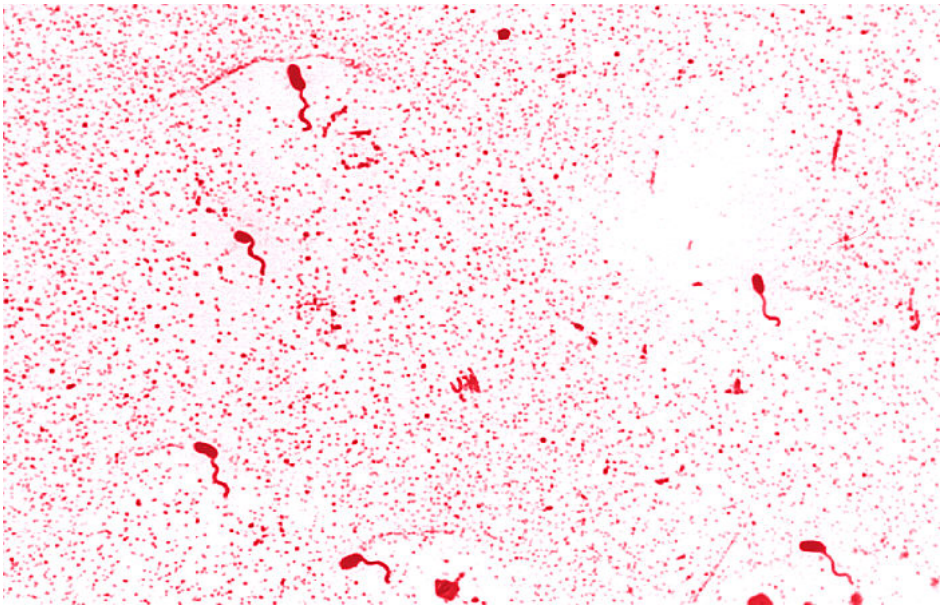
3I019 : Travaux Pratiques

Projet Analyse de génomes



Responsables. Hugues Richard (hugues.richard@upmc.fr) et Juliana Silva-Bernardes (juliana.silva_bernardes@upmc.fr).

Enseignants TP. Yasaman Karami (yasamankarami@gmail.com) et Ari Ugarte (ari.ugarte@gmail.com)

Introduction à la Bio-Informatique (3I019)

Le but de ce mini-projet est d'analyser de manière comparative le génome de deux bactéries : un organisme déjà bien connu *E. coli* (souche K12, qui n'est pas pathogène) et une bactérie pathogène des Entérococques (deux bactéries sont données au choix).

Le but du projet est d'analyser les propriétés générales du génome, de détecter les différentes régions de composition homogène, de pouvoir trouver les éléments génétiques qui lui confèrent sa toxicité, et de comprendre comment cette toxicité a été acquise.

Ce mini projet est organisé en trois parties successives pour pouvoir analyser les propriétés des génomes :

Partie A : propriété globales des génomes et détection des hétérogénéités dans la séquence.

Partie B : Analyse comparative, détection des îlots de pathogénicité, analyse des propriétés, annotation des gènes correspondants.

Partie C : Analyse de l'usage des codons et lien avec les propriétés des gènes et les îlots de pathogénicité.

Les deux génomes d'Entérobactéries sont les suivants :

- Groupe 1 : *Shigella dysenteriae*, une enterobactérie pathogène rencontrée chez l'homme (responsable de la dysenterie bacillaire, ou de la shigellose). Cette bactérie provoque des infections intestinales localisées essentiellement au gros intestin où les germes se multiplient en provoquant une inflammation de la muqueuse se traduisant par une diarrhée glaireuse et sanguinolente.
- Groupe 2 : Vibrio cholerae, est une bactérie gram négatif responsable du choléra chez l'Homme.

A - Première annotation des génomes

Dans un premier temps, nous allons utiliser les fonctions et les méthodes développées durant les TME précédents pour faire une première annotation des génomes.

Chaque groupe travaille avec le génome de *E. coli* et au choix le génome de *Shigella dysenteriae* ou *Vibrio cholerae* (juste un des deux, pas les deux !). Vous pouvez télécharger les séquences sur le site du module.

Préliminaires, propriétés de base

1. Après avoir téléchargé les deux génomes que vous analyserez par la suite, produire une liste des premières propriétés en rapportant dans une table pour l'organisme choisi (cf TME3)
    a. Le nombre de chromosomes et de plasmides, leur longueurs, la longueur totale du génome, le pourcentage en GC global et la composition en nucléotides.
    b. Découpez le génome en blocs 1kbp non chevauchants et calculer de % en GC dans chaque bloc. Faites un histogramme de la distribution du GC, que remarquez vous ? Est ce que les blocs avec un GC atypique ont tendance à être colocalisés ?

2. Annotons l'ensemble des ORFs du génome choisi avec la méthode vue au TME4 et comparons le résultat avec l'annotation disponible (fichier tab). Nous utiliserons un programme d'annotation automatique des gènes comme Glimmer.
    a. Produire un fichier d'annotation des gènes en utilisant Glimmer.
    b. Faites un histogramme des longueurs des gènes prédits avec Glimmer et un histogramme des longueurs des gènes du fichier tab. Y a-t-il des différences ?
    c. Adapter la fonction `compare_intervalle` pour comparer les annotations obtenues avec Glimmer avec le fichier tab d'annotation fourni sur les deux brins et renvoyer une matrice de confusion. donnez les sensibilités et spécificités de la prédiction.

On travaille à partir de maintenant et pour toute la suite avec les annotations du fichier tab.

    d. Extraire les séquences codantes des gènes et les écrire dans un fichier fasta `genes_codants.fasta`.
    e. Il y a des séquences qui possèdent un ou plusieurs stop codon en phase ou qui ne commencent pas par un start codon. Nous les analyserons plus tard. Pour l'instant séparez ces gènes dans un autre fichier que nous appellerons `genes_non_codants.fasta`.
    f. Calculer pour chaque gène son pourcentage en GC. Mettez à jour le fichier tab en rajoutant l'information dans une colonne.
    g. Faites un histogramme des pourcentages en GC des gènes. Que remarquez vous ? Comparez avec l'histogramme obtenu sur coli, commentez. Pouvez vous faire le lien avec l'histogramme obtenu à la question 1.b ?

Annotation par homologie avec BLAST

Nous allons maintenant tenter d'annoter la fonction des gènes du génome choisi en utilisant d'abord l'annotation du génome de *E. coli*.

3. Traduisez les séquences des gènes codants en séquences protéiques et faites un blast de toutes les protéines du génome choisi contre les protéines de *E. coli.* Quel type de BLAST doit-on utiliser ?
    a. A partir des résultats, fixer un seuil de E-value (10^{-3} par exemple) et ne garder que les alignement au dessous du seuil.
    b. Pour chaque alignement, vérifier la couverture de la séquence query et éliminer les alignement qui couvrent moins 80% du gène.
    c. Rajoutez dans le fichier tab le nombre d'alignements significatifs (0,1,2..) obtenus pour chacun des gènes. A quoi correspondent les cas où un gène a plus d'un alignement ?
    d. Combien de gènes n'ont aucun alignement ? Faire un histogramme de leur pourcentage en GC, que remarquez vous ?
4. Alignons maintenant les séquences des gènes non codants (fichier `genes_non_codants.fasta`)
    a. Faire un alignement avec blastn sur la base de données nr.
    b. Pour quelles familles de gènes obtenons nous des alignements ?
5. Assignation des catégories fonctionnelles pour les gènes codants avec COG