Travaux pratiques : Analyse de séquences

1 Présentation

1.1 Objectifs

Cette partie du TP a pour but de vous faire découvrir quelques outils utilisés pour **comprendre** et **décrire** des **molécules d'ADN** ou de **protéines**. En effet, à l'issue du séquençage d'un fragment de chromosome, ou plus largement du génome d'un organisme, les données obtenues ressemblent à ça (ici en version courte ! !) :

```
GGCTCTACTACCTGTTACCCTGGGGTAGTGTCTAGTGACAGCAAAAATAGAATAGTGTCAGTCAT
AGTCAGGTGGCTTTGGAAACTGAAGCTGCCTTCCCAGGACCTGGCCAGTCTCTTGTGGCAGCCAG
GGCTGACAGGCAACTCTTCCTCACAGTCACTGGTCCAGTCACGGAAAGCAGGGATTACTTCCGCT
ATGGCTACACGTACCTCTCTCAAGGATGAAGATCTGGTAAGTGACTCTTGGACATTT
```

Les questions que les biologistes se posent sont alors les suivantes :
— cette séquence contient-elle des gènes ?
— si oui, ces gènes codent-ils pour des protéines, pour des ARNs ?
— les protéines putatives codées par ces gènes sont-elles déjà connues ?

Dans cette première partie du TP, nous vous proposons d'endosser le rôle de chercheur pour **analyser un fragment de chromosome**, séquencé à partir de l'ADN d'une personne atteinte d'**une maladie génétique, la phénylcétonurie**. Cette maladie est caractérisée par un trouble du métabolisme de la phénylalanine. Il semblerait que ce soit une mutation dans le fragment d'ADN que nous allons étudier qui est responsable de la maladie.

1.2 Notions à maîtriser à la fin de cette partie

— structure des gènes eucaryotes ;
— sensibilité d'un programme de prédiction ;
— précision d'un programme de prédiction ;
— alignement de séquences ;
— Blast ;
— E-value ;
— homologie, paralogie, orthologie ;
— réseau métabolique ;
— polymorphisme ;
— navigateur de génomes ;
— logiciel et base de données.

Si certaines notions ne sont pas claires pour vous à la fin des 2 séances, n'hésitez pas à interroger votre enseignant !

1.3 Méthodes

Nous vous conseillons de créer un document et d'y enregistrer, tout au long de ces 2 séances, les adresses des sites web consultés ainsi que les résultats obtenus. Cela vous aidera à mémoriser les différentes analyses et à rédiger le compte-rendu.

La plupart des ressources informatiques (logiciels et bases de données) que vous allez utiliser sont disponibles dans `Liens Web` sur Claroline, mais nous vous conseillons fortement de faire comme les chercheurs : recherchez le programme ou la base de données à l'aide de votre moteur de recherche préféré (Google, DuckDuckGo...), le premier lien est très souvent le bon !

2 Rappel

Pour partir sur de bonnes bases...
→ *Représentez en un schéma bilan le dogme central de la biologie moléculaire (= "1 gène donne 1 ARN qui donne 1 protéine") pour un gène à 3 exons, sur lequel vous indiquerez les éléments suivants : ADN, préARNm, ARNm mature, protéine, gène, exon, intron, 5'UTR, 3'UTR, site d'initiation de la transcription, promoteur, épissage, transcription, traduction.*

3 Prédiction de gènes

Vous allez d'abord prédire la présence éventuelle de gènes dans la séquence étudiée et, si vous en trouvez, caractériser leur structure exons-introns. On distingue deux types d'approches pour prédire des gènes :

- les approches comparatives : elles prédisent des gènes dans une séquence en cherchant des similarités avec des gènes déjà connus (cf le programme "*Blast*" plus loin) ;
- les approches *ab initio* : elles ne comparent pas la séquence à des gènes connus, mais recherchent des éléments caractéristiques des gènes : cadre ouvert de lecture (=ORF), séquence promotrice consensus, séquences d'épissage... Cela nécessite de bien connaître ces éléments *a priori*.

Vous allez utiliser ici deux logiciels de prédiction *ab initio* : *Genscan* et *Augustus*. Comme il s'agit de prédictions, nous ne sommes jamais sûrs à 100% du résultat ; utiliser plusieurs programmes permet de croiser leurs résultats et d'identifier les prédictions auxquelles on peut faire le plus confiance.

La séquence à analyser est sur Claroline, dans le dossier : `TP Analyse de séquence`.

Enregistrez le fichier "`sequence.fasta`" dans votre répertoire de travail. Vous pouvez l'ouvrir grâce à un éditeur de texte (Wordpad, Notepad...).

Ce fichier est au format "fasta", qui est un format de fichier de séquences biologiques. Il se présente comme suit :

```
> nom de la sequence 1
ATTGCC... (séquence 1)
> nom de la séquence 2
GCGTTA... (séquence 2)
etc
```

3.1 Précision des programmes

Les programmes de prédiction sont caractérisés par leur **précision** et par leur **sensibilité**.

→ *A partir des notions de vrai/faux positif et vrai/faux négatif, rappelez les définitions de la "précision" et de la "sensibilité".*

Pour évaluer ces paramètres, il faut **tester les programmes** sur un **jeu de données connu** (ici, une séquence dans laquelle on sait exactement combien il y a de gènes et où ils se trouvent). C'est le cas des séquences ENCODE : il s'agit d'un projet de recherche international qui a pour but de décrire très précisément et expérimentalement une portion de notre ADN. On connaît donc, dans cette séquence, les gènes réels. Augustus et Genscan ont été testés sur ces données ; les résultats sont présentés ici : `Page d'accueil d'Augustus -> accuracy results`.

→ *Cherchez sur cette page le tableau "Accuracy results on human ENCODE regions (ab initio)". Quelle est la précision des 2 programmes en termes de prédiction d'exons ? Quelle est leur sensibilité ?*
Attention : le terme de spécificité est utilisé à tort dans le tableau, et devrait être remplacé par le terme de précision.

3.2 Utilisation de Genscan

Recherchez Genscan avec le moteur de recherche de votre choix et faites une prédiction de gènes dans la séquence `sequence.fasta`.
Attention : En entrée, Genscan ne prend pas de format fasta, mais uniquement la séquence brute, sans la ligne correspondant à son nom. Vous pouvez sélectionner et coller la séquence dans le formulaire de soumission. Les résultats apparaissent sous forme d'un tableau (sous la ligne "Predicted genes/exons").

Aidez-vous du document `GENSCAN_outputExplanation.pdf`, qui se trouve sur Claroline, pour comprendre les différentes lignes et colonnes de ce tableau.

→ *Combien de gènes Genscan prédit-il ?*
→ *Combien d'exons pour ce(s) gène(s) ?*
→ *Sur quel brin de l'ADN est-il / sont-ils situé(s) ?*
→ *Quelle est la longueur totale du transcrit (non mature) prédit ?*

La **probabilité** (colonne 'P....') calculée par Genscan pour chaque exon potentiel est la probabilité que cet exon soit correct, d'après le modèle utilisé par Genscan. Elle dépend à la fois de propriétés locales de la séquence, et de son environnement (correspondance de l'exon avec les exons voisins).

→ *D'après ces valeurs de probabilité, quels sont les exons qui ont le plus de chance d'être réels ? Lesquels au contraire constituent peut-être des faux-positifs ?*

Sauvegardez la séquence de la protéine prédite.

3.3 Utilisation d'Augustus

Soumettez la séquence au programme (onglet 'web interface'), en utilisant les paramètres par défaut.

La sortie est dans un format appelé "**gff**" : les lignes précédées d'un "#" sont des commentaires ; les autres correspondent chacune à un élément prédit. Les différentes informations sont séparées par des tabulations ; pour y voir plus clair, vous pouvez copier-coller ces lignes dans un tableur (LibreOfficeCalc par exemple, collage spécial "Text Unicode"), en choisissant "tabulations" comme séparateur de champs.

Voici le **contenu des différentes colonnes** :
1. Nom de la séquence ; 2. Programme ; 3. Type d'élément prédit (gène, transcrit...) ; 4. Début de l'élément dans la séquence ; 5. Fin de l'élément dans la séquence ; 6. Probabilité de l'élément ; 7. Brin ; 8. Phase par rapport au début du gène auquel l'élément appartient ; 9. Identifiant de l'élément.

Parmi les différents types d'éléments prédits par Augustus, on trouve des exons ('initial', 'internal' ou 'terminal'), ainsi que des "CDS" (= Coding Sequence), qui correspondent uniquement aux parties codantes des exons (ils ne comprennent donc par les UTRs, mais sont compris dans les exons).

En analysant vos résultats, répondez aux questions suivantes :
→ *Combien de gènes Augustus prédit-il ?*
→ *Combien de transcrits ?*
→ *Combien d'exons pour chacun de ces transcrits ?*
→ *Quelle est la longueur du transcrit t1 ?*
→ *Quels sont les exons les plus sûrs ?*

Sauvegardez les séquences des protéines prédites.

Vous pouvez visualiser les résultats en suivant le lien `graphical and text results -> graphical browsable results`.

3.4 Comparaison des résultats des deux programmes

→ *Comparez les résultats produits par les deux logiciels : sont-ils globalement compatibles ? Identifiez un exon prédit par l'une des deux méthodes uniquement. Présente-t-il une bonne probabilité ?*

Remarque : Les probabilités de chaque exon sont calculées différemment par les deux programmes ; vous pouvez donc les comparer pour différentes prédictions au sein d'un même programme, mais pas entre les deux.

4 Informations apportées par la recherche de protéines similaires

4.1 Recherche de protéines similaires à l'aide du programme Blast

Nous allons maintenant regarder si le gène étudié n'est pas déjà connu. Pour cela, nous allons comparer la séquence de la protéine prédite aux séquences de

protéines déjà connues, qui sont répertoriées dans des bases de données accessibles en ligne. C'est une **méthode très utilisée en bioinformatique** : si une séquence de protéine inconnue ressemble fortement à la séquence d'une protéine dont on connaît la fonction, on peut faire une hypothèse sur la fonction de la protéine inconnue !

Cette démarche repose sur l'**alignement de séquences**.

Il existe différentes bases de données de séquences ; UniProt/Swissprot et UniProt/TrEMBL en sont deux exemples. Leurs caractéristiques sont décrites sur la page d'accueil d'Uniprot (lien `Statistics`).

→ *Combien ces 2 bases contiennent-elles de séquences ('entries' en anglais) ? Quelle est la proportion de séquences pour lesquelles on a des "évidences" expérimentales ? et la proportion de séquences simplement prédites ? Qu'en pensez-vous ?*

Recherchez le formulaire de soumission de Blast sur le site du NCBI. Parmi les 5 versions de Blast possibles, choisissez celle qui permet de rechercher une séquence protéique dans une banque de protéines.

Utilisez le formulaire pour rechercher des protéines similaires au produit du transcrit 1 d'Augustus, dans la banque de données SwissProt (`Choose Search Set -> Database`).

Une description de Blast a été déposée sur Claroline (`"Cours_Blast_AMII.pdf"`). Consultez :

— la diapo 20 pour bien comprendre le formulaire de soumission de Blast ;
— les diapos 21 à 25 pour comprendre la page de résultats.

→ *Quelle est la protéine la plus similaire à la protéine prédite ? A quelle espèce appartient-elle ?*
→ *Quel est son score d'alignement ? Comment ce score est-il calculé ? (Relisez votre cours ! vous pouvez aussi vous aider des diapos 27 et 28 du pdf).*
→ *Quelle est la E-value de cet alignement ? Quelle est la signification de la E-value ? L'alignement observé est-il significatif, ou aurait-il pu être obtenu par hasard (c'est-à-dire sans que les séquences soient homologues) ?*
→ *Observez-vous des différences entre la protéine que vous avez soumise à Blast (issue de la prédiction de gènes), et la protéine la plus similaire trouvée dans Swissprot (et donc réelle) ? Comment les expliquez-vous ?*

Pour ce meilleur hit, cliquez sur le lien dans la colonne 'Accession' (ou sur le lien qui suit 'Sequence ID' au-dessus de l'alignement) puis sur 'FASTA'. Enregistrez cette séquence ; vous en aurez besoin pour l'étude du polymorphisme (Partie 5).

Refaites l'alignement avec la prédiction de Genscan.

→ *Là-encore, observez-vous des différences entre la protéine que vous avez soumise à Blast (issue de la prédiction de gènes), et la protéine la plus similaire trouvée dans Swissprot (et donc réelle) ? Ces observations vous semblent-elles cohérentes avec les paramètres globaux de sensibilité/précision des deux programmes ?*

4.2 Exploration de la fiche UniProt de la protéine trouvée par similarité

UniProt/SwissProt recense toutes les informations connues sur les protéines. Nous allons ici chercher des infos sur la protéine la plus similaire à celle qui a été prédite, afin de pouvoir faire des hypothèses sur la fonction de la protéine prédite elle-même.

Recherchez le gène d'intérêt dans la base de données UniProt/SwissProt. A partir de sa fiche, répondez aux questions suivantes :

→ *A-t-on une preuve expérimentale de l'existence de cette protéine ?*
→ *Quelle est sa fonction ?*
→ *Quel est le principal processus biologique dans lequel elle est impliquée ?*
→ *Où la protéine est-elle localisée dans la cellule ?*
→ *La protéine subit-elle des modifications post-traductionnelles ("PTM") ?*
→ *Combien sa structure comprend-elle de feuillets Beta ?*
→ *La protéine possède-t-elle des domaines particuliers ?*
→ *Combien de références bibliographiques concernant cette protéine existe-t-il ?*

Approfondissement :
→ ***A votre avis, quels sont les avantages d'une approche ab initio ? Ses limites ?***
→ ***De la même manière, quels sont les avantages de l'approche comparative ? Ses limites ?***

4.3 Etude de la famille de gènes de PAH

Une **famille génique** est un ensemble de gènes possédant la même origine, et présentant souvent de fortes ressemblances fonctionnelles et structurales. Ces gènes peuvent être dans la même espèce ou dans des espèces différentes.

→ *Redéfinir les notions d'homologue, orthologue, paralogue.*

On décide en général que deux séquences ont la même origine **à partir de leur alignement** : s'il est très bon, il est très probable que ces deux séquences aient une séquence ancestrale commune. Attention cependant, il existe toujours quelques cas particuliers : certaines séquences similaires ne sont pas orthologues (on parle alors d'homoplasie, ou convergence évolutive), et certaines protéines homologues ne sont plus très similaires (notamment si leur dernier ancêtre commun commence à dater...).

En pratique, les biologistes décident que 2 séquences sont homologues si leur **query coverage** est **supérieure à 50%** et leur **E-value inférieure à** 10^{-5}.

→ *Refaites un Blast contre SwissProt avec la protéine PAH enregistrée précédemment (pas la protéine prédite). En suivant ces critères, dénombrez le nombre d'homologues de cette protéine PAH dans la base de données Swissprot.*
→ *Combien possède-t-elle d'homologues chez l'homme ? (faites le même Blast contre SwissProt en précisant "human" dans la fenêtre "Organism")*

`Ensembl` est une base de données recensant des informations sur les génomes de différents organismes ainsi que sur les gènes qu'ils contiennent. Le **navigateur de génomes** d'`Ensembl` permet de visualiser l'environnement génomique d'un gène donné (notamment les gènes et séquences régulatrices voisins) ou de tout un fragment de chromosome.

Rendez-vous sur le site d'`Ensembl`, et sélectionnez le gène PAH de l'homme. En utilisant le menu à gauche "Comparative Genomics" répondez aux questions suivantes :

→ *Combien y a-t-il de paralogues répertoriés pour ce gène ? Retrouvez-vous le chiffre obtenu d'après le résultat du Blast ?*
→ *Combien d'orthologues sont-ils répertoriés dans Ensembl ?*

Visualisez l'**arbre de gènes** associé à cette famille de gènes. Il représente sous forme graphique le degré de similarité entre les différents gènes de la famille. Le nombre total de gènes dans l'arbre (indiqué au-dessus du graphique) donne le nombre total d'homologues répertoriés dans `Ensembl`.
Remarque : Vous remarquez que ce nombre est bien plus important que la simple somme des paralogues et des orthologues... Cela est lié au fait qu'`Ensembl` adopte une définition restrictive pour les orthologues d'un gène.

4.4 Analyse des régions conservées du gène

L'alignement de séquences est utilisé pour annoter un gène par similarité, mais aussi pour détecter des **régions conservées** entre espèces, indicatrices de sites fonctionnels. En effet, si une séquence est conservée sur de longues périodes évolutives, on peut faire l'hypothèse qu'il existe une **pression de sélection** qui la contraint.

Nous allons analyser la conservation du gène étudié à partir d'alignements multiples précalculés, accessibles sur le site `Vista` : `http://genome.lbl.gov/vista/index.shtml`. Cliquez sur VISTA-point. Sélectionnez la version février 2009 du génome humain, et la position PAH. Soumettez votre requête ; parmi les différentes possibilités qui sont proposées, choisissez celle correspondant au gène "RefSeq".

La courbe qui s'affiche a été calculée à partir d'un alignement entre le gène PAH de l'homme et celui de la souris. Elle correspond au degré de similarité local entre ces deux séquences. Les pics (notamment ceux qui passent la droite centrale) sont les régions les plus conservées.

Cliquez sur `"Alignment in PDF"`, qui affiche un pdf du graphique avec la légende des couleurs.

→ *A quoi correspondent les régions conservées homme/souris ?*

CNS = 'Conserved Non-coding Sequence', ou région conservée non-codante. Ces régions peuvent avoir une fonction régulatrice (fixation de facteurs de transcription, régulation de l'épissage...).

Affichez maintenant les courbes de comparaison avec le chimpanzé et avec la vache.

→ *Comment expliquez-vous le profil du chimpanzé ? Faites une hypothèse pour expliquer les zones "blanches", où la similarité chute brusquement.*

Approfondissement :
On peut également faire des alignement multiples avec des protéines. Il est possible d'obtenir l'alignement de séquences protéiques homologues à la protéine d'intérêt provenant d'autres espèces à partir du fichier "seq_famille_" (`Claroline`), en se servant de l'outil en ligne d'alignement multiple ClustalW (Claroline, Liens webs ; utiliser les options par défaut).
→ ***Trouvez des exemples de résidus conservés entre toutes les séquences. A quelle fonction particulière au sein de la protéine ces résidus pourraient-ils être liés ?***
→ ***Est-il possible de trouver des signatures dans l'alignement (substitution, insertion/délétion) partagées par certaines espèces seulement ? Cela a-t-il un rapport avec leur place relative dans l'arbre de la vie ?***

5 Etude du polymorphisme du gène

Le polymorphisme désigne la coexistence de plusieurs variants d'un même gène (et non pas de gènes paralogues) au sein d'une espèce ; ces variants sont appelés **allèles**. Les données de polymorphisme sont intéressantes car certains allèles peuvent être délétères ou conférer des prédispositions à certaines maladies.

La base de données de référence concernant le polymorphisme chez l'homme est la base `dbSNP` ; les données de cette base sont également accessibles sur le site `Ensembl`.

→ *Qu'est-ce qu'un SNP ?*

Recherchez à nouveau le gène PAH humain dans `Ensembl`. Dans le menu à gauche "Genetic Variation", affichez "Variation Table".

En haut à gauche de la table se trouve un bouton permettant de filtrer les résultats par type de variants. → *À l'aide du filtre, sélectionner les variants synonymes. Combien y en a-t-il ?*

→ *Citez un exemple de type de variant ayant un impact fonctionnel probablement important, et un exemple de variant sans impact (probable).*

Les SNPs peuvent également servir à diagnostiquer rapidement une maladie chez un individu, et à faire le lien structure-fonction entre un phénotype et une mutation.

En utilisant Blast avec l'option "Align two or more sequences", alignez la séquence saine de la protéine PAH (celle que vous avez enregistrée dans la partie 4.1), avec la séquence d'un individu malade (`Claroline, seq_variant_to_diagnostic`).

→ *Observez-vous une différence entre les deux séquences ? A quelle position ? De quelle nature ?*
→ *En vous aidant de la fiche d'annotation SwissProt/Uniprot, qui contient des informations sur les variants connus, trouvez la conséquence de ces changements sur la fonction de la protéine, notamment au niveau de sa structure.*

6 Etude du réseau métabolique de PAH

Le réseau métabolique est l'**ensemble des voies biochimiques** de synthèse/dégradation dans lesquelles une protéine est impliquée. Nous allons nous intéresser dans cette partie aux données que l'on peut trouver concernant le réseau métabolique de la protéine PAH.

Les enzymes sont caractérisées par un "numéro EC".
→ *Que signifie ce numéro ?*

Recherchez le numéro EC de l'enzyme PAH sur sa fiche SwissProt.

`KEGG` est une base de données recensant toutes les réactions métaboliques identifiées dans le Vivant, ainsi que les enzymes qui les catalysent.

Recherchez sur le site `KEGG` le réseau métabolique de la phénylalanine. Affichez le graphique de ce réseau, pour l'homme.

Ce schéma représente toutes les réactions potentielles impliquant la phénylalanine ou l'un de ses dérivés ; ces réactions ont pu être identifiées dans différentes espèces (le schéma est donc théorique). Les cercles représentent des métabolites, et les flèches des réactions chimiques. Chaque réaction chimique est catalysée par une ou plusieurs enzymes, indiquées dans les rectangles par leur numéro EC. Les enzymes connues chez l'homme sont indiquées en vert. Les enzymes qui ne sont pas en vert sont soit effectivement absentes chez l'homme, soit pas encore caractérisées dans cette espèce.

→ *Combien de réactions ont pour substrat la phénylalanine ?*
→ *A votre avis, pourquoi utilise-t-on la présence de phénylpyruvate dans les urines comme diagnostic de la phénylcétonurie ?*
→ *Pourquoi les individus atteints de phénylcétonurie ont-ils le teint pâle ? (un indice pour cette question pas évidente : la couleur de la peau est associée à la production de mélanine)*

Approfondissement :
Comment les cartes KEGG sont-elles établies ? En particulier, comment décide-t-on qu'une enzyme est présente ou absente chez un organisme ? Peut-on considérer que la carte KEGG du cheval est complète ? Pourquoi ?

7 Expression du gène PAH

Des données d'expression sont également disponibles en ligne. Elles permettent de voir dans quels tissus et à quels stades du développement les gènes

s'expriment. Là encore, cela permet de faire des hypothèses sur leur fonction.

→ *A l'aide du site BioGPS, qui recense de telles données d'expression, déterminez dans quels tissus/organes est exprimé le gène étudié.*

L'onglet `Correlation` donne accès à un formulaire de recherche de gènes présentant des profils 'corrélés', c'est-à-dire similaires.

→ *Quels autres gènes ont un profil d'expression similaire à celui de PAH ? Soumettez à nouveau ces gènes à BioGPS pour voir leur profil et vérifier que c'est bien le cas.*
→ *Quelle hypothèse pouvez vous faire sur des gènes qui ont des patrons d'expression similaires ?*