments alignés des deux génomes.
> - Les longueurs des alignements sont suffisamment élevées pour inférer l'homologie entre les fragments alignés.
> - Les E-valeurs indiquent qu'il est probable qu'un tel niveau de similarité résulte du hasard
> - Le fait d'avoir plusieurs fragments alignés entre les deux génomes indique une similarité significative entre les génomes
>
> **Conclusion:** Y a-t-il des insertions de séquences de HIV dans le génome de SARS-CoV-2 ?

## Exercice 11: BLAST, répétitions et alignements splicés

Nous allons étudier un fragment du génome humain dont la séquence est disponible ici.

#### Partie 1: BLAST et répétitions

### Recherche de la séquence codante par homologie de séquence

Dans un premier temps, nous allons comparer la **séquence génomique** aux **protéines** de la banque de données. Cela nous permettra peut-être de trouver des protéines de la famille de celle codée par notre séquence génomique.
Lancer la bonne version de BLAST **en filtrant** sur les séquences de l'homme (*Organism = Homo sapiens (taxid:9606)*).

> A quelle famille appartiennent les protéines trouvées par BLAST?
> Est-ce que ces protéines sont intéressantes?
> Quelles sont les positions des régions qui ressemblent à ces protéines ?

### Recherche des répétitions

Les protéines trouvées font partie de la famille des transcriptases reverses de type LINE. Les LINE (Long INterspersed repeated sequences) sont des répétitions très répandues sur le génome humain. Elles couvrent 14% du génome et mesurent 6 à 8 kb de long. Comme il y en a beaucoup dans le génome humain, la banque contient beaucoup de ces séquences. Notre séquence contient peut-être d'autres séquences codantes, mais elles sont masquées par les séquences de type LINE. Il faut donc masquer les séquences répétées connues puis relancer BLAST.
Le logiciel RepeatMasker compare une séquence à une banque de familles de séquences répétées. Il masque les régions qui ressemblent à des répétitions connues en les remplaçant par des N (lettre qui symbolise n'importe quel nucléotide).

> Est-ce que notre séquence contient beaucoup de répétitions ?
> Est-ce que les régions trouvées par BLAST ont également été trouvées par RepeatMasker ?

Vous trouverez la séquence avec les répétitions masquées parmis les pages de résultats de RepeatMasker. Il est maintenant possible de copier-coller cette séquence dans BLAST pour obtenir des séquences qui ressemblent aux régions non masquées. Relancez BLAST avec la séquence contenant les répétitions masquées (**sur tous les organismes de la banque**).

> Est-ce que d'autres protéines sont trouvées?
> Est-ce que les protéines trouvées ont toutes (ou presque) la même fonction?
> Si oui, quelle est cette fonction? De combien d'exons codants semble être composé le gène?
> Vérifiez que les différents exons sont sur le même brin. Pourquoi ne sont-ils pas dans la même phase de lecture?

Récuperez la séquence de la séquence protéique `EHH56202.1 Transcobalamin-1 [Macaca fascicularis]`.

#### Partie 2: alignements splicés

### Prédiction de la structure du gène

Maintenant que nous avons réussi à sélectionner des protéines qui ressemblent à celles codées par notre gène, nous pouvons utiliser GeneWise pour prédire la position des exons codants présents sur notre séquence.

> Est-ce que les résultats de GeneWise sont satisfaisants ?
> Combien d'exons codants sont prédits par GeneWise ?
> Quelles sont les positions de début et de fin des exons prédits ? Les résultats sont-ils les mêmes que ceux obtenus par BLAST ?

### Recherche des ARNm codés par notre gène

Il est également possible de comparer notre séquence génomique à des séquences d'ARNm humains. Pour cela, il est préférable d'utiliser la page de Blast dédiée aux génomes complets (partie *"BLAST Genomes"*) et plus précisemment, celle dédiée à l'humain (lien *"Human"*). Comme nous voulons comparer une séquence génomique humaine à des ARNm humains, nous pouvons utiliser megablast (prog proposé par défaut). Pour la banque on choisi : RefSeq RNA.

> Est-ce que des ARNm s'alignent avec notre séquence ?
> De combien d'exons notre gène semble être composé ?

Récupérez la séquence d'ARNm retournée par BLAST.

### Reconstruction de la structure du gène

Effectuez un alignement global avec needle sur Emboss Explorer entre la séquence génomique et l'ARNm.

> Est-ce que l'alignement vous semble pertinent? Pourquoi, d'après vous, certains exons sont-ils mal alignés?

Est2genome permet d'aligner l'ARNm trouvé à l'aide de Blast avec notre séquence génomique afin de prédire la structure complète du gène (y compris les régions 5' et 3' UTR). Utilisez les résultats obtenus à l'aide d'Est2genome afin de répondre aux questions suivantes :

> Combien d'exons sont prédits ?
> Quelles sont les positions des exons prédits ?
> Pourquoi les positions de début et de fin sont différentes de celles observées lors de

l'alignement avec la protéine?
Est-ce que les prédictions semblent fiables ?
Est-ce que les positions prédites par Est2Genome correspondent aux positions de début et de fin des alignements données par Blast ?

## Exercice 12: Alignement de protéines avec de longs gaps

Nous allons étudier trois protéines : une protéine de *Escherichia coli* qui porte deux fonctions (EC 4.1.1.48 et EC 5.3.1.24) et deux protéines de *Xylella fastidiosa* qui portent chacune une de ces deux fonctions :

```
>trpC, EC:4.1.1.48 et 5.3.1.2, E. coli
MMQTVLAKIVADKAIWVEARKQQQPLASFQNEVQPSTRHFYDALQGARTAFILECKKASP
SKGVIRDDFDPARIAAIYKHYASAISVLTDEKYFQGSFNFLPIVSQIAPQPILCKDFIID
PYQIYLARYYQADACLLMLSVLDDDQYRQLAAVAHSLEMGVLTEVSNEEEQERAIALGAK
VVGINNRDLRDLSIDLNRTRELAPKLGHNVTVISESGINTYAQVRELSHFANGFLIGSAL
MAHDDLHAAVRRVLLGENKVCGLTRGQDAKAAYDAGAIYGGLIFVATSPRCVNVEQAQEV
MAAAPLQYVGVFRNHDIADVVDKAKVLSLAAVQLHGNEEQLYIDTLREALPAHVAIWKAL
SVGETLPAREFQHVDKYVLDNGQGGSGQRFDWSLLNGQSLGNVLLAGGLGADNCVEAAQT
GCAGLDFNSAVESQPGIKDARLLASVFQTLRAY

>EC:5.3.1.24, xfa
MALAYGSECMNISPYRTRIKFCGMTRVGDVRLASELGVDAVGLIFASGSSRLLTVSAACA
IRRTVAPMVNVVALFQNNSADEIHTVVRTVRPTLLQFHGEEEDAFCRTFNVPYLKAIPMA
GAEAKRICTRTLYLKYPNAAGFIFDSHLKGGTGQTFDWSRLPIDLQHPFLLAGGITPENV
FDAIAATVPWGVDVSSGIELQPGIKDGDKMRQFVEEVRRADGRRLFGVA

>EC:4.1.1.48, xfa
MSNILTKIIAWKVEEIAERLLHVSQAELVARCADLPTPRGFAGALQATIAHGDPAVIAEI
KKASPSKGVLREDFRPAEIAISYELGGASCLSVLTDVHFFKGHDDYLSQARDACTLPVLR
KDFTIDPYQVYEARVLGADCILLIVAALDDAQLVDLSGLALQLGMDVLVEVHDIDELERA
IQISAPLIGINNRNLSTFNVSLETTLTMKGLVPRDRLLVSESGILTSADVQRLRAAGVNA
FLVGEAFMRATEPGESLREMFFIT
```

La protéine de *E. coli* possède la fonction enzymatique EC 4.1.1.48 au début et la fonction enzymatique EC 5.3.1.24 à la fin de sa séquence. Nous allons tester si les programmes d'alignement multiple retrouvent bien cette configuration.

Comparer les résultats obtenus avec ClustW et Clustal Omega.
Parmi les programmes suivants: Dialign, MAFFT, MUSCLE et T-Coffee quels sont ceux qui construisent l'alignement multiple attendu ?

**Note: Pour certains logiciels, il est possible de choisir l'ordre des séquences dans l'alignement produit (*Option ORDER, input=même ordre que le fichier d'entrée ou aligned=ordre produit par l'alignement*) ce qui peut s'avérer très utile pour comparer des alignements entre eux.**

## Exercice 13: Etude d'une famille de protéines

Nous allons étudier une famille de protéines au sein d'un même génome, avec un ensemble de séquences très conservées (duplication de gènes) et un gène ayant une fonction proche, mais une séquence éloignée.

Retrouvez sur le site du NCBI (banque *Protein*) les séquences qui portent les numéros d'accession : NP_015040 NP_013571 NP_014060 NP_010834 NP_011117 NP_015296

Quel quel organisme proviennent ces séquences ?
Quelle est la fonction de ces protéines ?

Mémorisez les séquences de ces protéines au format FASTA (*display settings -> fasta text*) et gardez la liste de résultats ouverte.

#### 1. Alignement multiple.

Effectuez un alignement multiple de ces séquences à l'aide des programmes suivants:
Dialign, ClustW, Clustal Omega, MAFFT, MUSCLE et T-Coffee.

Est-ce que les alignements trouvés sont identiques ?
Lesquels semblent les plus satisfaisants ?

**2. Qualité de l'alignement.**

Le meilleur moyen d'estimer la qualité d'un alignement est de vérifier si les régions connues pour avoir la même fonction biologique sont bien alignées entre elles.

Faites une recherche de motifs sur les 6 séquences en même temps dans Prosite (partie *Quick Scan mode of ScanProsite*).

Quels sont les domaines communs aux différentes séquences?
Quelles sont leurs positions sur chacune des séquences?

Afin d'évaluer la qualité des alignements, il faut repérer ces domaines dans les alignements obtenus, puis vérifier que les régions contenant ces domaines sont bien alignées les unes avec les autres.

Voici certains alignements avec les domaines colorés de *Hélicase ATP-Binding* et *Hélicase C Terminal*:

- ClustalW
- Clustal Omega
- Dialign

Quels sont les programmes d'alignement multiple qui alignent correctement les domaines fonctionnels ?

Nous allons maintenant évaluer les résultats obtenus avec MAFFT. Afin de pouvoir facilement obtenir les positions sur chaque séquence, nous allons utiliser un éditeur.
Télécharger SeaView (ou ici) ainsi que les résultats de MAFFT (bouton *Download Alignment File* ). Ouvrez SeaView (double-click) et charger l'alignement multiple.

Localisez les 2 domaines sur chacune des 6 séquences.
Est-ce que les domaines sont correctement alignés?

## Exercice 14 (facultatif): Etude d'une famille de protéines (EPB)

Le but de cet exercice est de comparer les résultats de plusieurs logiciels d'alignement multiple lors de l'alignement d'un ensemble de protéines qui sont identiques à quelques délétions près. Pour ce faire, nous allons utiliser un jeu de données contenant 11 protéines membranaires de l'érithrocyte chez l'humain produites par épissage alternatif (*EPB: erythrocyte membrane protein band 4.1*).
En plus des 11 séquences protéiques non-alignées disponibles ici, nous disposons également de l'alignement optimal pour ces séquences disponible ici.

Avec seaView, ouvrez le fichier contenant les séquences non-alignées.
**Note:** ici seaview est utilisé uniquement comme un éditeur de séquences, pas d'alignement!
Est-ce que les 11 séquences se ressemblent? Quelles sont les différences observées?

Avec seaView, ouvrez maintenant le fichier contenant l'alignement optimal.

Observez cet alignement et garder la fenêtre ouverte pour la suite de l'exercice.

Effectuez un alignement multiple de ces 11 séquences à l'aide des programmes suivants:
Dialign, ClustW, Clustal Omega, MAFFT, MUSCLE et T-Coffee.

Quel(s) logiciel(s) retourne l'alignement attendu?

**Note:** cet exercice représente simplement un test sur exemple précis. Sur un autre problème les performances des méthodes d'alignement seraient probablement très différentes. Cependant, cet exemple simple vous donne des indications sur les biais introduits par chaque méthode.

## Exercice 15: Modélisation de motifs biologiques

Cet exercice porte sur l'étude de motifs biologiques que ce soit dans les séquences ADN ou protéiques. Le fil conducteur de cette partie est l'étude d'une famille de facteurs de transcription qui possèdent un motif de type "basic leucine zipper" (bZIP). On notera que les protéines humaines appartenant à cette famille sont peu conservées.

### Détermination d'un motif caractéristique d'une famille de protéines

Il existe plusieurs représentations possibles pour un motif biologique (ex : pseudo-expression régulière, profile, HMM, alignement, ...). Nous allons essayer de construire un motif de type pseudo-expression régulière sur les 43 protéines suivantes :

- Non alignées, au format FASTA.
- Alignées avec Multalin, au format FASTA.
- Alignées avec Multalin (résultats obtenus à l'aide de l'interface du PBIL).

> Déterminer les positions approximatives de début et de fin de la région conservée entre les séquences de cette famille.
>
> Trouvez l'entrée Prosite correspondant au motif bZIP (partie "*Search*").
> Recopiez l'expression régulière modélisant le motif.

#### 1. Lecture de l'alignement à l'aide de WebLogo

Pour identifier plus facilement la conservation des colonnes, il est possible d'utiliser la représentation WebLogo.

Collez l'alignement multiple donné précédemment au format FASTA. Pour une meilleure lisibilité des résultats, nous allons limiter l'affichage à la région qui contient le motif bZIP à l'aide de l'option "Logo Range:" (positions définies précédemment). De plus, nous allons doubler la taille de l'image en indiquant les valeurs 36 et 10 dans l'option "Logo Size per Line".

> Est-ce que des colonnes bien conservées sont visibles ?
> Est-ce que l'on retrouve plus facilement l'expression régulière bZIP dans cette représentation ?
> Est-ce qu'une amélioration de l'alignement peut être envisagée pour se rapprocher du motif bZIP ?
> Gardez cette image ouverte.

#### 2. Méthode d'extraction d'un motif.

Pratt recherche des motifs communs à un ensemble de séquences ADN ou protéiques **non alignées**, sous la forme de pseudo-expressions régulières. Lancez Pratt sur les séquences non alignées de la famille bZIP fournies ci dessus, en choisissant "View PRATT output file" plutôt que "Directly submit best pattern to ScanProsite".

> Est-ce que Pratt retrouve des motifs qui vous semblent pertinents par rapport à ce qu'il peut être vu à l'aide de WebLogo (l'alignement) ? *Attention, les motifs d'intérêt trouvés par Pratt se situent dans la partie "Best Patterns (**after** refinement phase):"*
> Est-ce que l'expression régulière de bZIP est au moins partiellement trouvée par Pratt ?

#### 3. Vérification de la qualité d'un motif.

Pour vérifier si un motif est bien caractéristique d'une famille de séquences, il faut le tester contre une banque de séquences protéiques. Le plus simple est de choisir SwissProt, la banque de protéines annotées par des experts car la fonction des protéines est donnée systématiquement et est fiable. Les résultats attendus pour un bon motif sont :

- Il retrouve toutes les séquences de la famille considérée (ou presque).
- Il ne retrouve aucune séquence d'une autre famille (ou presque).

Nous allons tester le bon comportement des motifs trouvés par Pratt. Le site ScanProsite permet non seulement d'étudier une séquence protéique en cherchant les motifs de la banque Prosite qu'elle contient ; mais aussi de rechercher une expression régulière (même syntaxe que Pratt) sur toutes les protéines de SwissProt.

Testez le meilleur site déterminé par Pratt contre la banque SwissProt, limitée aux séquences qui proviennent de l'Homme (option "Filter(s):" -> "On taxonomy:" dans le Step 2). Dans la partie "Format", choisissez le mode "text" pour accélérer l'affichage des résultats.

> Combien d'entrées sont trouvées ?
> Le motif est-il représentatif de la famille étudiée?

Le motif vous semble-t-il pertinent?

Par défaut, Pratt recherche des motifs conservés dans toutes les séquences données en entrée. Mais, les motifs les plus pertinents ne sont pas toujours bien conservés dans l'ensemble des séquences de départ. Relancez Pratt en diminuant le pourcentage de séquence à apparier à 80%.

Est-ce les motifs trouvés semblent plus pertinents que ceux trouvés avec 100% des séquences à apparier ?
Relancer une recherche du meilleur motif trouvé par Pratt dans les séquences humaines de SwissProt (n'oubliez pas de prendre celui de la liste "Best Patterns (after refinement phase):"). Est-ce que la qualité du motif est meilleure?
Est-ce que les nouveaux motifs trouvés par Pratt correspondent au moins en partie au motif Prosite?

## Etude d'un site de fixation de facteur de transcription.

Maintenant que nous avons étudié les protéines, nous allons étudier le site de fixation d'un facteur de transcription de la famille bZIP : AP1_human.

#### 1. Récupération de la séquence.

Recherchez dans la banque ENA de l'EBI l'entrée portant l'identifiant AF077374.

De quel organisme provient cette séquence?
Notez la position du site AP-1 dans l'entrée AF077374.
Ce site a-t-il été validé expérimentalement?

#### 2. Détermination d'une expression régulière représentant le site.

Voici des séquences contenant le site de fixation du facteur de transcription de la famille bZIP et l'alignement multiple réalisé avec ClustalW. Faites le WebLogo à partir de l'alignement (au format clustal) en demandant également d'agrandir l'image à 36 X 10 cm pour une meilleure lisibilité. **Attention:** précisez qu'il s'agit de séquences nucléiques.

Est-ce que le motif est bien conservé sur toutes les postions ?
Quelle expression régulière peut-on définir à partir de cette représentation ?

#### 3. Recherche de l'expression régulière déterminée.

Nous allons rechercher l'expression régulière déterminée à partir du WebLogo contre la séquence de l'entrée AF077374 qui contient un site de fixation AP-1 déterminé expérimentalement. Pour ce faire, nous allons utiliser le logiciel Fuzznuc.

Vous pouvez lancer Fuzznuc avec votre expression régulière du type Prosite (**ATTENTION**: dans Fuzznuc le caractère jocker x est remplacé par le caractère N). Précisez également que la recherche doit être faite sur les deux brins ("Search complementary strand").

Combien de fois l'expression régulière est trouvée dans l'entrée ?
Est-ce que le site déterminé expérimentalement a été trouvé ?
Si ce n'est pas le cas, recherchez quelle en est la raison et modifiez l'expression régulière pour le trouver.
Combien de sites trouvez-vous à présent ?

#### 4. Construction et recherche d'un profil.

La représentation d'un site est plus fiable si l'on passe par un profil plutôt qu'une expression régulière. Toujours sur le site EMBOSS, construisez un profil du type Gribskov à partir de l'alignement à l'aide de Prophecy. Une fois le profil créé, vous pouvez le rechercher dans l'entrée AF077374 à l'aide de Prophet [lien alternatif pour Prophet].

Quelle est la taille du profil construit ?
Est-il plus long que l'expression régulière ? Pourquoi ?
Combien de fois le profil est trouvé dans la séquence de l'entrée ?
Est-ce que le profil est plus stringeant (strict) que l'expression régulière ?
Est-ce que tous les sites trouvés par le profil sont également trouvés par l'expression

régulière ?

Est-ce que le site déterminé expérimentalement est trouvé par le profil ?

## Exercice 16:

L'endonucléase III est une enzyme qui intervient dans la réparation de l'ADN. Elle présente une activité N-glycosylase et AP-lyase. Elle est présente dans beaucoup d'espèces. Nous allons étudier cette famille de protéines chez différents orgasnimes afin d'identifier s'il existe des domaines conservés entre les séquences de différentes espèces.
Nous allons d'abord etudier ces séquences chez les bactéries. Voici un fichier contenant des séquences d'endonucléase III provenant de différentes bactéries. Réalisez un alignement multiple de ces séquences à l'aide de Clustal Omega.

Est-ce que ces séquences présentent des régions conservées?

Les séquences étant assez proches il y a beaucoup de régions communes. Recherchez sur ScanProsite les domaines connus sur la séquence d' escherichia coli.

Quels sont les domaines identifiés par prosite? Quelles sont leurs positions?

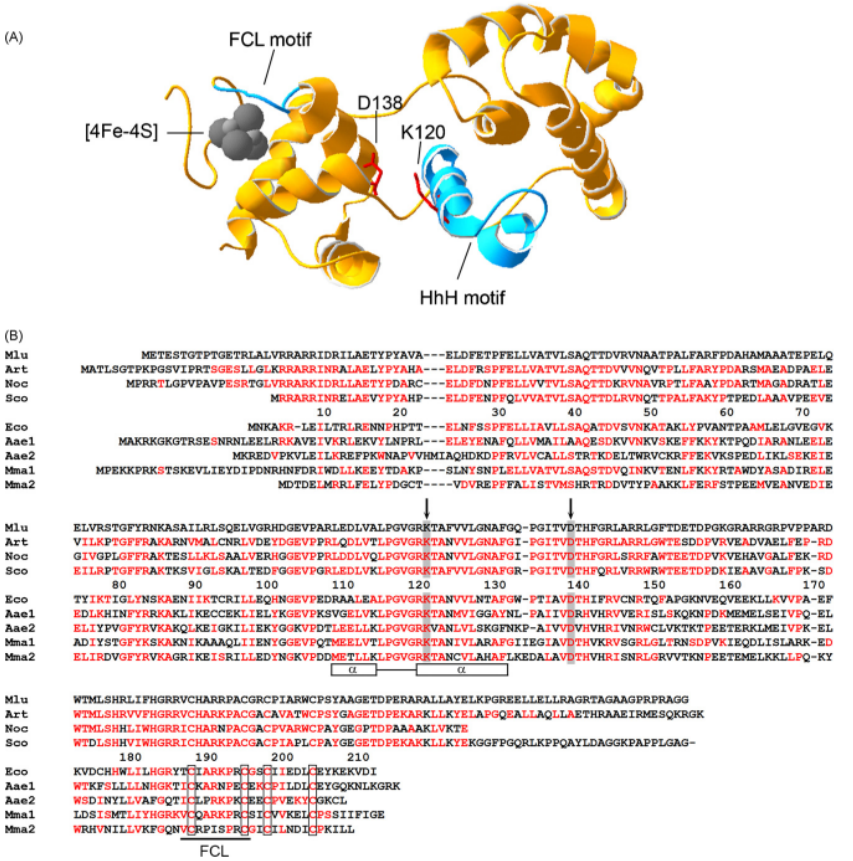


**Fig. 3 – Pyrimidine dimer-DNA glycosylase (31 kDa Mlu-Pdg). (A) Crystal structure of *E. coli* endonuclease III (Nth) (PDB:2ABK). The [4Fe–4S] cluster, the helix-hairpin-helix (HhH) motif and the iron-sulfur cluster loop (FLC) motif are indicated. Residues D138 and K120 which have been shown to be catalytically important are shown in red. (B) Alignment of the 31 kDa Mlu-Pdg protein and the Nth protein from *E. coli* with a selected number of homologous proteins. Residues in the different homologs that are similar to the corresponding residues in Mlu-Pdg are in red. The HhH and FCL regions are indicated. The four iron–sulfur ligating cysteine residues are boxed and the active-site residues are indicated with arrows. Mlu, *Micrococcus luteus*; Art, *Arthrobacter* sp. FB24; Noc, *Nocardioides* sp. JS614, Sco, *Streptomyces coelicolor* A3(2); Eco, *Escherichia coli* K12; Aae, *Aquifex aeolicus* VF5; Mma, *Methanosarcina mazei* Go1.**

source: Goosen et al. 2008

Il a été montré dans Goosen et al. 2008 que deux régions sont très importantes pour la fonction de cette protéine:

- des résidus impliqués dans le motif hélice-épingle à cheveux-hélice [helix-hairpin-helix (HhH)]: L-x-G-V-G-x-K
- des résidus impliqués dans la liaison avec 4Fe–4S: C-x6-C-x2-C-x5-C

Est-ce que ces motifs sont au moins partiellement retrouvés dans les motifs identifiés par prosite?

Ouvez l'alignement avec seaview.

> Est ce que les motifs L-x-G-V-G-x-K et C-x6-C-x2-C-x5-C sont complètement conservés dans toutes les séquences?

Nous allons maintenant étendre la comparaison avec des organismes très eloignés (bactéries, plantes, souris, humain...) afin de voir si ces motifs sont conservés.
Afin de récupérer des séquences homologues réparties dans l'arbre phylogénique du vivant, nous allons réalisé un Blastp en recherchant la séquence de l'endonucléase III d'Escheriachia coli contre la banque des organismes modèles **landmark**. Cette banque inclue 27 protéomes couvrant une large gamme taxonomique.
**Note:** l'utilisation de la banque landmark est possible car la séquence de l'endonucléase III est très conservées entre les espèces.
Sélectionnez une douzaine de séquences représentant une large diversité taxonomique et exportez ces séquences en Fasta [Menu Download puis Fasta (complete sequence)].
Réalisez ensuite un alignement multiple avec Clustal Omega. Vous pouvez ensuite ouvrir l'alignement avec seaview.

> Est-ce que les séquences se ressemblent? Est ce que les motifs L-x-G-V-G-x-K et C-x6-C-x2-C-x5-C sont complètement conservés dans toutes les séquences?