Travaux pratiques de bioinformatique

1 Introduction

Notion abordées : sensibilité, specificité, alignement de séquences, blast, E-value, structure d'un gne eucaryote, homologie, orthologie, paralogie, réseau métabolique, polymorphisme, navigateur de génomes.

1.1 Objectif et notation

Ce premier TP a pour objectif d'étudier certaines caractéristiques d'une séquence résultant du séquenage d'un fragment d'ADN génomique **humain**, qu'on sait impliqué dans une maladie génétique, la phénylctonurie. À la fin de ce TP, vous devrez rendre un compte-rendu rapportant les différentes manipulations que vous avez effectuées. Pour chaque partie du TP, vous devrez présenter dans votre compte-rendu : 1- les méthodes utilises 2- les résultats obtenus et 3- la conclusion biologique (qu'a-t-on appris sur la séquence analysée) et méthodologique (quelles sont les limitations de la méthode employée) que vous en tirez. Tout au long de l'étude de cette séquence, enregistrez donc tous les résultats obtenus (figures, pages web, documents, ...) vous paraissant pertinents pour les inclure par la suite dans votre compte-rendu.

Attention : il n'est pas demandé ici de détailler les réponses aux questions posées dans cet intitulé (qui sont plus un guide pour le déroulement du TP), mais bien de faire un rapport global sur les méthodes mises en oeuvre, les connaissances aquises et votre regard critique sur celles-ci.

Enfin, ce TP propose des approfondissements. Vous en choisirez un seul, pour lequel vous aurez faire quelques recherches. Vous incluerez le résultat de vos recherches dans votre compte-rendu. Ce compte-rendu *ne doit pas* dépasser 6 pages. Celui-ci sera produit avec le traitement de texte de votre choix – les comptes-rendus manuels sont *interdits*. Vous le convertirez impérativement au format PDF, puis vous le enverrez par courrier électronique. Vous nommerez impérativement votre fichier par vos nom de famille.

Vous pouvez donc créer un répertoire de travail dans lequel vous enregistrerez tous ces documents (en veillant à choisir des noms qui vous permettent de les retrouver facilement ensuite) : créez ce répertoire dans votre dossier personnel (accessible via le menu Raccourcis). Vous pouvez également prendre des notes au fur et mesure du TP dans un éditeur de texte pour vous éviter un travail de copie fastidieux par la suite.

La séquence que vous allez étudier au cours de ce TP se trouve sur PBIL. Pour y accéder, vous pouvez utiliser un navigateur web. Le fichier `sequence.fas` se trouve dans :

`ftp://pbil.univ-lyon1.fr/pub/cours/coimbra/TPAnalyseSequence5BB/`

Il comprend cette séquence au format FASTA (*un format standard de représentation d'une séquence en bioinformatique*). Enregistrez ce fichier dans votre répertoire de travail.

La plupart des ressources informatiques (logiciel et bases de donnes) que vous allez utiliser sont disponibles dans le fichier `LiensWeb.pdf`.

2 Prediction de gènes

Au cours de cette première partie, vous allez caractériser la structure exon-intron du gène étudié. Pour cela, vous allez utiliser deux logiciels : Genscan et Augustus.

Genscan est un programme de détection de gènes ab initio (seule la squence d'ADN est fournie en entre). Augustus permet quant à lui d'intégrer des données supplémentaires (ARN, introns...), mais nous allons l'utiliser comme un programme ab initio.

À l'aide de leur documentation, expliquez brièvement comment fonctionnent ces programmes.

Discutez de la confiance que vous pouvez accorder dans les prédictions de ces programmes. Pour cela, vous pouvez vous aider du tableau récapitulatif "Accuracy results on human ENCODE regions (ab initio)" de la documentation d'Augustus.

2.1 Utilisation de Genscan

Note : En entre, Genscan ne prend pas de format fasta, mais la séquence brute (i.e. retirer la ligne commenant par $>$)

Combien d'exons sont prédits par Genscan ? Quelle confiance pouvez-vous accorder dans ces prédictions ? Sauvegardez la séquence de la protéine prédite.

2.2 Utilisation de Augustus

Combien d'exons sont prédits par Augustus ? Combien de transcrits sont prédits par Augustus ? Quelle confiance pouvez-vous accorder à chacune de ces prédictions ? Sauvegardez la (ou les) séquence(s) de la (ou les) protéine(s) prédite(s). Qu'est-ce que le format fasta ? Qu'est-ce que le format gff ?

Vous pouvez visualiser et naviguer les résultats en utilisant les outils en ligne sur le site d'Augustus.

2.3 Comparaison des résultats

Comparez les résultats produits par les deux logiciels. Sont-ils compatibles ? (Quels sont les exons communément trouvés par les deux méthodes et ceux uniques à une méthode ?)

2.4 Schéma bilan

Pour récapituler les résultats obtenus, proposez un schéma bilan comprenant votre squence d'ADN, de pre-ARNm, ARNm mature et protéine.

3 Alignement de séquences

On va maintenant vérifier si le gène qu'on étudie n'est pas déjà connu. Nous allons donc comparer la séquence de notre protéine prédite aux séquences des protéines déjà connues.

On note que cette démarche d'alignement de séquences est très utilisé en bioinformatique et permet de remplir de multiples objectifs :

1. Identification : vérifier si la séquence qu'on étudie n'est pas déjà connue ;

2. Annotation par similarité : si la séquence étudiée n'est pas connue, est-elle similaire des séquences connues ?
3. Recherche de régions conservées : parmi toutes les séquences similaires à la nôtre, existe-t-il des régions conservées ?

Quelle est la différence entre Swissprot et TrEMBL ? Aidez-vous des statistiques descriptives de ces deux banques de donneés.

Alignez votre protéine prédite (produit du transcrit 1 de Augustus) en utilisant BLAST contre toutes les séquences de la base de données Swissprot.

Quelle est la protéine la plus similaire à votre protéine prédite ? À quoi correspond le score d'alignement ? Ce score est-il significativement élevé ?

Refaites l'alignement avec la prédiction de Genscan, et avec le transcrit 2 de Augustus. L'alignement est-il sur toute la longueur de la protéine ? Pourquoi ?

Recommencez l'alignement en utilisant cette fois UNIPROT. Qu'observez-vous ?

3.1 Exploration de la fiche Swissprot

1. Quelle est la fonction de la protéine identifiée ?
2. A-t-on une preuve expérimentale de l'existence de cette protéine ?
3. Dans quelle processus biologique est-elle impliquée ?
4. Quelle est sa structure ?
5. Combien de variants sont connus ?
6. Combien de références bibliographiques sont liées à cette protéine ?
7. La protéine subit-elle des modifications post-traductionnelles ?
8. Où est localisée la protéine dans la cellule ?
9. De quel organisme provient cette séquence ?

Approfondissement : Comment ce gène a-t-il t annoté ? Quel est l'avantage de l'approche ab initio ? Quel est l'avantage de l'approche comparative ?

3.2 Famille de gènes

Définir les notions d'homologue, orthologue, paralogue. Comment déterminer que deux protéines sont homologues ? Combien d'homologues possède la protéine sur laquelle vous travaillez ?

Rendez-vous sur le site d'Ensembl, et sélectionnez le gène PAH. En utilisant le menu à gauche "Comparative Genomics" répondez aux questions suivantes : Combien d'homologues sont-ils répertoriés pour ce gène ? Combien de paralogues ? Combien d'orthologues ? Visualisez l'arbre de gène associée à cette famille de gènes.

Dans l'interface web de Ensembl, sélectionnez "Orthologs" dans le menu à gauche de la fiche du gène PAH. Trouvez le gène orthologue chez la souris (*Mus musculus*). Vous pouvez visualiser les

deux gènes dans leur contexte chromosomique respectif en cliquent sur “Region Comparison”. Que remarquez-vous au voisinage des gènes PAH/Pah ?

Dans le menu à gauche de la page “Region Comparison”, vous pouvez choisir de visualiser les synténies, c’est à dire les régions chromosomiques conservées entre les différentes espèces. Quels sont les chromosomes de la souris qui partagent des régions d’homologie avec le chromosome 12 de l’homme ? Quel chromosome murin porte la région synténique avec le locus PAH humain ?

Dans le même menu à gauche vous avez accès aux paralogues du gène. Quels sont-ils ? Quelles sont les fonctions de leurs produits (protéines correspondantes) ?

3.3 Recherche de régions conservées

L’alignement de séquences est utilisé pour annoter un gène par similarité, mais aussi pour détecter des régions conserves entre espèces, indicatrices de sites fonctionnels. En effet, si une séquence est conservée au cours de l’évolution, c’est parce qu’il y a une pression de sélection qui la contraint.

Nous allons pour cette étape utiliser des alignements précalculs. Ce sont des alignements multiples. `http://genome.lbl.gov/vista/index.shtml` Cliquez sur “Vista Browser”. Sélectionnez la version février 2009 du génome humain, et la position PAH. Quelles sont les régions conservées ?

Approfondissement : On peut également faire des alignement multiples avec des protéines. Il est possible d’obtenir l’alignement de séquences protéiques homologues notre protéine provenant d’autres espèces partir du fichier “seq_famille.fas” en se servant de l’outil en ligne d’alignement multiple ClustalW (LiensWebs.pdf ; utiliser les options par défaut). Vous pouvez observer l’alignement sur la page ou grâce à l’outil Jalview. Trouvez des exemples de résidus conservés. Ces résidus sont-ils lis une fonction particuliére au sein de la protéine ? Est-il possible de trouver des signatures dans l’alignement (substitution, insertion/deletion) qui caractérise des groupes d’organismes ? Cela a-t-il un rapport avec leur place dans l’arbre de la vie ?

4 Polymorphisme

Dans cette partie, vous allez utiliser un navigateur de génomes. C’est un outil communément employé en bioinformatique pour visualiser un gène dans son contexte génomique. On peut en outre afficher des informations complémentaires. On s’intéresse ici aux données de polymorphisme, issues de la base de données dbSNP.

Allez sur le site de l’UCSC. Affichez le gène PAH. Sur quel chromosome se situe-t-il ? Quels sont les gènes qui sont voisins ?

Affichez la piste SNP. Qu’est-ce qu’un SNP ? Quels sont les significations des couleurs associées aux SNPs ? Un SNP modifie-t-il toujours la séquence de la protéine ? Un SNP a-t-il toujours un impact fonctionnel ?

Les SNPs peuvent également servir à diagnostiquer rapidement un individu, et à faire le lien structure-fonction entre un phénotype et une mutation. En utilisant BLAST, alignez une séquence protéique d’un individu malade (“seq_variant_to_diagnostic.fas”) avec sa contrepartie saine présente dans les bases de données. Trouvez-vous des changements dans la séquence protéique ? En

vous aidant de la fiche d'annotation SwissProt contenant des informations sur les variants connus, pouvez-vous trouver la conséquence des changements sur la fonction de la protéine, notamment au niveau structurel ?

5 Réseau métabolique

À partir de la fiche Swissprot, notez le numéro EC du gène PAH. Que signifie ce numéro ?
À partir du site KEGG, recherchez la voie de dégradation de la phénylalanine. Affichez la voie métabolique chez l'homme. Combien de réactions ont pour substrat la phénylalanine ?

Pourquoi utilise-t-on la présence de phénylpyruvate dans les urines comme diagnostic de la phénylcétonurie ?

Pourquoi les individus atteints de phénylcétonurie ont le teint pâle ?

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

6 Expression du gène

À l'aide du site BioGPS, déterminer dans quels tissus est exprimé le gène étudié. Quels autres gènes ont le même profil d'expression ?

Approfondissement : Vous pouvez voir dans BioGPS que l'expression du gène PAH dépend beaucoup du tissu que l'on considère. En réalité, l'expression d'un gène dépend de très nombreux facteurs. Une des techniques maintenant classiques pour étudier l'expression d'un gène dans de nombreuses conditions différentes est l'usage de microarrays, ou "puce à ADN" en français. Les résultats de ce type d'étude sont disponibles sur le site `http://www.ebi.ac.uk/gxa/`. Après avoir réalisé une recherche pour un gène d'intérêt, ce site donne un résumé des expériences de mesure d'expression réalisées sur le gène : on peut avoir accès au fait que le gène soit surexprimé (code couleur rouge) ou sous-exprimé (code couleur bleu) dans différents tissus, en fonction de différentes maladies affectant le patient, ou encore dans différents types cellulaires.

1. **Expliquez de manière concise le principe de fonctionnement d'une puce à ADN : que permet-elle de mesurer ? Les mesures obtenues sont-elles des valeurs d'expression absolues ou bien des comparaisons ? Quel avantage voyez-vous à ce type d'expérience par rapport à des PCR quantitatives ?**
2. **En étudiant les profils d'expression du gène PAH dans la base de données ArrayExpress, que pouvez-vous dire sur l'expression de ce gène ? Est-elle très stable dans les différents tissus ? Les différentes conditions expérimentales ? Globalement, concluriez-vous que le niveau d'expression d'un gène est plutôt très variable ou très stable ?**