Comparaison de séquences 2 à 2

Liens utiles

- GQuery
- Galaxy, ou Emboss-Explorer lorsque nécéssaire (paramétrage avancé, logiciel indisponible)
- PRSS
- Prosite
- Interpro

Petits exercices

Logiciels de dotplot

Sur le site Galaxy vous disposez d'un certain nombre de logiciels dont nous avons besoin. Répondez aux questions suivantes sur l'utilisation des paramètres de ces logiciels :

- dottup :

  Question 1

  Quel paramètre modifier pour changer la taille des mots servant à la comparaison ? Quelle est la valeur par défaut ? Quelle est la taille minimale ?

- dotmatcher :

  Question 2

  Que signifie le paramètre `Stretch plot` ?

  Question 3

  Quelle est la matrice de scores par défaut ? Sur cette matrice, donnez le scores des appariements et mésappariement des nucléotides A C G T, et expliquez les autres symboles présents.

Exemple de dotplot et interprétation

Voici un séquence d'ADN courte et singulière:

```
tcgcgcgcgcgctgtaagagtgtgtgtggct
```

Tracer (avec dottup) le dotplot de la séquence, en utilisant la taille minimale de mot. Vous réaliserez (dans un premier temps) ce dotplot en comparant la séquence originale contre elle-même.

Question 4

Qu'observez vous sur la diagonale principale ? A votre avis pourquoi ?

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

Qu'observez vous d'autre ? à quoi cela correspond-t-il sur la séquence ?

Réaliser un nouveau dotplot (même parametres) en comparant la séquence originale contre son **complémentaire inversé** (utilisez le programme `revseq` disponible sur Galaxy pour obtenir le complémentaire inversé).

Question 5

Que n'observez vous désormais plus sur la diagonale principale ? A votre avis pourquoi ?

Quel régularité observez vous ? A quoi cela correspond t-il sur la séquence originale ? Justifiez.

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

Alignement manuel

Voici deux séquences artificielles : `AAGTCATTCGCACATCG` et `AACACATCG` et trois alignements possibles à partir de ces séquences :

```
AAGTCATTCGCACATCG
::  :::
AACACATCG
```

```
AAGTCATTCGCACATCG
::  ::: ::
AACACAT-CG
```

```
AAGTCATTCGCACATCG
::       :::::::
AA--------CACATCG
```

Question 6

Calculez le score de chaque alignement avec le jeu de paramètres proposé par défaut par la plupart des programmes d'alignement :

- La matrice de scores :

```
   A  T  C  G
A  5 -4 -4 -4
T -4  5 -4 -4
C -4 -4  5 -4
G -4 -4 -4  5
```

- Ouverture de gap = -10
- Extension de gap = -0,5
- Pas de pénalité pour les gaps en début et en fin d'alignement (alignement semi-global)

Question 7

Quel sera l'alignement retenu par un programme d'alignement avec ces paramètres ?

Question 8

Changer la valeur d'un seul paramètre pour que le premier alignement possède le meilleur score ; puis pour que ce soit le second.

Logiciels d'alignement

Nous allons utiliser aussi deux logiciels d'alignement de séquences : `water` et `stretcher`

Question 9

Lequel des deux fait des alignements locaux ?

Nous ne savons à priori pas si `stretcher` réalise des alignements globaux ou semi-globaux. Pour le trouver, nous choisissons d'aligner la séquence AAGG contre la séquence AG avec ce programme, en fixant un coût d'ouverture de gap de 10 et un coût d'extension de gap de 5.

Question 10

Quel est le résultat (type d'alignement, score) donné par `stretcher` ? Qu'en déduisez vous alors ?

Utilisez également le logiciel `needle` avec les **mêmes parametres** pour valider votre précédente conclusion (Notez que chacun des deux est, soit global, soit semi-global...). Que constatez vous alors ?

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*.

Question 11

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

Dotplots

Question 12

Avec le logiciel `dotpath` réalisez un dotplot. Que voyez-vous ? Combien comptez-vous de diagonales ? A quoi correspondent-elles ?

Question 13

Utilisez maintenant le logiciel dottup afin de déterminer la taille du plus petit exon.

Alignement

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

Question 14

Quel logiciel d'alignement choisissez-vous ? Lancez le logiciel avec les paramètres par défaut ? Retrouvez-vous ce à quoi vous vous attendiez ? Sinon, modifier les pénalités associées aux gaps (eg gap_open:100.0, gap_ext:0.5) jusqu'à obtenir le bon résultat.

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

Question 15

Expliquez les résultats des trois programmes de dotplot sur cette séquence.

Région de faible complexité

Question 16

Observez la séquence contenue dans le fichier falciparum.fasta. Vous devez observer quatre tâches en utilsant dotmatcher, et en faisant varier le score de la fenêtre. A quoi chacune correspond-elle ?

Conservation de domaine

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

Question 17

Construisez un dotplot avec dotmatcher de ces deux séquences.

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

Question 18

Comparez ensuite les deux séquences avec un alignement local en utilisant `matcher`. Retrouver le résultat précédent.

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

Question 19

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

Recherche de domaine

Nous allons comparer 3 enzymes. Le but de l'exercice est de décider sil il existe un ou plusieurs domaines communs à ces trois enzymes.

Question 20

Obtenez les séquences des trois protéines `PDC1_MAIZE`, `ILVB1_TOBAC` et `ILVB_ARATH`.

Question 21

Grâce à des dotplots (à vous de choisir le logiciel le mieux approprié), faites-vous une idée sur les deux séquences les plus proches parmi les trois.

Question 22

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

Question 23

Grâce à la banque de données de domaines Interpro, identifiez les domaines.

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
```

Question 24

Réaliser un alignement local entre les séquences d'ADN. Est-ce que ces séquences se ressemblent ? Quel est le pourcentage d'identité entre les séquences ? L'alignement est-il selon vous significatif ?

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

Question 25

Vous veillerez à prendre une taille de fenêtre de *shuffle* suffisament grande (150), à selectionner la bonne matrice selon la comparaison réalisée (EDNA / Blosum62), et à ne pas coller l'entête fasta avec votre séquence.

Sur combien d'alignements le test a été réalisé ? Combien a-t-on trouvé d'alignements de score 50 ? Combien en attendait-on ? Quelle est le score du meilleur alignement local ? Combien de fois un score meilleur est-il attendu ? Est-ce que l'alignement est significatif ?

Question 26

Réaliser un alignement local entre les séquences protéiques. Est-ce que ces séquences se ressemblent ? Quel est le pourcentage d'identité entre les séquences ?

Question 27

De la même manière que pour l'ADN, estimez la significativité de l'alignement des séquences protéiques. Est-ce que l'alignement est significatif?

Question 28

Comparer les valeurs de significativité trouvées pour l'ADN et les protéines. Quel alignement a la meilleure ? 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 ?

Question 29

Comparer l'alignement obtenu avec les protéines à celui obtenu avec les gènes. Vous allez faire une analyse *fine*, en localisant le domaine sur votre alignement Protéique, puis Nucléique.

Pour vous aider, vous pouvez utilisez un logo du domaine conservé : vous l'obtiendrez à partir d'Interpro (prenez le lien PFxxxx, puis obtenez le *logo HMM* associé).

Y'a t'il des mutations synonymes sur l'ADN ? A quelles positions ? Commenter ...

Page réalisée par les membres des équipes Bonsai et EEP, mise à jour en octobre 2016