Tri de données – TP de Bioinformatique
Partie 1 – A l'échelle du gène

Licence professionnelle Génomique – automne 2015

Résumé

Ce premier TP a pour objectif d'étudier certaines caractéristiques d'une séquence résultant du séquençage 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 utilisées 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, ...) PERTINENTS pour les inclure par la suite dans votre compte-rendu. Le compte-rendu ne devra pas dépasser si pages, sans annexes.
Attention : il n'est pas demandé ici de simplement 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. N'hésitez donc pas à proposer vos propres démarches et perspectives liées aux questions.
Les directives deviennent volontairement de moins en moins strictes tout au long du TP, afin de vous permettre de devenir autonome vis à vis de l'utilisation des différents logiciels.
Vous pouvez 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). 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.

Documents nécessaires :
Tous les documents dont vous aurez besoin sont disponibles sur mon GitHub, un service web d'hébergement et de gestion de développement de logiciels, qui se veut aussi utile pour le partage de fichiers. Pour y accéder, taper l'adresse dans votre navigateur web : `https://github.com/tdelhomme`. Dans l'onglet "Popular Repository", cliquer sur "TP_bioinfo_L3_genomique". Les fichiers disponibles sont listés au centre de la page, vous pouvez les télécharger un par un en cliquant droit sur le lien du fichier, puis en cliquant sur "Save link as...", mais aussi les télecharger sous forme de zip (dossier compressé contenant tous les fichier), en cliquant sur "Download ZIP sur la droite de votre écran.

Exercice 1 : Prédiction 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 entrée). 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.

1. Utilisation de GenScan

Vous trouverez le programme à cette adresse : `http://genes.mit.edu/GENSCAN.html`.
En entrée, Genscan ne prend pas de format fasta, mais la séquence brute (*i.e.* retirer la ligne commençant par $>$).
Combien d'exons sont prédits par Genscan ? Quelle confiance pouvez-vous accorder dans ces prédictions ? Sauvegardez la séquence de la protéine prédite.

2. Utilisation d'Augustus

Vous trouverez le programme à cette adresse : `http://bioinf.uni-greifswald.de/augustus/submission.php`.
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 les résultats et naviguer entre eux en utilisant les outils en ligne sur le site d'Augustus.

3. Comparaison des résultats

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

4. Schéma bilan

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

Exercice 2 : 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 ?

1. Alignement

Qu'est-ce-que Swiss-Prot et TrEMBL ?
Quelle est la différence entre Swiss-Prot 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'aligne ment est-il sur toute la longueur de la protéine ? Pourquoi ?
Recommencez l'alignement en utilisant cette fois UNIPROT. Qu'observez-vous ?

2. 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 ?

3. 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. 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. Pour cela nous utiliserons l'outils de génomique comparative VISTA disponible ici : `http://genome.lbl.gov/vista/index.shtml`.
Cliquez sur Browser". Sélectionnez la version la plus récente du génome humain, et la position PAH. Quelles sont les régions conservées ?

Exercice 3 : Réseau métabolique

À partir de la fiche Swiss-Prot, 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 ?

Exercice 4 : 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 ?