Mise au point d'outils bioinformatiques d'analyse métagénomique : application à la caractérisation des communautés Rhizobiales au sein des microbiomes des sols forestiers de Côte d'Ivoire.

A Thesis
Presented to
The Division of Sciences Agronomiques et Genie Rural
Ecole Doctorale Polytechnique

In Partial Fulfillment
of the Requirements for the Degree
Philosophia Degree

Ediman Theodore Anicet Ebou

Janvier 2020

Approved for the Division
(Bioinformatics)

Dominique Koua	Adolphe Zeze

Table of Contents

List of Tables

List of Figures

Présentation générale

La microbiologie environnementale est un domaine intimement relié a l'écologie microbienne, et qui s'intéresse aux interactions entre les micro-organismes et l'environnement qui les entoure. La métagénomique, est la partie de la microbiologie environnementale qui vise à séquencer l'ensemble des ADN d'un microbiome. La métagénomique est l'un des domaines qui par l'essor des nouvelles technologies de séquençage de l'ADN à connue une avancée majeure. Anciennement tributaire des milieux de culture et méthodes d'isolation des micro-organismes pour la caractérisation des microbiomes, l'ère moléculaire a permis l'avènement de la génomique environnementale. De fait, la génomique environnementale permet l'investigation des champignons et bactéries de l'environnement à une plus grande échelle en scannant des millions de séquences nucléotidiques en une seule analyse.

L'analyse des données issues du séquençage environnemental constitue un véritable challenge pour la recherche. C'est dans ce contexte que les travaux de recherche envisagés visent le développement d'outils d'analyse et de décision pour faciliter la compréhension des communautés microbiennes de notre environnement notamment les communautés de bactéries symbiotiques des plantes cultivées, les Rhizobiales. Les travaux qui seront réalisés dans le cadre de nos travaux de thèse devraient permettre en particulier : (i) de proposer des outils bioinformatiques permettant d'exploiter des jeux de données génomiques importants et (ii) de confirmer des hypothèses biologiques expliquant, au niveau génomique, la complexité des communautés microbiennes du phytomicrobiome. De tels outils sont nécessaire à la mise en œuvre d'une agriculture durable qui requiert, entre autres, une meilleure compréhension et caractérisation fonctionnelle des micro-organismes. Nos résultats permettront une meilleure compréhension des populations qui façonnent les écosystèmes agricoles et qui expliquent les variations de rendement et de l'état de santé des plantes présentes dans des conditions de culture similaires.

Au regard des progrès fulgurants réalisés dans le domaine du séquençage, la génération de données de séquençage environnemental ne constitue plus le problème majeur. Le principal défi à l'heure actuelle est d'analyser et de traiter les grandes quantités de données moléculaires rendues disponibles. L'analyse et le traitement des données, sont, dans la plupart des cas, réalisés avec des méthodes non adaptées à la question de recherche posée. Ainsi, les résultats d'analyse sont difficile d'utilisation, les données et annotations externes ne sont généralement pas intégrées et les résultats obtenus ne sont pas transformés en outils de décision. Pour répondre à ces défis, les travaux envisagés comporteront trois axes. Premièrement, il n'existe à ce jour aucune base de

données annotée et maintenue pour les séquences de 16S de Rhizobiales. Cette insuffisance conduit les chercheurs à utiliser des bases de données plus grandes et surtout plus généralistes, ce qui a pour effet d'augmenter le temps d'analyse et de réduire la précision des résultats obtenus. Deuxièmement, la plupart des outils d'analyse sont encore proposés en ligne de commande. Non seulement cela rend l'analyse inaccessible aux non initiés, mais en plus, cela rend l'utilisateur lambda tributaire d'un expert capable de faire l'analyse. Bien que la présence d'un expert soit indispensable pour la réussite de tels projets, des outils d'analyses performants et adaptées , proposés avec une interface graphique permettraient certainement d'améliorer et de faciliter la recherche et la découverte dans ce domaine. Enfin, troisièmement, la plupart des résultats obtenus utilisent des pipelines d'analyse non spécifiques à la question de recherche posée. Les mêmes protocoles d'analyse sont donc généralement utilisés pour l'analyse des micro-organismes quels que soient leur origine ou leur milieu de vie.

Nous proposons donc, en premier lieu, de développer une base de données annotée de séquences spécifiques aux Rhizobiales. En second lieu, nous proposerons un pipeline d'analyse des données moléculaires issues des collectes d'échantillons environnementales. L'outil qui sera développé devrait fournir des séquences annotées, des graphiques et des alignements multiples de séquence prêts pour publication. Enfin, nous effectuerons une analyse comparative des communautés de Rizobiales en Afrique de l'ouest en utilisant l'outil qui sera développé.

Introduction

Les être vivants présent dans l'environnement qui nous entoure ne vivent généralement pas de façon disparate mais plutôt en communautés et/ou en ayant des relations entre eux. Une grande partie des êtres vivants sont en fait des micro-organismes ou des organismes invisibles à l'œil nu. Les communautés qu'ils forment sont appelés microbiomes et sont à l'origine de plusieurs modifications environnementales observable macroscopiquement. La microbiologie environnementale est la branche de la microbiologie qui se charge de l'étude des micro-organismes présent dans le sol, l'eau, l'air et les sédiments.

Chapter 1

Etat de l'art: méthodes d'analyses des données des communautés microbiennes

Chapter 2

Base de données métagénomique pour les Rhizobiales bénéfiques et non bénéfiques

Chapter 3

Amélioration de la classification des taxons issus du metabarcoding par utilisation des profils de Markov, des profils généralisés et de l'association phage-hôte inféré par le système CRISPR-Cas/anti CRISPR

Chapter 4

Outils d'analyse en interface graphique pour l'analyse de données des communautés microbiennes

Conclusion

Appendix A

The First Appendix

This first appendix includes all of the R chunks of code that were hidden throughout the document (using the `include = FALSE` chunk tag) to help with readibility and/or setup.

In the main Rmd file

```
# This chunk ensures that the thesisdown package is
# installed and loaded. This thesisdown package includes
# the template files for the thesis.
if (!require(remotes)) {
  if (params$`Install needed packages for {thesisdown}`) {
    install.packages("remotes", repos = "https://cran.rstudio.com")
  } else {
    stop(
      paste('You need to run install.packages("remotes")",
            "first in the Console.')
    )
  }
}
if (!require(thesisdown)) {
  if (params$`Install needed packages for {thesisdown}`) {
    remotes::install_github("ismayc/thesisdown")
  } else {
    stop(
      paste(
        "You need to run",
        'remotes::install_github("ismayc/thesisdown")',
        "first in the Console."
      )
    )
  }
}
library(thesisdown)
```

```
# Set how wide the R output will go
options(width = 70)
```

In Chapter ??:

Appendix B

The Second Appendix, for Fun

Appendix C

Algorithmes et outils

Appendix D

Publications

References

Angel, E. (2000). *Interactive computer graphics : A top-down approach with opengl*. Boston, MA: Addison Wesley Longman.

Angel, E. (2001a). *Batch-file computer graphics : A bottom-up approach with quicktime*. Boston, MA: Wesley Addison Longman.

Angel, E. (2001b). *Test second book by angel*. Boston, MA: Wesley Addison Longman.