Traitement de données génomiques avec Perl Orienté Objet et BioPerl

IUT Génie Biologique, Option Bioinformatique (2e année)

Bérénice Batut

Février 2016

Introduction

Les séances de TP vont s'organiser autour d'un mini-projet qui vous permettra de vous familiariser avec le traitement de séquences en utilisant des classes, des références et BioPerl.

Principe du projet

L'objectif du projet est traiter très simplement un jeu de données métagénomique, c'est-à-dire un jeu de données contenant de nombreuses séquences dont on ne connait pas l'origine. Les traitements passent un contrôle de la qualité des séquences et un essai d'affiliation des séquences à des séquences connues.

Ce projet vous permettra de vous familiariser avec ce type de données pour mieux les utiliser dans le projet du cours de Métagénomique et d'avoir une vue rapide de ce qu'il y a derrière les outils de traitement de ce type de données.

Organisation

Pour ce projet, vous vous mettrez par binome. Les deux membres du binome doivent participer au projet, pour les codes mais aussi la rédaction du compte-rendu. D'ailleur, il serait bien qu'au moins un des membres du binome amène son ordinateur personnel. De plus, la personne qui "code" et la personne qui "rédige" ne doivent pas être toujours les même.

La note du projet portera sur le compte-rendu, les codes fournis et le suivi des instructions (2 points).

Instructions

L'organisation est importante tout le temps, même en information et en bioinformatique. En particulier dans un projet bioinformatique, il est important de bien centralisé à un seul endroit (dossier) tous les fichiers concernant le projet afin de garder une trace. De plus, les fichiers au sein de ce dossier doivent être organiser de façon méthodique afin de faciliter votre travail mais aussi celui de vos collaborateur.

Ainsi, tous les fichiers liés à ce projet sont rassemblés dans un dossier comprenant

- Un dossier `data` pour toutes les données
- Un dossier `results` pour les résultats des analyses (graphiques, fichiers générés, ...)

- Un dossier `src` pour le script principal (nommé ") et les modules Perl développés
- Un dossier `doc` pour les notes et en particulier le compte-rendu du TP

Le compte-rendu dans le cadre de ce projet correspond à un "cahier de notes", où les différentes étapes de l'analyse (graphiques, résultats, échecs, ...). En effet, quand on travaille sur un projet, tenir un tel "cahier" permet de garder une trace de ce qui a été fait, ce qui a marché pour pouvoir expliquer la démarche, faciliter la reprise du travail par quelqu'un, ... Dans ce "cahier", vous noterez les différentes étapes, les grandes lignes des codes, les méthodes particulières utilisées (et pourquoi), les graphiques, les liens vers les fichiers utiles ainsi que les sites. Pour faciliter la rédaction, un modèle de cahier en `markdown` doit être utilisé. Il est disponible sur bebatut-edu.github.io.

L'organisation et la collaboration passent aussi par le rédaction de codes "propres". Un code "propre" permet de faciliter la prise en main du code par quelqu'un de nouveau, mais aussi par vous dans quelques mois. Ecrire un code "propre" implique de suivre quelques règles

- Documenter le code (avec POC en Perl)
- Faire attention à l'indentation
- Ne pas mettre trop de caractères par ligne (généralement, moins de 80)

Deux modèles sont disponibles sur bebatut-edu.github.io : un modèle pour les classes et un modèle pour les scripts. Ils doivent être utilisé dans ce projet. Ces modèles intègrent de la documentation qui doit être mise à jour régulièrement et visualiser avec `perldoc`.

Pour visualiser l'enchainement nécessaire des analyses et garder une trace des paramètres choisis, toutes les commandes utilisées seront exécutées depuis un script Perl principal nommé `traitement_donnees.pl`. Cependant, comme toutes les étapes ne seront pas à exécuter à chaque fois, il faut penser à mettre des conditions de test pour les exécutions.

Récupération des données

Création d'une classe `Sequence`

Parcours d'un fichier de séquences

Contrôle de la qualité des séquences

Assignation des séquences à des séquences connues