Mémoire de stage
TITRE DE L'INTITULE DE STAGE

Gabriel Toublanc, Master IGIS ITA, 2^{eme} année

25 Juin 2017

Table des matières

Chapitre 1

Introduction

1.1 Cadre

Ce mémoire a été rédigé dans le cadre de mon stage de fin de Master IGIS ITA, dans le but de réaliser un rapport complet sur l'ensemble du travail effectué. Ce stage s'est déroulé du 1^{er} avril au 30 juin 2017, au sein de l'équipe TIBS du laboratoire LITIS de l'Université de Rouen, dans le batiment Monod puis CurieB sur le campus de Mont-Saint-Aignan. Les travaux de cette équipe portent principalement sur la recherche, l'indexation et l'extraction d'informations pertinentes dans des données biologiques et des systèmes d'information en santé, et offrent donc de nombreuses applications dans ces domaines. Ce stage a été financé par le LITIS, grâce à des fonds alloués par l'Université de Rouen, destinés à permettre l'accueil de stagiaires de seconde année de Master Recherche.

1.2 Contexte

L'analyse de séquençage ADN appliqué à la Bio-Informatique a connu un développement important dans les années 2000.
Le séquençage ADN est effectué de deux façons différentes, par le biais de lectures sur des brins d'ADN :

$\longrightarrow$ **Les lectures longues** (Nanopore, PacBio, ...)
$\longrightarrow$ **Les lectures courtes** (Illumina, Roche, ...)

Les **lectures longues** ont l'avantage de nécessiter que d'**une petite quantité** d'entre elles pour quadriller le génome d'un être vivant, mais le gros désavantage d'être **hautement imprécises** (15 à 30 % d'erreurs)

Les **lectures courtes**, en revanche, ont le net avantage d'être **très précises**, avec un taux d'erreur de moins de 1 %. Leur inconvénient principal est qu'il faut **de nombreuses lectures** courtes afin de reconstruire le génome, ce

qui est coûteux et peu pratique.

Dans les faits, le génome peut être reconstruit grâce à un nombre déraisonnable de lectures courtes, ou en corrigeant les lectures longues grâce aux lectures courtes. Plusieurs **méthodes de corrections** ont été développées à ce sujet [1].

Afin de pouvoir se passer des lectures courtes, et d'ainsi corriger les lectures longues d'elles-mêmes, une branche de la bioinformatique s'intéresse à l'étude des **k-mers**.

Chapitre 2

K-mers

Chapitre 3

K-mers espacés

Chapitre 4

kmerCount

Chapitre 5

kmerExpand

Chapitre 6

kmerDel

Chapitre 7

MAWS

Chapitre 8

LCS

Chapitre 9

Conclusion

Chapitre 10

Références

Bibliographie

[1] Pierre Morisse, Thierry Lecroq, and Arnaud Lefebvre. Hg-color : Hybrid-graph for the error correction of long reads. In *Actes des Journées Ouvertes Biologie Informatique et Mathématiques*, 2017.