Université Claude Bernard Lyon 1
Département de BIOLOGIE
Master Bioinformatique moléculaire
Année universitaire 2017/2018

UMR CNRS 5558 - LBBE
"Biométrie et Biologie évolutive"
UCB Lyon 1 - Bât. Grégor Mendel

CAHIER DES CHARGES

Développement d'un pipeline d'assemblage et d'annotation de génome bactérien déployable sur le cloud

Auteurs :
Cécile Hilpert
Sumaira Javaid
Lucía Castro Garcia
Krystian Valenducq

Encadrants :
M. Philippe Veber
M. Stéphane Delmotte

Table des matières

1 Présentation du projet

1.1 Problématique

Grâce aux techniques de séquençage haut-débit, la recherche scientifique est aujourd'hui capable d'acquérir dans un intervalle de temps très court, des volumes massifs de données génomiques, transcriptomiques, etc... De ce fait, le problème actuellement n'est plus la production de données, mais leur analyse. Cette production massive de données nécessite de méthodes d'analyses plus sophistiquées qui sont le plus souvent très coûteuses, que ce soit en temps de calcul ou mémoire.

La majorité des analyses sont aujourd'hui dites "à façon", conçues spécifiquement pour fonctionner avec certaines données expérimentales. Cependant, dans les protocoles d'analyse de données il existe des étapes récurrentes que l'on peut automatiser. Par exemple, l'assemblage du génome et son annotation, se basant sur des outils existants.

1.2 Objectifs

L'objectif de ce projet est de rendre automatique ces étapes récurrentes en analyse génomique, en développant un outil disponible sur le web (cloud) d'assemblage et annotation automatiques de génome, spécifiquement de génome bactérien. Le but de la création de cet outil n'est pas uniquement d'assister les travaux de recherche des bioinformaticiens, mais aussi de faciliter leur utilisation par des scientifiques n'ayant pas de compétences en informatique, avec par exemple une interface web très "user-friendly". Un autre objectif est de valider une méthodologie permettant de transformer une chaine de traitement en un outil accessible visant à réduire significativement le travail logiciel.

1.3 Description de l'existant

Il existe des outils d'analyse automatique de données qui sont disponibles sur le web. Néanmoins, ces outils ne répondent pas complètement à nos besoins. Une liste de ces outils sera décrite ci-dessous, avec une description correspondant à chaque outil et les raisons pour lesquelles il n'est pas adapté aux besoins présentés dans la problématique.

Galaxy
Il s'agit d'une plateforme d'analyse de données génomiques, métagénomiques et transcriptomiques en ligne. Il permet par exemple le trimming de séquences, l'alignement contre un génome de référence, mais aussi d'effectuer une analyse CHIP-seq, de faire du reséquençage et de l'assemblage (RNA, séquences PE et DNA), avec une vérification de la qualité de ce dernier. On y trouve également d'autre fonctionnalités comme l'annotation de SNPs

et ou encore la prédiction de leur effet. Cependant, les logiciels d'annotation qui y sont disponibles (ANNOVAR, GEMINI, SNPEff) servent uniquement à annoter les variants génétiques. De plus, il est nécessaire d'implémenter le pipeline "soi-même", c'est-à-dire qu'il faut lancer d'abord l'assemblage puis l'annotation. De plus, la conception de nouveaux pipelines ou l'ajout d'outils sont compliqués à implémenter pour les bioinformaticiens et la lecture des résultats est difficile pour les grosses analyses de données.

MicroScope
Il s'agit d'une plateforme web qui permet de faire de l'analyse comparative de génome microbien et de l'annotation fonctionnelle manuelle de ces derniers.

MyPro
Cet outil permet l'assemblage et l'annotation de génomes procaryotes. Les seules façons d'utiliser le pipeline sont : (1) Télécharger et compiler les codes sources et télécharger les différents logiciels nécessaires. Pour cela, il est nécessaire d'être familiarisé avec l'environnement Unix et les outils en lignes de commande. (2) Utiliser une version VirtualBox avec tous les logiciels installés. Cependant, une installation de ce type est bien plus lourde qu'uniquement utiliser un navigateur web pour accéder au pipeline.

MEGAnnotator
Il s'agit d'un pipeline d'assemblage, vérification de la qualité d'assemblage et d'annotation user-friendly pour les génomes procaryotes. Le logiciel comprend une interface graphique facile d'utilisation mais l'étape d'installation de cet outil peut s'avérer compliquée pour les utilisateurs sans compétences informatiques (surtout en langage bash). L'utilisation peut se faire uniquement dans environnement Unix ou en installant une VirtualBox mais ceci diminue la mémoire disponible et la vitesse d'exécution.

2 Expression des besoins

2.1 Besoins fonctionnels

Données de départ

Dans un premier temps, les données de départ seront des reads courts (Illumina). En fonction du temps disponible, la modification du pipeline pour pouvoir y intégrer l'analyse de reads longs (PacBio, MinIon) pourra être envisagée.

2.1.1 Pipeline d'assemblage et d'annotation

— Nettoyage des données brutes (élimination de contamination avec du DNA d'autres organismes). Élimination d'adaptateurs et de lectures de mauvaises qualités
— Assemblage de novo du génome bactérien choisi
— Annotation des gènes codants, 2 stratégies possibles : (1) Aligner le génome contre une base de données protéiques (blastx) pour identifier les gènes codants pour des protéines identifiées dans la littérature. (2) Détecter les gènes de novo et les annoter par alignement

2.1.2 Vérification de la qualité du pipeline

— Comparaison de l'assemblage avec la référence disponible par alignement
— Comparaison des gènes annotés avec ceux annotés dans le génome de référence
— Mettre en place une métrique permettant de qualifier la qualité de l'assemblage par rapport à la(les) référence(s)

2.1.3 Environnement d'exécution du pipeline

— Ecriture du pipeline en langage OCaml avec la bibliothèque Bistro
— Isolement du pipeline et de ses ressources (installations des logiciels bioinformatiques) dans un «container» créé avec docker sur un cloud
— Création d'un formulaire web permettant l'interaction des utilisateurs avec notre pipeline

2.1.4 Point de vue utilisateur

Pour lancer le pipeline, l'utilisateur devra se connecter à une machine virtuelle sur un cloud, se connecter à cette machine à l'aide d'un navigateur web puis utilisera un formulaire pour renseigner les paramètres et les fichiers d'entrée.

2.2 Besoins non-fonctionnels

Exigence de qualité

— Aide contextuelle (documentation accessible en ligne sous forme de tutoriel et dans un document à part).

Exigence de performances

— Temps d'exécution raisonnable en rapport avec la taille des données
— Elimination de bugs informatiques

Accesibilité

— Rendre l'outil disponible via le cloud

Internationalisation

— Les commentaires à l'intérieur du code et la documentation seront en anglais

2.3 Informations relatives aux contenus

Ressources générées

Fichiers de sorties contenant le génome assemblé (ensemble de contigs ou dans le meilleur des cas, un seul contig au format MULTIFASTA) et le fichier GFF contenant les annotations des gènes

Obtention des données générées

Les fichiers créés seront disponibles via le navigateur web, où ils pourront être téléchargés par l'utilisateur

3 Contraintes

3.1 Coûts

Nous n'avons pas de coûts limitants pour ce projet. Moyens matériels : Le projet sera réalisé sur les ordinateurs personnels des étudiants impliqués. Le cloud girofle du LBBE sera utilisé pour tester notre pipeline.

3.2 Délais

Date de rendu du cahier des charges : 20 octobre 2017

Date de rendu du projet : 19 décembre 2017

Soutenance de projet : 21 décembre 2017

3.3 Contraintes d'implémentation (souhaitées par le client)

Pour les développeurs

— Utilisation du langage de programmation Ocaml et de la librairie bistro développée par le LBBE
— Utilisation de docker pour virtualiser notre pipeline

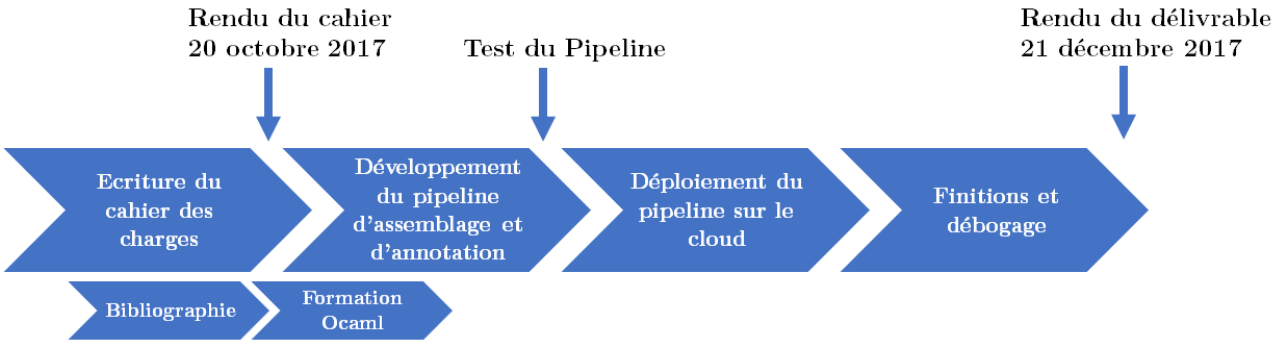
4 Déroulement du projet

4.1 Plannification

Le projet se décompose essentiellement en trois parties :

— Bibliographie pour déterminer les programmes à utiliser, et rédaction du cahier des charges
— Développement et évaluation du pipeline sur un jeu de données de référence
— Déploiement sur le cloud

Le projet comprendra également, une formation au langage Ocaml, prise en main de la librairie bistro. Les différentes tâches seront partagées entre les étudiants. Certaines personnes se concentreront plutôt sur déploiement du pipeline sur le cloud et d'autres sur le développement et le test du pipeline.



4.2 Description des étapes du pipeline et logiciels utilisés

4.2.1 Contrôle qualité et trimming des séquences brutes

FASTQScreen : analyse de contamination

FASTQC-trimmomatic : contrôle qualité et trimming de lectures courtes

LORDEC : contrôle qualité et trimming de lectures courtes

4.2.2 Assemblage des lectures

SPAdes [1] : assembleur conçu pour l'assemblage de novo de génomes bactériens

QUAST [3] : contrôle qualité de l'assemblage (si un génome de référence est disponible)

4.2.3 Alignement contre génome de référence

BOWTIE2 [4] : aligneur de reads longs contre un génome de référence indexé

4.2.4 Annotation du génome

ORFinder : détection de gènes de novo

BLAST [2] : annotation comparative et détection de gènes homologues

PROKKA : annotation de génomes procaryotes

4.3 Plan d'assurance qualité

Pour s'assurer de la qualité du pipeline d'assemblage/annotation, il sera testé sur un jeu de données test de séquençage d'une espèce bactérienne bien annotée dans les bases de données. A priori, les données d'Escherichia coli seront utilisée. Un jeu de données de séquençage Illumina devra être trouvé.

4.4 Responsabilités

Maîtres d'ouvrage

Philippe Veber et Stéphane Delmotte (LBBE) sont les maîtres d'ouvrage de ce projet. Les contraintes de temps sont imposées par le Master 2 Bioinformatique de l'université Lyon 1.

Maîtres d'oeuvre

Les quatres étudiants concernés par le projet sont maîtres d'oeuvre, à savoir Lucía Castro García, Cécile Hilpert, Sumaira Javaid et Krystian Valenducq.

Références

[1] Anton Bankevich, Sergey Nurk, Dmitry Antipov, Alexey A Gurevich, Mikhail Dvorkin, Alexander S Kulikov, Valery M Lesin, Sergey I Nikolenko, Son Pham, Andrey D Prjibelski, et al. Spades : a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology*, 19(5) :455–477, 2012.

[2] Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden. Blast+ : architecture and applications. *BMC bioinformatics*, 10(1) :421, 2009.

[3] Alexey Gurevich, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. Quast : quality assessment tool for genome assemblies. *Bioinformatics*, 29(8) :1072–1075, 2013.

[4] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. *Nature methods*, 9(4) :357–359, 2012.