Algorithmes pour la reconstruction de séquences de marqueurs conservés dans des données de métagénomique

Pierre Pericard

École Doctorale SPI

UMR 9189 - CRIStAL

Thèse

Présentée pour l'obtention du grade de DOCTEUR
DE L'UNIVERSITE DE LILLE
par

Pierre Pericard

Algorithmes pour la reconstruction de séquences de marqueurs conservés dans des données de métagénomique

---

Spécialité : Informatique

Soutenue le 27 Octobre 2017 devant un jury composé de :

Rapportrice	**Claudine Médigue**	(DR CNRS, Genoscope)
Rapporteur	**Dominique Lavenier**	(DR CNRS, IRISA/INRIA)
Examinatrice	**Laetitia Jourdan**	(Prof. Univ. Lille, CRIStAL)
Examinateur	**Pierre Peyret**	(Prof. Univ. Clermont Auvergne)
Directrice de thèse	**Hélène Touzet**	(DR CNRS, CRIStAL)
Co-encadrant de thèse	**Samuel Blanquart**	(CR Inria, CRIStAL)

Affiliations :

Thèse effectuée au sein de **l'UMR 9189 - CRIStAL**
de l'Université de Lille
Bâtiment M3 extension
avenue Carl Gauss
59655 Villeneuve d'Ascq Cedex
France

Résumé

Les progrès récents en termes de séquençage d'ADN permettent maintenant d'accéder au matériel génétique de communautés microbiennes extraites directement d'échantillons environnementaux naturels. Ce nouveau domaine de recherche, appelé *métagénomique*, a de nombreuses applications en santé, en agro-alimentaire, en écologie, par exemple. Analyser de tels échantillons demande toutefois de développer de nouvelles méthodes bio-informatiques pour déterminer la composition taxonomique de la communauté étudiée. L'identification précise des organismes présents est en effet une étape essentielle à la compréhension des écosystèmes même les plus simples. Cependant, les technologies de séquençage actuelles produisent des fragments d'ADN courts et bruités, qui ne couvrent que partiellement les séquences complètes des gènes, ce qui pose un véritable défi pour l'analyse taxonomique à haute résolution.

Nous avons développé MATAM, une nouvelle méthode bio-informatique dédiée à la reconstruction rapide et sans erreurs de séquences complètes de marqueurs phylogénétiques conservés, à partir de données brutes de séquençage. Cette méthode est composée d'une succession d'étapes qui réalisent la construction et l'analyse d'un graphe de chevauchement de lectures. Nous l'avons appliquée à l'assemblage de la petite sous-unité de l'ARN ribosomique sur des métagénomes simulés, synthétiques et réels. Les résultats obtenus sont de très bonne qualité et améliorent l'état de l'art.

Algorithms for conserved markers sequences reconstruction in metagenomics data

Abstract

Recent advances in DNA sequencing now allow studying the genetic material from microbial communities extracted from natural environmental samples. This new research field, called *metagenomics*, is leading innovation in many areas such as human health, agriculture, and ecology. To analyse such samples, new bioinformatics methods are still needed to ascertain the studied community taxonomic composition because accurate organisms identification is a necessary step to understand even the simplest ecosystems. However, current sequencing technologies are generating short and noisy DNA fragments, which only partially cover the complete genes sequences, giving rise to a major challenge for high resolution taxonomic analysis.

We developped MATAM, a new bioinformatic methods dedicated to fast reconstruction of low-error complete sequences from conserved phylogenetic markers, starting from raw sequencing data. This methods is a multi-step process that builds and analyses a read overlap graph. We applied MATAM to the reconstruction of the small sub unit ribosomal ARN in simulated, synthetic and genuine metagenomes. We obtained high quality results, improving the state of the art.

Remerciements

Je tiens tout d'abord à remercier mes deux encadrants, Hélène Touzet et Samuel Blanquart, pour leur support et leur patience.

Hélène, en tant que directrice de thèse, a su m'aider à maintenir le cap sur mon sujet, et m'a apporté la structure dont j'avais durement besoin pour éviter parfois de m'éparpiller. Je la remercie aussi pour m'avoir laissé le temps nécessaire pour comprendre mes propres erreurs et pour me former sur les compétences qui pouvaient me manquer. Enfin, je remercie grandement Hélène pour être restée positive et à l'écoute tout au long de ma thèse.

Samuel a aussi été indispensable à la réussite de cette thèse. Je le remercie particulièrement pour avoir toujours été disponible lorsque j'en avait besoin, et pour les longues discussions scientifiques que j'ai pu avoir avec lui. Samuel m'a permis de débattre et d'explorer de nouvelles idées, mais il m'a aussi appris à ne pas faire de compromis scientifiques et d'aller jusqu'au bout de mes raisonnements.

Je souhaite remercier Claudine Médigue et Dominique Lavenier pour avoir accepté d'être rapporteurs de mon manuscrit de thèse, pour le temps qu'ils ont passé à le lire, et pour les retours pertinents qu'ils m'ont apporté. Je remercie aussi les membres de mon jury, Claudine Médigue, Laeticia Jourdan, Dominique Lavenier et Pierre Peyret pour s'être déplacés et participer à l'évaluation de ma thèse.

Un immense merci à tous les membres actuels et passés de l'équipe BONSAI pour m'avoir chaleureusement accueilli pendant 4 ans. Un remerciement tout particulier à Yoann, pour sa contribution à l'algorithme de compression des graphes de chevauchements, mais aussi pour tout le temps passé à discuter d'algo, de sciences, et pour les moments associatifs partagés. Merci aussi à Loïc, qui prend maintenant la relève dans le développement de MATAM, et qui a été déterminant dans l'implémentation des derniers modules et dans la génération des résultats les plus récents.

Un grand merci à tous les choristes d'ACDJ et d'Azimuts, pour avoir été là pendant mes années à Lille. On sous-estime parfois l'importance d'un système de support moral pour les doctorants. Dans mon cas, je peux clairement dire que la complétion de ma thèse n'aurait pas été possible sans les moments de chants, de musique, et de fête que j'ai partagé avec vous, et qui m'ont permis de décompresser même dans les moments les plus durs.

Merci Papa et Maman. Gros bisous Marie.

Table des matières

Table des figures

Introduction

La compréhension de l'organisation du vivant et de sa diversité a toujours été un sujet scientifique important. C'est un domaine central en biologie, qui relève à la fois de la *taxonomie* et de la *systématique* : observer les êtres vivants et les classer.

Historiquement, les organismes vivants ont d'abord été classés sur la base des caractères observables à l'œil nu, tels que des caractères morphologiques. La pratique a ensuite évolué sous l'impulsion de progrès technologiques majeurs, renouvelant ainsi les critères utiles pour établir ces classifications. C'est ainsi que l'apparition du microscope optique dans la deuxième partie du XVIIe siècle a permis les premières observations d'organismes microscopiques, un préalable à leur intégration aux classifications du vivant et bien d'autres bouleversements scientifiques. C'est ensuite la découverte de l'ADN et le développement des premières technologies de *séquençage* au XXe siècle qui ont initié une deuxième révolution du domaine. Ces avancées ont permis l'analyse de caractères *biomoléculaires*, tels que des mutations au sein des gènes. Pour des raisons techniques, jusqu'au début des années 2000, cet effort de séquençage s'est porté exclusivement sur des organismes isolés et cultivés en laboratoire.

La généralisation des technologies de *séquençage haut débit* à laquelle on assiste actuellement, accompagnée du développement de nouvelles méthodes bio-informatiques pour l'analyse des données de séquençage sont en train de mener à une troisième révolution : le séquençage massif de communautés composées de micro-organismes non cultivés, échantillonnées dans des environnements naturels. On appelle ce domaine d'étude la *métagénomique*. L'effort de classification du vivant et ces nouvelles données permettent alors de répondre à un autre problème scientifique majeur : *caractériser la structure écologique d'un milieu*. On parle alors de l'analyse taxonomique d'un échantillon environnemental.

Les enjeux fondamentaux et appliqués de la métagénomique sont nombreux. Par exemple, l'étude de la flore intestinale chez l'humain, et la compréhension de la dynamique des organismes qui l'habitent, peuvent entraîner des progrès majeurs pour le traitement de pathologies comme l'obésité ou des troubles digestifs chroniques. De même, l'analyse régulière des micro-organismes peuplant les océans peut informer sur l'impact global des changements climatiques ou de la pollution.

D'un point de vue méthodologique, ces données de séquençage métagénomique posent de nouveaux problèmes algorithmiques passionnants. Il s'agit en effet des jeux

de millions de courtes séquences d'ADN, bruitées et mélangées, issues des génomes inconnus d'individus, présents en nombre inconnu, appartenant à un nombre d'espèces inconnu séparées durant leur évolution par des taux de divergence génétique inconnus. Accéder à l'ensemble de l'information consiste alors à reconstruire les séquences complètes des génomes de chacun des organismes de l'échantillon environnemental, à l'image de nombreux puzzles mélangés partageant des pièces similaires.

La reconstruction, ou *assemblage*, des génomes des micro-organismes d'échantillons environnementaux dans leur intégralité à partir du séquençage *métagénomique complet* de l'ensemble de ces génomes est un sujet particulièrement difficile pour lequel il n'existe pas encore de solution satisfaisante. En effet, la nature des données (courts fragments de séquences d'ADN), la complexité des matériels génétiques (répétitions, duplications génomiques, hybridation, transferts horizontaux, insertions, transpositions, etc. de matériel génétique) et la similarité des génomes des espèces proches rendent difficile le développement de méthodes et d'algorithmes capables de résoudre ce problème.

On peut toutefois envisager de résoudre efficacement des problèmes plus simples. À l'opposé du problème de l'assemblage de génomes complets, les premières approches de *métagénomique ciblée* proposaient de se focaliser sur le séquençage de régions restreintes et bien connues des génomes de ces organismes, appelées *marqueurs conservés*, c'est-à-dire des gènes partagés et dérivés chez un grand nombre d'espèces. On tâche ensuite de classer les séquences obtenues en groupes d'espèces proches, eux-mêmes identifiés taxonomiquement par leur similarité aux séquences connues les plus proches. Ces approches ont été abondamment explorées, et de nombreuses méthodes bio-informatiques performantes existent pour l'analyse de ce type de données. Toutefois, cette stratégie possède des biais et des limites intrinsèques qui empêchent la distinction entre des gènes très similaires issus d'espèces proches.

Une troisième voie, intermédiaire entre l'assemblage des génomes et le ciblage de matériel génétique, propose de se focaliser sur la reconstruction de marqueurs génétiques conservés à partir des fragments de séquences d'ADN produits par un séquençage métagénomique complet. On espère ainsi permettre une analyse taxonomique moins biaisée qu'en métagénomique ciblée et une identification plus précise des espèces présentes dans un échantillon environnemental. Un certain nombre de méthodes pour cette dernière approche existent déjà, mais nous avons identifié une marge de progression claire qui reste à explorer.

Ce manuscrit présente donc les travaux réalisés au cours de cette thèse, qui ont consisté au développement d'une nouvelle méthode de reconstruction des séquences de marqueurs conservés dans des données de séquençage métagénomique complet, appelée MATAM.

Ce manuscrit est organisé en quatre chapitres.

Le premier chapitre présente le contexte biologique de la thèse. Nous propo-

sons une introduction à la biologie environnementale, un rappel sur les technologies de séquençage de l'ADN, sur leur application à l'étude des communautés de micro-organismes dans le cadre de la métagénomique, et enfin sur les différents types de données générés dans ce contexte.

Le deuxième chapitre dresse un état de l'art des méthodes bio-informatiques développées pour l'analyse des données de séquençage génomique, puis des données de métagénomique. Nous finissons par présenter les outils et algorithmes existants pour la reconstruction de séquences de marqueurs conservés dans des données de séquençage métagénomique complet, dédiées au problème de leur analyse taxonomique.

L'objectif de ces deux premiers chapitres est de donner au lecteur quel qu'il soit, biologiste ou informaticien, les éléments nécessaires à la compréhension de nos travaux.

Dans le troisième chapitre, nous décrivons en détail la méthode que nous avons développée et implémentée sous la forme de l'outil MATAM. Nous illustrons aussi chaque étape des calculs réalisés par la méthode sur un jeu de données de test, simple et pédagogique, pour permettre au lecteur de bien en comprendre le fonctionnement et les choix que nous avons faits.

Finalement, dans le quatrième chapitre, nous présentons les résultats de l'évaluation de MATAM et de ses concurrents directs sur trois types de jeux de données différents, simulés, synthétiques et réels. Nous discutons enfin des limites et améliorations possibles pour de futures recherches.

Chapitre 1

Contexte Biologique

Ce premier chapitre a pour objectif de présenter le contexte biologique de la thèse, ainsi que de décrire les données sur lesquelles nous allons travailler et les méthodes de biologie moléculaire qui ont permis de les générer. Nous avons choisi de nous focaliser en particulier sur les problématiques liées à l'étude d'environnements complexes composés d'organismes microscopiques dans le cadre de la biologie environnementale.

Dans une première partie, nous introduisons la biologie environnementale et ses enjeux. Ensuite, nous expliquons comment le séquençage à haut débit et les technologies de biologie moléculaire modernes ont radicalement modifié la biologie environnementale. Enfin, nous présentons le domaine de la métagénomique, et les données qui sont générées.

1.1 Introduction à la biologie environnementale

1.1.1 Bref rappel historique, non exhaustif

La volonté d'étudier les environnements naturels et les organismes qui peuplent ces environnements remonte aux origines de la civilisation humaine. En effet, la caractérisation des organismes macroscopiques, comme les animaux et les plantes, a toujours été une compétence essentielle à la survie des humains dans les multiples environnements qu'ils ont occupés. On trouve d'ailleurs des tentatives de classification de ces organismes depuis les civilisations égyptienne et grecque antiques jusqu'à la Renaissance, en passant par le Moyen Âge. Ces classifications restaient toutefois principalement descriptives et se concentraient surtout sur les animaux et plantes utiles dans l'agriculture ou la médecine [57].

Pour ce qui est des organismes microscopiques, ceux-ci restent inconnus jusqu'au XVII^{e} siècle. C'est en 1668 que Antoine van Leeuwenhoek les observe pour la première fois grâce à un microscope de son invention. Il appelle ces organismes *animalcules* ; le terme *bactéries* (1.1.6.2), utilisé dans le langage courant d'aujourd'hui et dérivant du

grec pour « bâtonnet », n'apparaît qu'en 1838.

Les premières classifications scientifiques du vivant sont apparues au cours du XVIII^{e} siècle, avec notamment les travaux du botaniste Carl von Linné qui introduit une classification à sept niveaux basée sur des caractéristiques observables (1.1.6.1). Encore à l'époque, cette classification s'intéresse surtout aux organismes macroscopiques (animaux, plantes, champignons, etc.), que l'on qualifierait maintenant d'eucaryotes pluricellulaires (1.1.6.2). Des dérivées de cette classification, dite « classique », sont encore utilisées aujourd'hui.

En 1859, les travaux de Louis Pasteur sur les micro-organismes inaugurent la microbiologie, avec notamment la découverte des mécanismes de réplication, la mise au point de milieux de culture et de méthodes de destruction de ces micro-organismes. C'est à la fin du XIX^{e} siècle que les premières classifications de bactéries apparaissent, avec entre autres l'invention de la coloration de Gram en 1884, qui permet de partitionner les bactéries en deux groupes, bactéries Gram positifs et Gram négatifs, en fonction de propriétés de la paroi bactérienne.

En 1925, sur la base d'observations microscopiques, Edouard Chatton propose une classification des organismes cellulaires en deux types : les procaryotes (cellule sans noyau) qui comprennent toutes les bactéries, et les eucaryotes (cellule avec noyau) qui comprennent tous les organismes pluricellulaires visibles à l'œil nu, mais aussi des micro-organismes unicellulaires.

Il faudra attendre 1938 pour que les procaryotes soient élevés au rang de domaine, au même niveau que les eucaryotes. Et c'est en 1950 que les premières observations au microscope électronique permettent de mieux décrire la diversité de forme de ces micro-organismes et de définir des embranchements plus précis en fonction de caractéristiques morphologiques.

En 1977, Carl Woese est le premier à utiliser des caractères moléculaires pour classifier des organismes vivants. Grâce à l'analyse de l'ARN ribosomique 16S (1.1.5.1), il divise les procaryotes en deux domaines : les *Eubacteria* (« vraies » bactéries, renommées en *Bacteria* en 1990) et les *Archaebacteria* (renommées en *Archaea*) [99]. Cette stratégie de *classification phylogénétique* (1.1.6.1) est rapidement étendue aux eucaryotes, par l'utilisation de l'ARN ribosomique 18S, et devient l'approche de référence pour la classification des organismes vivants. Avec l'intérêt croissant pour l'étude des micro-organismes d'environnements très variés, et la découverte d'un nombre exponentiel de nouvelles espèces et souches microbiennes, l'intérêt de l'approche basée sur des caractères biomoléculaires devient considérable. C'est pourquoi depuis le début du XXI^{e} siècle, de nouvelles approches basées sur le séquençage haut débit (1.2.3) de ces organismes sont apparues et continuent d'être développées. C'est dans ce courant-là que cette thèse s'inscrit.

Glossaire

Dans ce manuscrit, nous utiliserons les termes de **bactéries** et **archées** au sens moderne du terme, pour désigner les domaines *Bacteria* et *Archaea*.
Des termes du langage courant, parfois sans réelle signification phylogénétique, sont utilisés dans ce document :

micro-organisme Organisme invisible à l'œil nu, unicellulaire (procaryote ou eucaryote) ou pluricellulaire (eucaryote).

espèce microbienne / microbe Micro-organisme unicellulaire.

procaryotes Terme désuet désignant l'ensemble des espèces de bactéries et d'archées.

virus Organisme microscopique non cellulaire et ne possédant pas de ribosome.

microbiote Communauté écologique de micro-organismes et de virus spécifiques à un environnement (appelé *microbiome*).

microbiome Environnement spécifique, dans lequel évolue un *microbiote*. En anglais, le terme *microbiome* se réfère à l'ensemble des génomes d'un microbiote donné.

1.1.2 L'exploration des micro-organismes dans leur environnement

Dans tous les environnements, on retrouve un mélange d'espèces microbiennes, appelé *microbiote*, qui évolue dans cet environnement, appelé *microbiome*, et qui peut être composé d'organismes des trois domaines du vivant (1.1.6.2), *bactéries*, *archées*, et *eucaryotes*, mais aussi de *virus*. C'est le cas des environnements les plus familiers comme la surface de notre peau ou le sol de notre jardin, tel que nous le développerons dans les sections 1.1.7.2 et 1.1.7.3. C'est également le cas d'environnements plus exotiques ou moins accessibles comme les grands fonds marins [9], les profondeurs d'une mine d'or [13], les flancs de la Station Spatiale Internationale [83], ou encore les nuages [7].

En fonction de l'environnement, le microbiote peut présenter des caractéristiques très différentes. Le principal facteur de complexité du microbiote est l'*abondance* respective des espèces qui le composent (Figure 1.1). Une espèce très *abondante* est largement majoritaire dans le microbiote (>50% du nombre d'individus), tandis qu'une espèce peu abondante, ou *rare*, sera présente en faibles proportions (< 0,1%). Par conséquent, un microbiote sera qualifié de simple lorsqu'il est composé d'une ou deux espèces largement majoritaires et de quelques espèces rares, alors qu'un microbiote très complexe peut être composé de milliers d'espèces et ne possède pas d'espèce majoritaire.

En pratique, quantifier la complexité des microbiotes est un problème ouvert. Ainsi, nous pouvons supposer que la plupart des microbiotes environnementaux peuvent

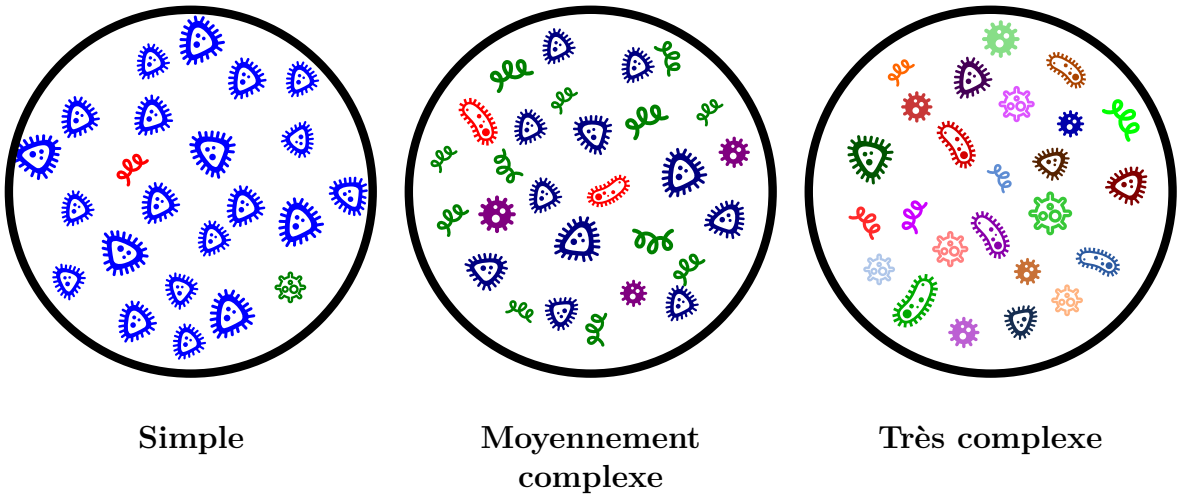


FIGURE 1.1 – Complexité d'un microbiote. Un microbiote *simple* est composé d'une espèce largement majoritaire et, optionnellement, de quelques espèces peu abondantes. Un microbiote *moyennement complexe* peut être composé de 2 à 3 espèces abondantes et d'une dizaine d'espèces peu abondantes. Un microbiote *très complexe* ne possède pas d'espèce majoritaire et peut être composé de milliers d'espèces d'abondances variables.

être qualifiés de (très) complexes, et sont composés de quelques organismes abondants (par exemple, qui représentent plus de 40% du microbiote), de dizaines d'organismes moyennement abondants, et potentiellement de plusieurs centaines ou plusieurs milliers d'organismes rares (avec des abondances $< 0{,}001\%$ du microbiote). Il est important de noter que l'abondance d'un organisme n'est pas forcément liée à l'importance de son rôle dans la communauté. Un organisme rare peut en effet avoir un impact majeur sur son environnement. On peut aussi souligner que la complexité d'un microbiote n'est pas statique, mais fluctue au gré des équilibres trophiques de l'écosystème, des saisons ou encore de variations moins régulières des paramètres physico-chimiques de l'environnement.

L'exploration des microbiotes dans des environnements variés représente un véritable enjeu pour la communauté scientifique, avec des applications potentiellement révolutionnaires en écologie, santé, et pour l'industrie (1.1.7). Par exemple, ces dernières années, des études chez l'humain ont permis de montrer un lien entre notre microbiote intestinal et des prédispositions à l'obésité [36], ainsi que la participation de ce microbiote intestinal à notre système immunitaire [8].

Afin de pouvoir analyser un microbiote et comprendre son fonctionnement, il faut commencer par identifier précisément les organismes présents dans l'environnement. Or, cette première étape reste difficile, et constitue à elle seule un enjeu. Les estimations les plus récentes suggèrent que plus de 99,999% de la biodiversité microbienne est encore inconnue [51]. On donne à cette biodiversité inconnue le surnom de *matière*

noire biologique. Mais même l'identification d'organismes déjà connus est délicate. En effet, à cause du manque de connaissances critiques sur leur biologie, on ne sait pas cultiver la majorité de ces micro-organismes dans des conditions de laboratoire [93], ce qui est souvent un préalable obligatoire à leur description.

La combinaison de la biodiversité inconnue et de la non-cultivabilité de la majorité des organismes rend donc l'identification précise des organismes sauvages difficile, même pour les environnements les plus simples ou les plus communs. Actuellement, la seule manière d'identifier les organismes d'un microbiote consiste à analyser directement le matériel génétique, son ADN (1.1.3) ou son ARN (1.1.4), obtenu par séquençage (1.2) d'un échantillon de cet environnement. On parlera alors de l'*analyse taxonomique* des échantillons environnementaux (1.1.6), et de l'*assignation taxonomique* des lectures issues du séquençage du matériel génétique.

1.1.3 L'ADN

Tous les organismes vivants partagent des systèmes moléculaires communs, dont l'ADN, la molécule qui sert de support de l'information génétique pour les organismes cellulaires.

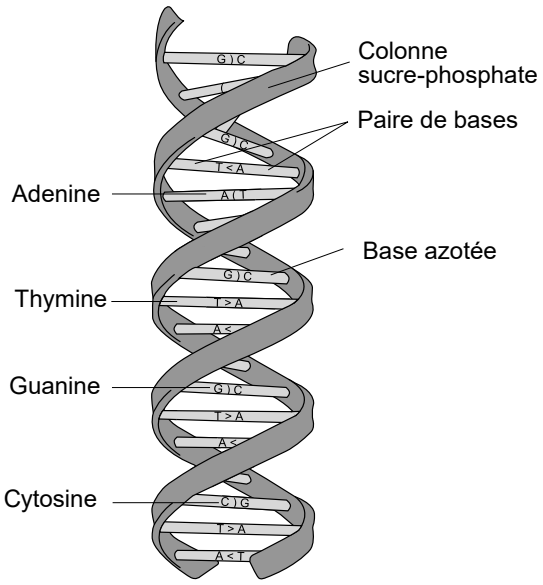


FIGURE 1.2 – Structure de la double hélice d'ADN [1]

L'*acide désoxyribonucléique* (ou ADN) est un long polymère composé de monomères appelés *nucléotides*, ou *bases azotées*, que sont les adénines (A), thymines (T), guanines (G), et cytosines (C). L'ADN des organismes cellulaires est formé de deux brins *antiparallèles* enroulés en forme de double hélice (Figure 1.2). Dans un ADN

1. https://commons.wikimedia.org/wiki/File:DNA_structure_and_bases_FR.svg (consulté 27/04/2017)

double brin, l'adénine (A) s'apparie avec la thymine (T) au moyen de deux liaisons hydrogène, et la guanine (G) avec la cytosine (C) au moyen de trois liaisons hydrogène. Les deux brins sont ainsi dits *complémentaires* parce qu'il est possible de déduire un brin à partir de l'autre. C'est d'ailleurs cette redondance qui est exploitée par le mécanisme de réplication de l'ADN, au cours duquel une enzyme, l'*ADN polymérase*, synthétise un brin d'ADN complémentaire à partir d'un brin matrice. De plus, l'ADN est orienté, et cette synthèse se déroule toujours dans le sens 5' vers 3', soit de l'extrémité portant un groupe phosphate (5') vers l'extrémité portant un groupe hydroxyle (3').

Glossaire

ADN polymérase Enzyme responsable de la réplication de l'ADN.

réplication Processus de duplication de l'ADN par l'ADN polymérase qui utilise un brin matrice pour synthétiser son brin complémentaire.

sens 5' vers 3' Sens de synthèse de l'ADN, de l'extrémité portant un groupe phosphate (5') vers l'extrémité portant un groupe hydroxyle (3').

gène Région d'ADN transcrite en ARN et située sur un des deux brins.

génome ensemble du matériel génétique propre à un individu.

homopolymère Séquence continue du même nucléotide, généralement supérieure à 3 nucléotides (ex. : `AAAAAAAA`).

bp, kbp, Mbp, Gbp, Tbp On mesure la longueur d'une séquence d'ADN en nombre de nucléotides, ou paires de bases (*bp* pour *base pairs*). On utilise ensuite les préfixes du Système International d'unité (*kbp* pour kilobase, *Mbp* pour mégabase, *Gbp* pour gigabase et *Tbp* pour térabase).

1.1.4 L'ARN

L'*acide ribonucléique* (ARN) est une molécule simple brin similaire à l'ADN, composée des mêmes nucléotides à l'exception de la tyrosine (T) qui est remplacée par l'uracile (U) dans l'ARN. Chez tous les organismes cellulaires, un brin d'ARN est généré par la *transcription* d'une partie de l'ADN. On parle de l'expression d'un gène (ADN) sous forme de copie transcrite (*ARN messager*) traduite en *protéine* par le ribosome (1.1.5).

1.1.5 L'ARN ribosomique

Le ribosome est un complexe d'ARN et de protéines qui réalise la traduction des ARN messager en protéines, remplissant ainsi une fonction indispensable à l'ensemble

des organismes cellulaires connus. L'importance de cette fonction et son caractère universel font, en raison de la pression de sélection, que les ribosomes de toutes les espèces connues sont structurellement similaires, ce qui implique en particulier la conservation des séquences qui participent à sa composition.

Le ribosome est un complexe, ainsi composé de molécules d'ARN, appelées *ARN ribosomique* (ARNr), et de protéines organisées en deux sous-unités principales :

— la grande sous-unité, constituée de deux à trois molécules d'ARNr (5S, 28S et 5,8S chez les eucaryotes ; 23S et 5S chez les procaryotes) et de plusieurs dizaines de protéines ;

— la petite sous-unité, constituée d'une molécule d'ARNr (18S chez les eucaryotes, 16S chez les procaryotes) et de plusieurs dizaines de protéines.

1.1.5.1 L'ARNr de la petite sous-unité du ribosome (*ARNr SSU*)

L'ARNr de la petite sous-unité du ribosome (*ARNr SSU*, pour *Small Sub-Unit*) est considéré comme la référence pour l'assignation taxonomique des espèces d'organismes cellulaires. En plus d'être universel, il est constitué d'une succession de régions conservées et de régions variables qui en font le candidat idéal pour la construction d'arbres phylogénétiques basés sur des alignements de séquences (2.3.1). Les régions conservées correspondent à des structures indispensables à la fonction de traduction. D'un point de vue pratique, elles facilitent l'alignement des séquences et permettent la construction d'amorces « universelles » pour les approches de séquençage ciblé (1.2.3.1). Les régions hypervariables (au nombre de neuf chez les procaryotes) varient en taille de 30 à 100 bp (Figure 1.3) et correspondent à des régions moins contraintes dans la structure tertiaire du ribosome. Par conséquent, ces régions contiennent un signal phylogénétique suffisant pour discriminer les séquences d'espèces proches.

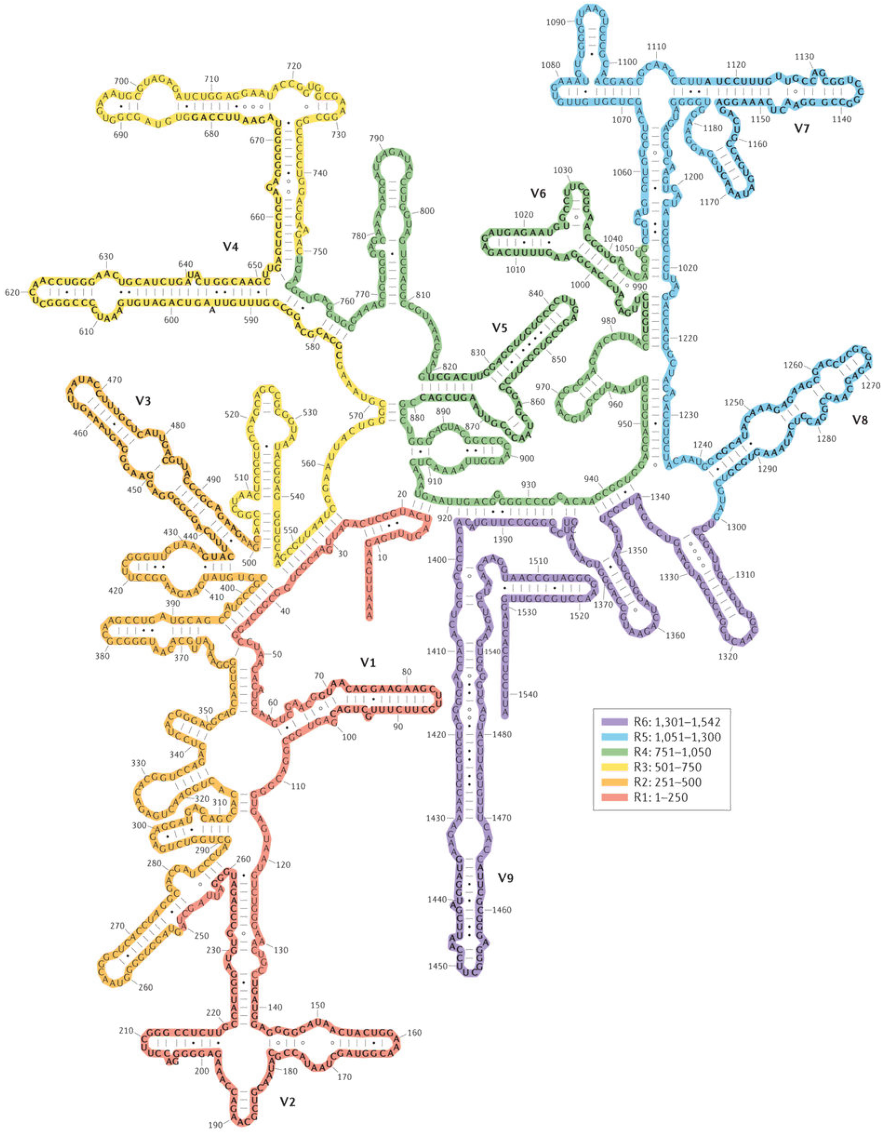


FIGURE 1.3 – Structure secondaire de l'ARNr 16S chez *Escherichia coli* [103]. Les 9 régions hypervariables (V1 à V9) sont indiquées en gras.

1.1.6 L'arbre du vivant

1.1.6.1 La classification taxonomique du vivant

La classification taxonomique du vivant consiste à catégoriser les organismes sur la base de critères scientifiques. Cette classification, ou *taxonomie*, organise des catégories d'organismes (*taxons*) de manière à retracer leur histoire évolutive. On décrit alors la phylogénie des organismes. Dans un *arbre phylogénétique*, les espèces les plus proches dérivent des ancêtres communs les plus récents et partagent le plus de caractères communs.

Nous présentons ici deux types de classifications taxonomiques, qui diffèrent prin-

cipalement par les critères utilisés pour la classification : critères physiques pour la *classification classique*, aujourd'hui obsolète pour la plupart des taxons, dont les micro-organismes, et critères biomoléculaires pour la *classification phylogénétique*.

Qu'est-ce qu'une espèce ?
La notion d'espèce est définie chez les eucaryotes pluricellulaires comme un groupe d'individus interféconds, pour qui la transmission du matériel génétique est dite « verticale », c'est-à-dire de génération en génération au sein d'une même espèce. Dans ce cas, les génomes des individus d'une même espèce sont plus proches génétiquement que ceux d'individus d'espèces différentes.
Ce concept d'espèce basé sur la transmission verticale est toutefois difficilement généralisable aux procaryotes qui peuvent aussi s'échanger du matériel génétique « horizontalement », c'est-à-dire entre individus d'espèces différentes, par les processus de *transformation*, *transduction* et *conjugaison*.
En l'absence de meilleure suggestion, il a été proposé en 2005 de définir une espèce procaryote comme un groupe d'individus partageant une similarité de séquence de l'ARNr 16S supérieure à 97% d'identité [37], et définissant ainsi un OTU (*Operational Taxonomic Unit*). Ce seuil de similarité arbitraire ne fait pas forcément de sens biologiquement, mais il est toujours utilisé dans les approches de métagénomique ciblée pour la construction d'OTU (2.3.1). Toutefois, à la suite des efforts de séquençage de la biodiversité inconnue ces dernières années, cette définition commence maintenant à être remise en question [66] et est appelée à évoluer.

La classification classique Apparue au XVIII[e] siècle et initiée par le naturaliste Carl von Linné, la classification classique propose d'organiser les êtres vivants sous la forme d'un arbre à sept niveaux (ou rangs de *taxons*) successifs, dont l'espèce est l'unité de base, puis faisant intervenir le genre, la famille, l'ordre, la classe, l'embranchement, le règne et le domaine (Figure 1.4). Les critères utilisés pour la catégorisation des espèces reposent principalement sur la présence ou l'absence de caractères physiques (présence d'organes, formes des organes et des organismes, etc.).

Cette classification basée sur les caractères visibles bute toutefois sur l'horizon eucaryote/procaryote et classe tous les procaryotes dans un même domaine (voir ci-après 1.1.6.2).

La classification phylogénétique Le développement de la biologie moléculaire et de la génétique dans la deuxième moitié du XX[e] siècle a introduit un changement de

2. https://commons.wikimedia.org/wiki/File:Taxonomic_hierarchy.svg (consulté le 13/07/2017)

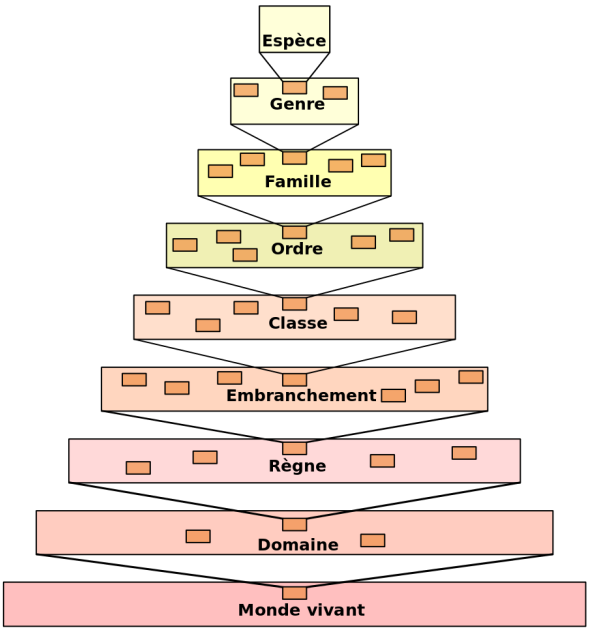


FIGURE 1.4 – Hiérarchie du vivant selon la classification classique [2]

paradigme dans le domaine de la taxonomie, et a mené à la création de la classification phylogénétique. C'est notamment le microbiologiste Carl Woese, qui en 1977 sépare les procaryotes en deux domaines distincts sur la base de l'analyse phylogénétique de l'ARNr 16S [99].

Cette classification repose sur la comparaison de caractères biomoléculaires, au premier rang desquels le matériel génétique, et vise à reconstruire l'histoire évolutive des organismes et leurs liens de parenté sur cette base. Par conséquent, un arbre phylogénétique ne possède pas un nombre de niveaux fixé, comme dans la classification de von Linné. On définit alors un *clade* comme un ensemble d'espèces qui partagent un ancêtre commun.

En pratique, les classifications phylogénétiques modernes reposent sur la comparaison de marqueurs conservés dans le génome des organismes. Et c'est notamment l'ARNr SSU (16S pour les procaryotes, 18S pour les eucaryotes) (1.1.5) qui est utilisé en raison de son caractère d'universalité au sein de l'arbre du vivant. L'enregistrement d'une nouvelle espèce dans les bases de données requiert d'ailleurs la séquence 16S/18S correspondante. Un exemple de classification phylogénétique récente est présenté dans la Figure 1.5 à la section 1.1.6.2.

Choisir sa classification La classification classique est maintenant considérée comme obsolète, mais les classifications phylogénétiques modernes conservent parfois encore certaines de ses caractéristiques. Notamment, il est courant de trouver des taxonomies hybrides dans lesquelles des taxons phylogénétiques raffinent une structure basée sur

les taxons classiques. Par exemple, un ensemble de « genres » d'une même « famille » peuvent être identifiés phylogénétiquement comme ne partageant pas un ancêtre commun, et cette « famille », bien que mentionnée dans une taxonomie hybride, n'a pas de sens historique.

Dès lors que l'on travaille sur des questions liées à la biologie environnementale, il est donc indispensable de préciser la taxonomie que l'on utilise. Par exemple, la taxonomie du NCBI [3] n'est pas la même que celle de SILVA [75], ou encore celle de Greengenes [19]. Le « Tree of Life Web Project » (ToL) [96] est un exemple d'une des taxonomies phylogénétiques les plus complètes. C'est aussi l'une des plus difficiles à utiliser parce que très complexe et continuellement mise à jour.

Un grand nombre d'institutions et la majorité de la communauté des biologistes préfèrent par conséquent utiliser des taxonomies partiellement fausses, car n'intégrant pas les résultats les plus récents concernant les parentés des clades, plutôt que de les modifier trop souvent. Ces modifications régulières n'ont en fait que peu d'intérêt quand on s'intéresse aux organismes déjà connus.

Pour le reste de ce manuscrit, et en particulier dans le chapitre III, nous utiliserons exclusivement la taxonomie SILVA, basée sur la comparaison des séquences d'ARN ribosomique, et disponible ici : https://www.arb-silva.de/

3. https://www.ncbi.nlm.nih.gov/taxonomy

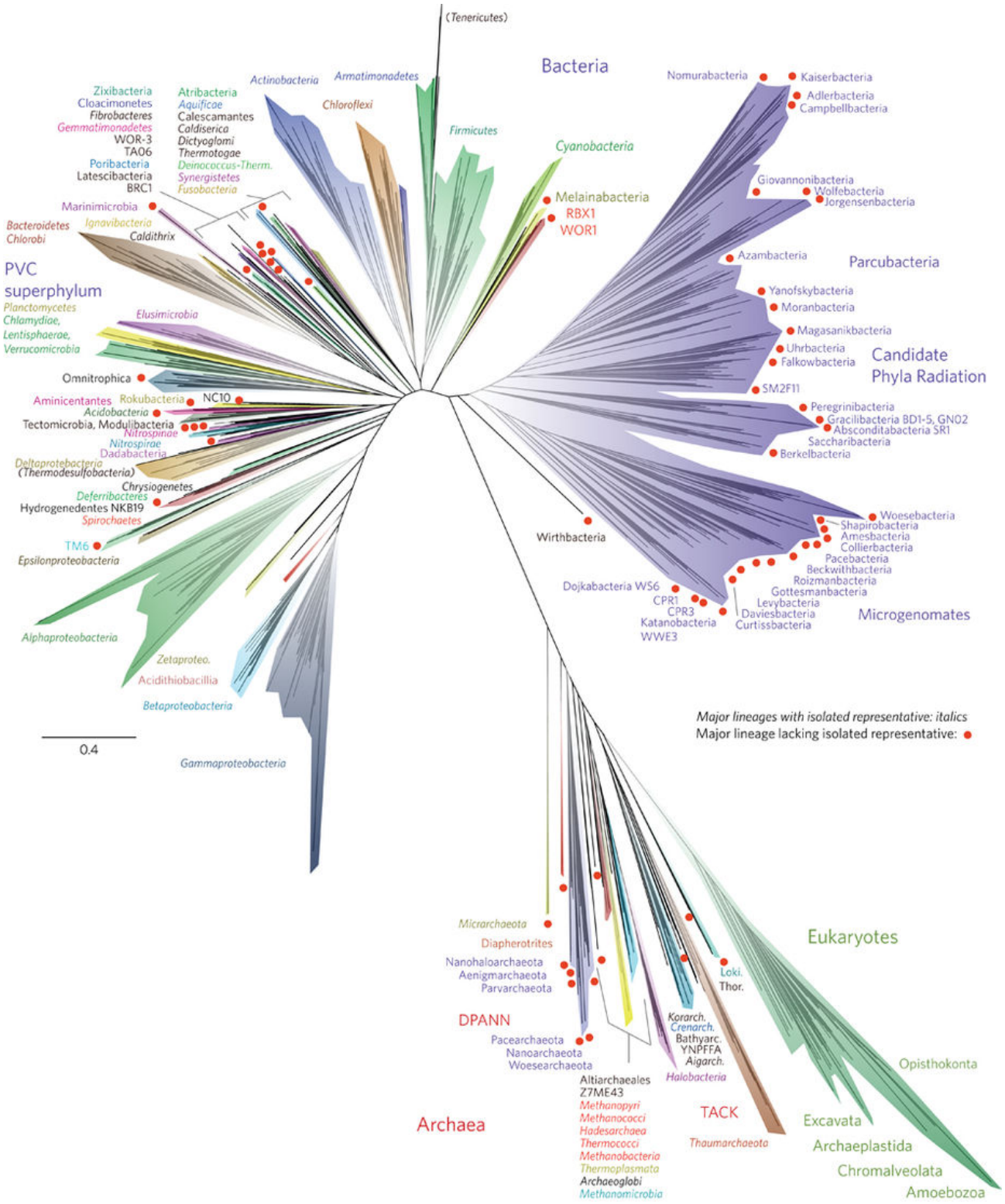


FIGURE 1.5 – Une vision récente de l'arbre du vivant [34]. La phylogénie présente une grande partie de la diversité connue et mentionne les noms des clades principaux les plus divergés. Cet arbre propose l'identification d'un super-clade bactérien constitué exclusivement de bactéries non cultivables. Ces travaux illustrent à quel point l'identification et la classification de la biodiversité inconnue sont susceptibles de transformer radicalement les conceptions de la taxonomie et de l'histoire du vivant.

1.1.6.2 Les domaines du vivant

Depuis le changement de paradigme introduit par Carl Woese en 1977 (1.1.6.1), l'arbre du vivant est organisé en trois grands domaines : les eucaryotes (*Eucaryota*), les bactéries (*Bacteria*), et les archées (*Archaea*) [99, 100]. Comme vu précédemment, on regroupe sous le terme de procaryotes les bactéries et les archées pour des raisons historiques et par souci de simplification du langage (notamment parce que les organismes de ces deux domaines possèdent des ARNr 16S).

Les eucaryotes Les eucaryotes représentent tous les êtres vivants visibles à l'œil nu. Il s'agit des animaux, des plantes, des champignons, mais aussi d'un ensemble d'autres organismes pluricellulaires comme les algues rouges et brunes, ou d'organismes unicellulaires aussi nommés *protistes* dans les premières taxonomies, comme les amibes ou encore les radiolaires (qui font partie du zooplancton). Les eucaryotes sont caractérisés par la présence dans leurs cellules d'un vrai noyau (*eu* « vrai » et *karuon* « noyau »), ainsi qu'un ensemble de compartiments spécialisés appelés *organites*.

Les bactéries Les bactéries sont regroupées au sein du domaine le plus ancien des trois domaines du vivant. Elles sont extrêmement abondantes et diversifiées. Il s'agit exclusivement d'organismes unicellulaires sans noyau. Les bactéries sont présentes dans tous les environnements terrestres. Par exemple, les bactéries des genres *Bacteroides* et *Prevotella* (super-clade *Bacteroidetes/Chlorobi*) ainsi que du genre *Ruminococcus* (super-clade des *Firmicutes*) colonisent les intestins humains (1.1.7.3). Leur taille peut varier de 0,05µm à plus de 2µm. Elles peuvent prendre de très nombreuses formes et réaliser un très grand nombre de fonctions métaboliques différentes.

Les archées Les archées sont des micro-organismes unicellulaires sans noyau. Initialement considérées comme des bactéries extrêmophiles, les archées sont classées comme un domaine à part à la suite des travaux de Carl Woese en 1977 [99]. Si les premières archées ont bel et bien été identifiées dans des environnements exotiques comme les sources chaudes volcaniques ou les lacs salés, les récents projets d'analyse des microbiotes environnementaux ont révélé la présence d'une très grande quantité d'espèces d'archées non cultivables dans la plupart des milieux (sol, eau de mer, flore intestinale, etc.). D'un point de vue génétique et biomoléculaire, elles sont aussi différentes des bactéries que des eucaryotes [10].

La question de l'enracinement de l'arbre du vivant Historiquement, archées et bactéries formaient un clade, et il était ainsi proposé que la racine de l'arbre du vivant se situe avant la division de cet arbre en procaryotes et eucaryotes. Toutefois, en 2008, l'étude de 45 gènes universels change cette vision et propose une organisation de l'arbre du vivant à deux domaines que seraient les bactéries, et un domaine regroupant archées

et eucaryotes [15]. Cette vision est maintenant d'autant plus renforcée qu'il semblerait que les eucaryotes aient en fait évolué à partir des archées [16].

1.1.7 Les enjeux

1.1.7.1 Dans les différents domaines de la recherche ou de l'industrie

L'étude des environnements complexes et de la biodiversité inconnue est un enjeu actuel majeur dans de nombreux domaines, et les possibilités ouvertes par une meilleure compréhension des microbiotes sont nombreuses :

— en santé, humaine et animale, l'importance du microbiote commence à être identifiée, au point de le considérer comme un nouvel organe [5]. Son étude pourra ainsi permettre de mieux comprendre l'impact de ce microbiote sur le maintien en bonne santé, ou l'apparition de maladies ;

— pour l'industrie, l'analyse massive des nouveaux gènes révélés au sein de cette diversité inconnue permettra la découverte de nouvelles protéines et molécules qui pourraient potentiellement être utilisées comme médicaments ou catalyseurs ;

— en écologie, où l'on cherche à avoir une meilleure compréhension des microbiotes environnementaux, de leur évolution dans le temps et des interactions entre organismes, ces connaissances nouvelles permettront entre autres de comprendre l'évolution d'un environnement et l'impact de facteurs externes (pollution, réchauffement climatique, etc.) sur ces environnements ;

— finalement, en recherche fondamentale, cela permettra de mieux comprendre l'organisation de l'arbre du vivant et de se rapprocher d'une compréhension fine des mécanismes de l'histoire de la vie. Par exemple, l'analyse de sédiments océaniques en 2015 a permis l'identification d'un nouvel embranchement des archées, les *Lokiarchaeota*, qui pourrait permettre de mieux comprendre l'émergence des eucaryotes [92].

Pour mieux comprendre les enjeux importants de ce domaine d'étude, nous allons les illustrer au travers de projets de recherche majeurs, tous pilotés au sein de *consortia* de recherche qui regroupent des laboratoires et chercheurs du monde entier.

1.1.7.2 Exploration marine : Tara Océans

D'initiative française, le projet Tara Océans est piloté par la fondation Tara Expéditions. Celle-ci met à la disposition de la communauté scientifique internationale un bateau, la goélette Tara, pour réaliser des expéditions d'exploration et d'étude des océans du monde entier. Le projet Tara Océans s'est déroulé de 2009 à 2013 et correspond aux 8^{e} et 9^{e} expéditions qui ont permis de réaliser 50 escales sur l'ensemble des océans de la planète (Figure 1.6).

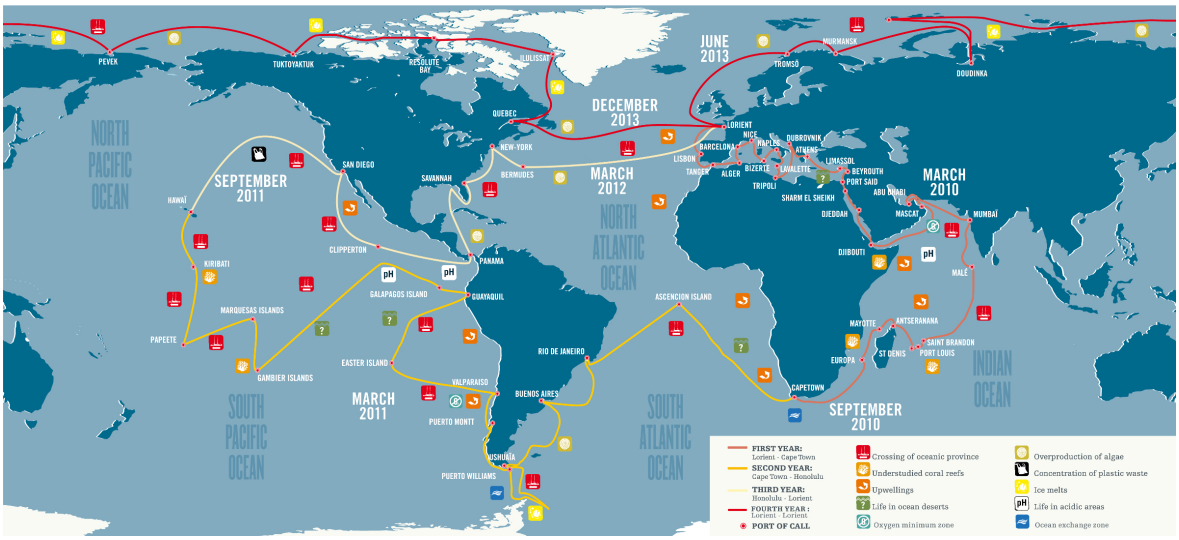


FIGURE 1.6 – Carte des expéditions Tara Océans

L'objectif du projet Tara Océans est la caractérisation la plus complète possible des océans de notre planète, qui sont encore très mal connus. Pour cela, à chaque escale, un ensemble d'analyses et de prélèvements ont été réalisés (plus de 35 000 au total) :

— l'enregistrement d'un ensemble de paramètres physico-chimiques depuis la surface jusqu'à des profondeurs de 2000 m ;

— le déploiement de filets avec des mailles de tailles variables, dont un filet spécialisé dans la récolte des plastiques de surface ;

— le prélèvement des micro-organismes présents entre 10m et 120m de profondeur. L'utilisation de tamis et filtres successifs a permis la séparation de ces microbiotes en fractions de tailles correspondant respectivement au zooplancton, aux protistes (dont le phytoplancton), aux bactéries, et aux virus ;

— la prise de vidéos *in situ*, ainsi que la prise d'images pour les fractions de tailles visibles au microscope optique.

Le séquençage des échantillons microbiens (que nous présentons dans la section 1.3.4.1) et l'analyse postérieure des prélèvements ont ainsi permis de dresser une première image de référence de l'état écologique et géophysique des océans de la planète au moment de ces deux expéditions [94]. Il s'agit du premier inventaire métagénomique à l'échelle planétaire. Des projets ultérieurs, comme le projet Investissements d'Avenir Oceanomics [4], pourront ensuite permettre de suivre l'évolution de ces écosystèmes, par exemple en étudiant l'impact des changements climatiques.

4. http://www.oceanomics.eu/

1.1.7.3 Analyse du microbiote humain

L'analyse du microbiote humain est un sujet majeur de recherche depuis la fin des années 2000. Notamment, deux projets d'étude du microbiote humain à grande échelle ont démarré en 2008 : MetaHIT [5], un projet européen focalisé sur l'étude du microbiote intestinal et son rôle dans diverses pathologies ; et le *Human Microbiome Project* (*HMP*) [6], un projet américain visant à caractériser de manière la plus exhaustive possible les microbiotes de multiples sites corporels sur un groupe de plusieurs centaines d'individus sains. Plus récemment, des projets comme l'*American Gut Project* [7] et le *British Gut Project* [8] visent à caractériser les microbiotes d'un panel d'individus le plus large et le plus varié possible (plus de 10 000 individus) afin d'explorer la diversité de ce microbiote et d'identifier des corrélations entre ce microbiote et la santé des individus.

Nous présentons brièvement ci-après les projets MetaHIT et HMP, leurs enjeux respectifs et les résultats scientifiques qu'ils ont déjà permis d'obtenir.

MetaHIT MetaHIT est un projet européen qui s'est déroulé de 2008 à 2012, et qui a regroupé des partenaires académiques et industriels de huit pays. Focalisé sur l'étude du microbiote intestinal humain, l'objectif principal de ce projet était d'établir des associations entre les gènes de ce microbiote et la santé et pathologies humaines. Plus particulièrement, le projet s'est intéressé à deux pathologies majeures : la maladie inflammatoire chronique de l'intestin (MICI), et l'obésité. Pour cela, l'analyse du microbiote provenant d'échantillons fécaux a été réalisée pour 124 individus européens.

Le projet MetaHIT a généré deux résultats majeurs :

— la création d'un catalogue de 3,3 millions de gènes bactériens présents dans le microbiote intestinal humain [74] ;

— la découverte d'une partition de la population mondiale en 3 groupes. Chaque groupe est caractérisé par une population microbienne intestinale prédominante différente (*Bacteroides*, *Prevotella* et *Ruminococcus*). On parle d'entérotypes intestinaux [3].

De plus, ces résultats ont ouvert de nombreuses perspectives, que ce soit dans la détection précoce de certaines maladies (maladie de Crohn, obésité), ou encore la mise au point d'une nutrition et d'une médecine personnalisée en fonction des organismes présents dans le microbiote intestinal.

Human Microbiome Project Le *Human Microbiome Project* (*HMP*) est un projet américain financé par le fonds commun du *National Institutes of Health* (NIH) et initié en 2008. Son principal objectif est la caractérisation la plus exhaustive possible des

5. http://www.metahit.eu
6. http://hmpdacc.org/
7. http://americangut.org/
8. http://britishgut.org/

microbiotes humains et l'analyse de leurs rôles dans la santé humaine. Contrairement au projet MetaHIT qui s'est limité à l'étude du microbiote intestinal, le projet HMP a réalisé l'analyse de 15 à 18 sites corporels différents (bouche, nez, peau, intestin, etc.) pour 300 adultes sains.

L'analyse des données du projet HMP est encore en cours en 2017, mais le consortium HMP a publié deux articles majeurs en 2012 [14, 59] qui présentent les résultats suivants :

— l'isolement et la culture de 1300 souches bactériennes de référence, qui ont été séquencées ;

— le séquençage ciblé (1.3.1) et le séquençage métagénomique complet (1.3.2) de plus de 11 000 échantillons, ainsi qu'une première analyse taxonomique de chacun de ces échantillons. Ces analyses ont ainsi permis de générer une estimation de la complexité des communautés microbiennes pour chaque site corporel.

Pour l'évaluation de notre méthode, présentée dans le chapitre III de ce manuscrit, nous avons utilisé des jeux de données issus du projet HMP. Nous décrivons dans la section 1.3.4.2 les différents jeux de données générés par ce projet. Puis, dans le chapitre IV, section 4.4, nous présentons les résultats que nous avons obtenus sur ces jeux de données.

1.2 Les technologies de séquençage de l'ADN

Nous avons vu dans la section précédente que l'analyse du matériel génétique était devenue une étape primordiale pour comprendre un microbiote et en déterminer la composition taxonomique. L'acquisition de ce matériel passe par le *séquençage*, terme qui désigne le processus permettant de lire les séquences d'ADN. Pour cela, plusieurs procédés existent, qui ont progressé au fil du temps. Nous en faisons une présentation rapide.

1.2.1 Définitions

Le *séquençage* de l'ADN consiste à déterminer l'ordre des nucléotides d'un fragment d'ADN. On appelle *lecture de séquençage* la séquence d'un fragment d'ADN lue par un séquenceur.

Dans le cadre du séquençage à haut débit (1.2.3), on appelle *librairie* l'ensemble des fragments d'ADN qui sont issus de la préparation biomoléculaire du matériel génétique, obtenus avec plusieurs stratégies possibles (1.2.3.1).

Un *run de séquençage* est l'expérience qui consiste à séquencer une librairie de séquençage donnée sur le même séquenceur, en une seule fois. Lors de ce séquençage,

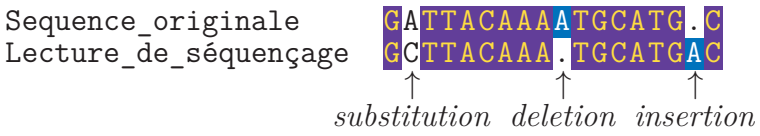


FIGURE 1.7 – Les 3 types d'erreurs de séquençage

peuvent se produire des *erreurs de séquençage*, c'est-à-dire l'apparition d'une différence entre la séquence d'une lecture et la séquence originale du fragment d'ADN. Les erreurs de séquençage peuvent être du type *substitution* (transformation d'un nucléotide en un autre dans la lecture de séquençage), ou du type *indel* (*insertion* ou *délétion* d'un nucléotide dans la lecture) (Figure 1.7). Les lectures sont accompagnées de scores de *qualité* estimés par le séquenceur, qui associent à chaque nucléotide lu une valeur Q, exprimée en *Phred*, représentant la probabilité p d'avoir identifié le mauvais nucléotide par erreur, avec $p = 10^{-Q/10}$.

1.2.2 Perspective historique

Les premières méthodes de séquençage ont été inventées au début des années 1970 (Wu et Cornell, 1970 ; Gilbert et Maxam, 1973), mais c'est la méthode par synthèse enzymatique inventée par Frederick Sanger en 1977 qui posa les bases technologiques du séquençage moderne [79].

La méthode de Sanger, commercialisée en premier par Applied Biosystems, a été la méthode de séquençage la plus utilisée pendant près de 40 ans. Inspirée par le processus naturel de réplication de l'ADN (1.1.3), elle repose sur la polymérisation contrôlée *ex vivo* de simples brins d'ADN par des enzymes ADN polymérases. Grâce à l'introduction de nucléotides modifiés (didésoxyribonucléotides), la polymérisation se termine prématurément de manière aléatoire. La séquence complète peut être ensuite lue sur un gel ayant servi à faire migrer les produits de la polymérisation selon leur taille.

Cette méthode de séquençage produit des lectures d'environ 700 bp en moyenne, avec un très faible taux d'erreur (entre 0,001% et 1%) [29]. Toutefois, son débit limité (∼100 kbp/h pour les technologies les plus rapides, en 2008) la rend difficilement applicable dans le cadre du séquençage de génomes, qui nécessite alors l'utilisation de stratégies de séquençage par ordonnancement hiérarchique[9]. C'est cette approche qui a été initialement utilisée pour le séquençage du premier génome humain en 2001 [40].

9. https://fr.wikipedia.org/wiki/Séquençage_de_l'ADN#Ordonnancement_hi.C3.A9rarchique

1.2.3 Le séquençage à haut débit

Les premières technologies de séquençage à haut débit dites de « nouvelle génération » (ou NGS pour *Next Generation Sequencing*) sont apparues à la fin des années 1990. Ces technologies automatisées et massivement parallèles permettent de produire des millions de lectures par *run* de séquençage, à faible coût.

Une expérience de séquençage à haut débit commence par l'extraction du matériel génétique d'un échantillon biologique. Ensuite, en fonction de la question biologique posée, plusieurs choix sont possibles : stratégies de construction des librairies (ciblée ou globale), type de séquençage (obtention de lectures simples, appariées, ou *mate-pairs*), et enfin plateforme de séquençage.

1.2.3.1 Approche ciblée ou globale

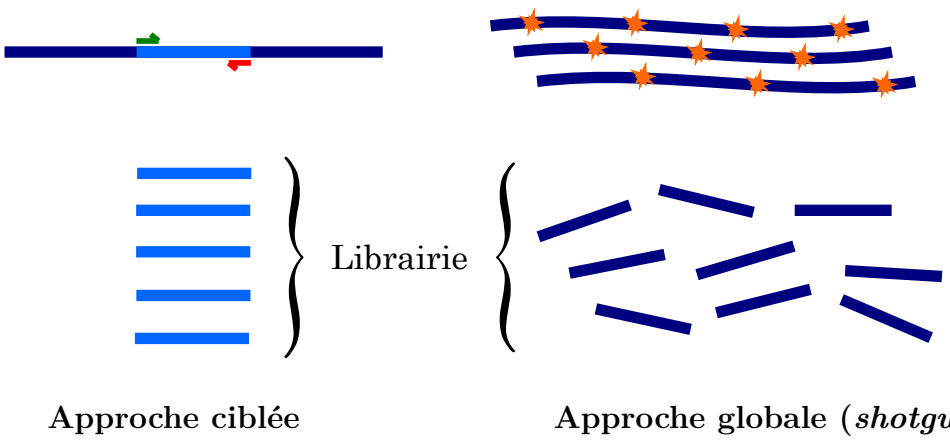


FIGURE 1.8 – Principe du séquençage par approche ciblée et par approche globale (*shotgun*).

Le choix entre approche ciblée et approche globale (*shotgun*) va principalement dépendre de la taille de la région que l'on souhaite étudier (Figure 1.8).

Séquençage ciblé Pour séquencer une courte région d'intérêt de quelques dizaines à quelques milliers de nucléotides, on réalise une étape d'amplification *PCR* (*Polymerase Chain Reaction*) préalable au séquençage. Cette méthode biomoléculaire très classique s'inspire du processus naturel de réplication de l'ADN (1.1.3) et utilise l'ADN polymérase pour répliquer une région délimitée par des séquences connues (appelés *amorces*).

L'avantage principal de cette méthode est sa capacité à multiplier le nombre de copies de la région d'intérêt, permettant ainsi son séquençage même lorsque peu de matériel génétique est initialement disponible. Toutefois, si plusieurs copies similaires de la région d'intérêt existent dans le génome, il y a possiblement des artefacts appelés *biais d'amplification PCR*. L'étape d'amplification peut alors d'une part générer

des *séquences chimériques*, dont une partie correspond à une copie de la région d'intérêt et l'autre partie correspond à une autre copie. D'autre part, certaines copies de la région d'intérêt risquent d'être davantage répliquées que d'autres, on parle alors d'*amplification préférentielle*.

Méthode globale, dite *shotgun* Pour séquencer un génome complet, la méthode globale consiste à fragmenter de manière aléatoire ce génome en de très nombreux *fragments* de quelques dizaines à plusieurs centaines de paires de bases. Cette méthode repose sur la redondance du matériel génétique et nécessite de grandes quantités initiales de matériel génétique dans l'échantillon biologique.

On parle de *couverture de séquençage* pour désigner le nombre de copies identiques d'un génome séquencé. En pratique, on peut estimer la couverture moyenne par le rapport de la taille totale des lectures sur la taille attendue du génome. Cette estimation fait l'hypothèse que le séquençage est homogène le long du génome, ce qui n'est toutefois pas toujours le cas à cause de *biais de séquençage préférentiel*.

1.2.3.2 Lectures simples et appariées

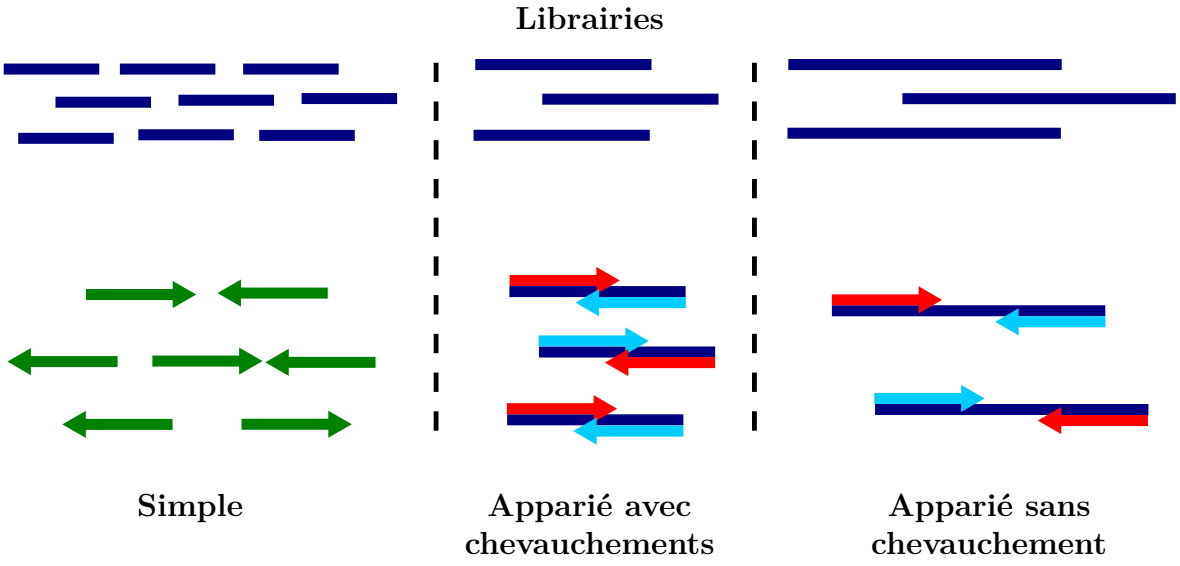


FIGURE 1.9 – Stratégies de séquençage simple et apparié

Une librairie donnée peut ensuite être séquencée de plusieurs manières (Figure 1.9) :

— un séquençage simple, au cours duquel les fragments sont lus dans toute leur longueur. Chaque lecture correspond donc à un fragment de la librairie ;

— un séquençage apparié, au cours duquel chaque fragment est lu deux fois, à partir de l'extrémité 5' et à partir de l'extrémité 3', et génère des *lectures appariées*.

Cette stratégie de séquençage est pertinente pour séquencer des fragments de taille supérieure à la taille des lectures. De plus, si la taille du fragment est inférieure à deux fois la taille des lectures, les lectures seront *chevauchantes*. En pratique les tailles des fragments séquencés par cette stratégie sont de l'ordre de quelques centaines de nucléotides ;

— une troisième stratégie de séquençage dite *mate-pair* permet de séquencer les extrémités 5' et 3' de fragments bien plus grands, de l'ordre de plusieurs kilobases, et nécessite une préparation spécifique de la librairie.

1.2.3.3 Les plateformes de séquençage à haut débit (HTS)

TABLE 1.1 – Description des principales technologies de séquençage à haut débit [10]. Les données indiquées dans ce tableau sont à titre consultatif et représentent l'état des technologies en 2016-2017. Parce que les technologies de séquençage progressent vite, elles sont susceptibles d'évoluer rapidement.

Technologie	**Longueur lectures**	**Précision**	**Nb de lectures par run**	**Coût ($US/Mbp)**
Sanger	400-900 bp	99,9%	N/A	2400 $
Illumina	MiSeq : 50-600 bp HiSeq : 50-500 bp	99,9%	MiSeq : 1-25M HiSeq : 0,3-2 G	0,05-0,15 $
Roche-454	700 bp	99,9%	1 M	10 $
Ion Torrent	jusqu'à 400 bp	98%	jusqu'à 80 M	1 $
Pacific Biosciences	10-15 kbp moy max > 40 kbp	87%	500-1000 M	0,13-0,60 $
Nanopore	jusqu'à 500 kbp	~92-97%	variable	~0,5 $

Séquençage de deuxième génération Parmi les plateformes HTS dites de « deuxième génération », les plus notables sont les technologies Illumina et Roche-454 (Table 1.1). Chacune de ces technologies utilise une approche physico-chimique différente et possède donc des avantages et défauts qui lui sont propres. On peut les comparer en matière de débit, de coût, de longueur de lecture et de taux d'erreur.

La technologie Illumina permet de générer jusqu'à plusieurs milliards de lectures avec un très faible taux d'erreur (de l'ordre de 0,1%), et les erreurs sont exclusivement du type substitution. De plus, la taille des lectures est fixée pour un *run* de séquençage

10. Adapté de https://en.wikipedia.org/wiki/DNA_sequencing#High-throughput_methods (consulté le 05/05/2017).

donné. Toutefois, en pratique, afin de maintenir un faible taux d'erreur, les séquençages Illumina sont habituellement paramétrés pour générer des tailles de lecture entre 100 bp et 250 bp. Avec un coût de l'ordre de 0,1 $US/Mbp, la technologie Illumina est la technologie la moins chère sur le marché en 2017 et reste la technologie la plus utilisée pour les projets de séquençage actuels.

La technologie Roche-454, qui n'est plus développée depuis 2013, permet quant à elle de produire des lectures entre 400 bp et 500 bp en moyenne, avec un faible taux d'erreur. Toutefois, le débit de cette technologie reste limité (1 Mbp par *run*) et le coût élevé (10 $/Mbp). De plus, le principal défaut de la technologie Roche-454 est sa propension à générer des erreurs de type *indels* dans les homopolymères (1.1.3), ce qui complexifie l'assemblage postérieur des lectures (2.2)

Séquençage de troisième génération Depuis la fin des années 2000, une nouvelle vague de technologies de séquençage à haut débit dites de « troisième génération » est arrivée sur le marché. Ces plateformes, comme Oxford Nanopore (ONT) et Pacific Biosciences (PACBIO), permettent de séquencer de longs fragments d'ADN (supérieurs à 10 kbp) d'un seul tenant, avec un débit raisonnable (encore inférieur toutefois aux technologies de seconde génération). Le principal défaut des technologies de troisième génération est leur taux d'erreur encore élevé (de l'ordre de 10%).

1.3 La métagénomique

Dans la dernière section de ce chapitre, les microbiotes rencontrent le séquençage à haut débit. Nous présentons les stratégies de séquençage utilisées dans le cadre de l'analyse taxonomique des microbiotes et la description des données ainsi générées. Ces données sont au cœur de cette thèse, et nous expliquerons dans le chapitre II quelles sont les méthodes bio-informatiques nécessaires pour leur analyse.

1.3.1 Le séquençage ciblé (amplicon ARNr SSU)

Dans le cadre de la métagénomique, le séquençage ciblé (1.2.3.1), et plus particulièrement celui de l'ARNr SSU, a pour objectif de permettre l'identification rapide et peu coûteuse des espèces microbiennes présentes dans un échantillon biologique complexe. Cette approche se déroule en trois étapes :

— une ou plusieurs régions hypervariables contiguës sont sélectionnées (1.1.5.1) et des amorces « universelles » sont définies ;

— la séquence cible est amplifiée par PCR à partir du matériel génétique extrait de l'échantillon. Les amorces « universelles » permettent d'amplifier cette séquence chez tous les organismes présents dans l'échantillon ;

— le produit d'amplification, nommé *amplicon*, est séquencé à l'aide d'une technologie de séquençage de seconde génération.

Puisqu'en pratique les régions amplifiées mesurent généralement quelques centaines de paires de bases (∼400 bp dans la majorité des cas), c'est la technologie de séquençage 454 qui a longtemps été employée pour cette étape. Avec le développement de la technologie Illumina, il est maintenant possible (et courant) de séquencer des amplicons de 400 bp avec un Illumina MiSeq et une librairie appariée de deux fois 250 bp.

Cette approche possède comme principal avantage d'être facile à mettre en œuvre, rapide et peu coûteuse. De plus, grâce à l'étape d'amplification, elle permet le séquençage d'organismes très peu abondants. En contrepartie, cette approche comporte un grand nombre de biais. On peut citer :

— la *non-universalité des amorces*. Le nombre d'espèces connues augmentant, des études récentes ont montré que les régions dites « conservées » sont moins partagées que ce qui était précédemment accepté [55]. Par conséquent, des séquences d'espèces encore inconnues suffisamment divergentes pourraient ne pas être reconnues par les amorces, et ne seraient pas amplifiées. Pour pallier ce problème, il est possible de multiplier les amorces utilisées (*amorces dégénérées*) [49], ce qui augmente aussi le coût du séquençage ;

— la création de *chimères d'amplification PCR* (section 1.2.3.1). L'étape d'amplification risque de générer des lectures dont une partie correspond à une espèce et l'autre partie correspond à une autre espèce ;

— enfin, *l'amplification préférentielle* de certaines séquences due aux biais d'amplification PCR [76] (section 1.2.3.1) peut fausser la quantification des espèces présentes dans l'échantillon.

Ainsi, avec l'approche amplicon, on risque de séquencer plutôt des espèces déjà connues en biaisant leur abondance dans l'échantillon, et de créer des lectures chimériques ne correspondant à aucune réalité biologique. Ces biais nuisent à une analyse taxonomique précise et sans *a priori* des échantillons environnementaux.

De plus, la faible longueur des lectures obtenues rend leur identification difficile. On parlera d'un manque de signal phylogénétique : le faible nombre de sites informatifs génère un fort risque d'erreur stochastique lors de l'attribution de la lecture à un taxon.

1.3.2 Le séquençage métagénomique complet (*shotgun*)

Dans le cadre de la stratégie *shotgun* couplée aux technologies de séquençage de deuxième et troisième génération, le matériel génétique est extrait de l'échantillon, puis fragmenté et séquencé tel quel (1.2.3.1). Cette approche offre plusieurs avantages importants par rapport au séquençage par amplicons. Elle permet de s'affranchir de l'étape d'amplification PCR et des biais associés. Elle produit en outre une quantité

d'information supérieure. Au lieu de se focaliser seulement sur une séquence de quelques milliers de nucléotides, un séquençage métagénomique complet fournit des lectures pour l'ensemble des génomes. Tout cela explique que le séquençage complet apparaisse de plus en plus comme une alternative souhaitable au séquençage par amplicon.

Cependant, un séquençage complet est, par nature, plus coûteux qu'un séquençage ciblé. C'est d'autant plus vrai que pour pouvoir séquencer les espèces de faible abondance, on a besoin d'importantes quantités de matériel génétique afin d'obtenir de plus grandes couvertures de séquençage, ce qui entraîne un coût proportionnel à la couverture.

Enfin, le principal goulot d'étranglement à la généralisation de projets de séquençages métagénomiques complets reste le manque crucial d'outils et de méthodes pour analyser ces données. Nous détaillerons dans le chapitre II les développements réalisés ces dernières années pour proposer de nouveaux outils capables d'analyser de telles données, et nous présentons dans ce manuscrit une nouvelle méthode pour l'assignation taxonomique des lectures de séquençages métagénomiques complets.

1.3.3 Les technologies de séquençage en métagénomique

TABLE 1.2 – Nombre d'entrées métagénomiques sur SRA en fonction des technologies de séquençage (au 09/05/2017).

Nb entrées SRA	Total	Illumina	454	IonTorrent	Autre
Amplicon	372 705	213 486	139 388	17 348	2 483
Métagenome complet	138 202	125 563	10 525	1 435	679
RNA-Seq	12 146	8 115	3 794	144	93

Que ce soit dans le cadre des analyses amplicons ou de séquençages métagénomiques complets, c'est la technologie Illumina qui est majoritairement employée par les projets actuels de métagénomique. Une interrogation de la base de jeux de données de séquençages SRA [11] [42] indique que plus de 57% des jeux de données d'amplicons et plus de 90% des jeux de données métagénomiques complets ont été générés par la technologie Illumina. Cette adhésion massive peut s'expliquer par le faible coût de séquençage et le très faible taux d'erreur de cette technologie, mais aussi par le manque d'outils bio-informatiques capables de travailler sur de longues lectures avec beaucoup d'erreurs (3^{e} génération).

11. https://www.ncbi.nlm.nih.gov/sra

1.3.4 Exemples d'applications en métagénomique

1.3.4.1 Retour sur le projet Tara Océans

Nous avons décrit dans la section 1.1.7.2 la nature et les objectifs du projet Tara Océans. Comment cela se traduit-il en matière de données métagénomiques ? Plus de 35 000 prélèvements ont été collectés, chacun correspondant à une position géographique, une profondeur, et une fraction de taille (zooplancton, protistes, bactéries, et virus). Les prélèvements ont ensuite été mis à disposition de 23 laboratoires et instituts scientifiques autour du monde pour analyse.

Différentes stratégies d'analyse ont été mises en place sur tout ou partie des échantillons. Pour le séquençage, on peut notamment citer :

— le séquençage systématique de la région V9 de l'ARNr 18S et son analyse bioinformatique par des approches de partitionnement (2.3.1.1) pour caractériser la diversité du plancton eucaryote [97]. Le même type de stratégie a aussi été mis en place pour le séquençage de l'ARNr 16S des procaryotes ;

— le séquençage métagénomique complet de 243 échantillons (plateforme Illumina, lectures appariées) pour un total de 7,2 Tbp. L'assemblage et l'analyse de ces jeux de données ont permis la création d'un catalogue de plus de 40 millions de gènes de référence d'eucaryotes, de procaryotes et de virus [94] ;

— le séquençage de l'ARN total d'une partie des échantillons.

Étant donnée la quantité d'échantillons récoltés par les expéditions Tara Océans, le séquençage et l'analyse de ces derniers se poursuit encore aujourd'hui. Par ailleurs, certains échantillons ont été prélevés dans des zones pauvres en micro-organismes, et la rareté du matériel biologique limite parfois la profondeur possible des différents séquençages.

1.3.4.2 Retour sur HMP

Dans le cadre du projet HMP, une cohorte de 300 individus sains a été échantillonnée pour cinq sites corporels (bouche, nez, peau, intestin, vagin), subdivisés en 15 à 18 sous-sites corporels. Chaque individu a été échantillonné à l'occasion de 1 à 3 visites étalées dans le temps. À chaque échantillon (un individu, un site corporel, une visite) est attribué un numéro unique PSN (*Primary Sample Number*). Pour chaque échantillon individuel, une ou plusieurs extractions d'ADN ont été réalisées, auxquelles sont attribués des numéros *Sample Number*.

Deux types de jeux de données ont ensuite été générés à partir des échantillons décrits ci-dessus : des données d'amplicons 16S et des données de séquençage métagénomique complet (de type shotgun).

Données d'amplicons 16S Pour l'ensemble des échantillons du projet HMP, une analyse taxonomique par une approche amplicon a été réalisée. Le séquençage des amplicons a ciblé les régions variables V1-V3 et V3-V5 (1.1.5) et a été réalisé sur une plateforme de séquençage Roche-454 (1.2.3). Les données d'amplicons ont ensuite été analysées à l'aide de deux pipelines d'analyse différents, QIIME [11] et mothur [81] (2.3.1.1), afin de permettre l'identification d'OTU et de fournir une analyse taxonomique de chaque échantillon. Notamment, pour l'analyse QIIME, le projet HMP met à disposition toutes les données intermédiaires, dont les séquences représentatives d'OTU inférés pour chaque échantillon.

Données de séquençage métagénomique complet Pour plus de 1200 échantillons, un séquençage métagénomique complet a été réalisé. Chaque échantillon a été séquencé sur une plateforme de séquençage Illumina afin de fournir des lectures appariées de 101 bp (1.2.3). Les lectures ont été ensuite nettoyées sur la base de leur qualité, et celles qui correspondent au génome humain ont été supprimées (environ 49% des lectures). Le projet HMP met à disposition les jeux de données Illumina ainsi nettoyés pour tous les échantillons.

1.3.4.3 Mais tellement d'autres aussi (*data-flood*)

Ces dix dernières années ont vu proliférer les projets métagénomiques à large échelle (12 974 projets dans SRA au 07/08/2017 [12]). Comme nous l'avons évoqué, ces projets étudient des environnements aussi divers que les microbiomes humains ou animaux, les océans, les sols, les nuages, ou même les microbiomes urbains [95].

Les données générées par ces projets sont très variées, et leur quantité déjà importante ne va vraisemblablement faire qu'augmenter dans les années à venir. Il émerge donc un besoin important d'analyser les structures écologiques des échantillons métagénomiques. Cependant, le goulot d'étranglement, tant qualitatif que quantitatif, pour l'étude de ces environnements se situe maintenant au niveau de leur analyse bio-informatique.

Dans le chapitre II, nous présentons les méthodes bio-informatiques existantes pour l'analyse taxonomique des jeux de données métagénomiques complets, puis, dans le cadre plus précis de l'assemblage de marqueurs conservés, typiquement les ARNr SSU, nous ferons un état de l'art des méthodes existantes ainsi que de leurs limitations.

12. https://www.ncbi.nlm.nih.gov/sra/?term=metagenome

Chapitre 2

Les méthodes bio-informatiques

[de-novo] Metagenomic assembly is impossible

Mihai Pop, sous forme de running-gag

Dans le chapitre précédent, nous avons présenté la métagénomique, de la problématique générale de la biologie environnementale à la production des données. Dans ce chapitre, nous décrivons les approches bio-informatiques existantes pour l'analyse de ces données. Nous commençons par rappeler les méthodes universelles et atemporelles de l'analyse de séquences, à travers l'alignement, l'assemblage, puis présenterons plus en détail les approches développées ces dernières années pour l'analyse des données de métagénomique, avant leurs points forts et leurs limites.

2.1 Comparaison de séquences

Dans cette section, nous allons présenter les méthodes bio-informatiques élémentaires nécessaires à la manipulation de séquences. Cela concerne la représentation des séquences, les formats de données et les méthodes de base permettant de comparer les séquences.

2.1.1 Les formats de séquences : FASTA/FASTQ

En bio-informatique, les séquences nucléiques/protéiques sont généralement stockées sous la forme de fichiers texte, au format FASTA ou FASTQ.

Le format FASTA est relativement souple. Une entrée commence par un chevron supérieur, suivi d'un chapeau optionnel. Le chapeau est composé typiquement de l'identifiant de la séquence et d'informations complémentaires optionnelles. La séquence nucléotidique/protéique commence à la ligne suivante et peut couvrir plusieurs lignes continues.

```
>[identifiant] [informations complémentaires]
CATGCTGATCGTGANNNNTTGGC
ACACTACTAGCTATGCTAGCTAG
ACTGgacttcgatgcATCGAT
```

Le format FASTQ est le format utilisé pour la représentation des lectures de séquençage. En plus d'un identifiant et de la séquence, il permet de stocker pour chaque position une information sur la qualité de la lecture à cette position, exprimée par un score de qualité en *Phred*, qui correspond à une probabilité d'erreur de séquençage associée à chaque nucléotide. Pour des raisons de compression d'information, les qualités sont représentées par des caractères ASCII, avec deux échelles possibles (Phred+33 ou Phred+64).

```
@[identifiant] [informations complémentaires]
ACGTACTAGTCGATCGCTA
+[même identifiant, optionnel]
eed]‘]_Ba_^__[YBBBB
```

Plus d'information sur ces formats pourra être trouvée sur leurs pages Wikipédia respectives [1] [2].

2.1.2 Définitions du problème d'alignement

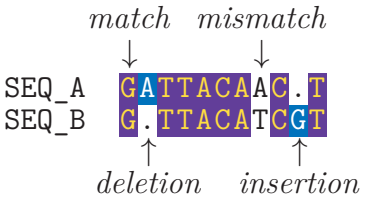


FIGURE 2.1 – Exemple d'alignement entre deux séquences

Un *alignement* est une représentation de la similarité entre deux séquences (Figure 2.1). Chaque position d'un alignement peut être : un *match* lorsque les nucléotides des deux séquences sont identiques, ou une *édition* lorsqu'il y a une différence. Les différentes opérations d'éditions sont le *mismatch*, lorsqu'à une position donnée la substitution d'un nucléotide par un autre permet de transformer une séquence en l'autre, ou encore l'*indel* (ou *gap*, noté « - »), lorsqu'à une position donnée la délétion ou l'insertion d'un nucléotide permet de transformer une séquence en l'autre. On définit le *pourcentage d'identité* entre deux séquences comme le rapport du nombre de *matchs* sur la longueur de l'alignement. Deux séquences identiques s'aligneront donc avec 100% d'identité.

1. FASTA : https://fr.wikipedia.org/wiki/FASTA_(format_de_fichier)
2. FASTQ : https://fr.wikipedia.org/wiki/FASTQ

Scores d'alignement La recherche d'un alignement passe par la maximisation d'un score ou réciproquement, la minimisation d'une distance. Une approche classique consiste à attribuer un score positif aux *matchs*, et à attribuer des scores négatifs aux *mismatchs*, à l'ouverture d'un *gap*, et à l'extension d'un *gap* déjà ouvert. Par exemple, le paramétrage par défaut de BLASTN (2.1.3) est le suivant : *match*=+2, *mismatch*=-3, ouverture d'un *gap*=-5, extension d'un *gap*=-2. On peut aussi utiliser une matrice de similarité entre nucléotides afin d'attribuer des scores différents aux différentes substitutions (par exemple, les substitutions A↔G et C↔T, nommées transitions, sont plus fréquentes que les autres substitutions possibles).

Types d'alignements On parle d'*alignement global* lorsque l'on cherche à aligner sur toute leur longueur deux séquences de tailles comparables et que l'on souhaite pénaliser des éditions en amont ou en aval des séquences.

Un *alignement local* entre deux séquences consiste à trouver l'alignement de meilleur score entre deux sous-séquences des séquences de départ. On utilise généralement l'alignement local lorsque l'on cherche à aligner une petite séquence sur une séquence bien plus grande, comme dans le cas de l'alignement d'une lecture de séquençage sur un génome.

Un alignement *semi-global* (ou *glocal*) cherche à trouver le meilleur alignement qui inclut le début et la fin de l'une ou l'autre des séquences. Il consiste à comparer le préfixe d'une séquence avec le suffixe de l'autre séquence, et inversement. Dans le cadre de l'assemblage avec graphe de chevauchement, c'est le type d'alignement qui est généralement utilisé pour la recherche de chevauchements de lectures avec erreurs (2.2.3).

2.1.3 Algorithmes d'alignement

Les algorithmes mis en œuvre pour construire des alignements de séquences dépendent principalement de la taille des données à aligner.

Gène contre gène Pour un cas d'application simple, visant à aligner deux séquences relativement courtes (quelques milliers de nucléotides), il est possible d'utiliser des méthodes exactes par programmation dynamique (Figure 2.2).

— pour l'alignement global : l'algorithme de Needleman-Wunsch [65], qui possède une complexité quadratique ;

— pour l'alignement local : l'algorithme de Smith-Waterman [91], dont la complexité est aussi quadratique.

En pratique, ces deux algorithmes sont très similaires, et l'algorithme de Swith-Waterman, plus généraliste, est plus souvent utilisé. Une extension de cet algorithme [26]

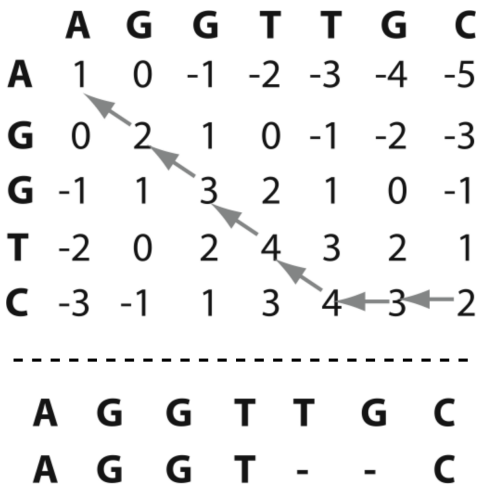


FIGURE 2.2 – Exemple d'une matrice de programmation dynamique calculée pour l'alignement global de 2 séquences. Pour cet exemple, les coûts sont les suivants : *match*=+1, *mismatch*=-1, *gap*=-1. La valeur de chaque cellule de la matrice correspond au score maximal atteignable parmi tous les alignements des préfixes des séquences se terminant à cette cellule. La cellule terminale de la matrice (en bas à droite) contient donc le score du meilleur alignement global possible. L'alignement est ensuite reconstruit par rebroussement en parcourant le chemin de plus fort score depuis la cellule terminale vers la cellule initiale.

permet de définir des coûts affines pour les gaps, afin de privilégier de grands gaps contigus plutôt qu'un ensemble de petits gaps isolés, ce qui correspond mieux au modèle biologique de création d'*indels*. Pour un historique plus détaillé des algorithmes exacts pour l'alignement par programmation dynamique, nous vous invitons à lire la note récente de Heng Li à ce sujet [44].

Gène contre génome Une autre application classique consiste à aligner un ensemble réduit (de l'ordre de quelques milliers au maximum) de petites séquences (gènes, ARN, etc.) contre un génome complet. Dans ce cas, il n'est plus possible de faire appel aux méthodes exactes, trop longues, notamment dans le cas de génomes eucaryotes pouvant mesurer plusieurs gigabases. Il est nécessaire d'utiliser des *heuristiques*. Parmi les stratégies les plus courantes, complémentaires, on retrouve :

— les techniques à base de *graines*, qui consistent à chercher de courtes sous-séquences présentes dans les deux séquences à aligner avec une forte similarité. L'alignement est ensuite étendu par programmation dynamique (Figure 2.3). Les graines peuvent être des mots de taille k (k-*mers*), des *graines espacées* [52], voire des graines avec erreurs ;

— l'indexation des génomes, qui permet d'accélérer la recherche de graines dans le

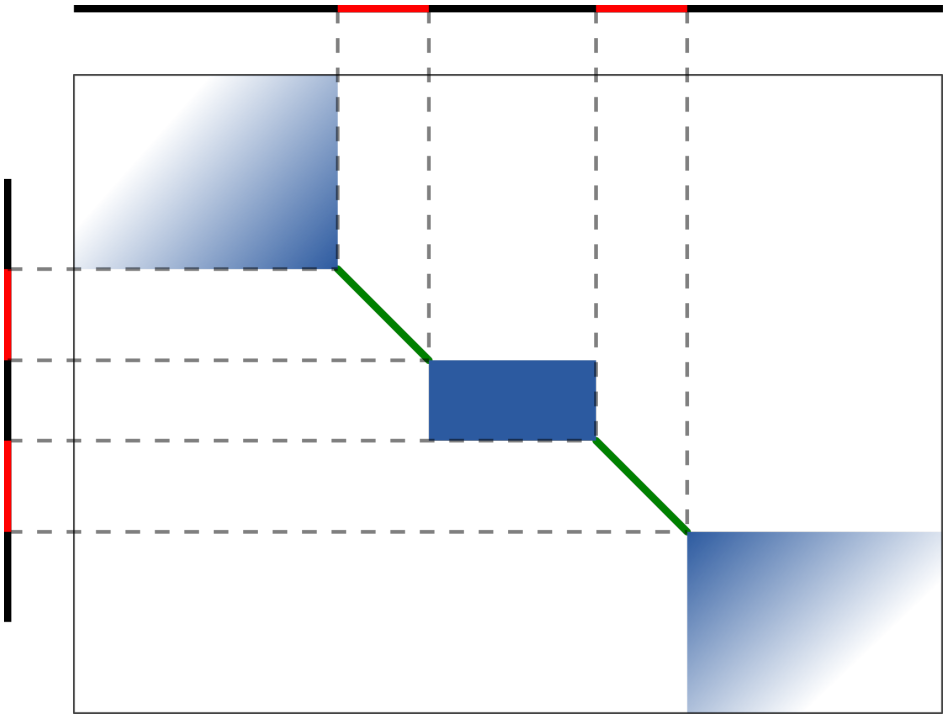


FIGURE 2.3 – Exemple d'application d'une heuristique à base de graines pour l'alignement de 2 séquences. L'extension (rectangles bleus) entre les graines (régions en rouge) est calculée par programmation dynamique, avec toutefois une forte réduction de la complexité par rapport à une méthode exacte. Un seuil minimal de score (généralement calculé à partir du score maximal) est utilisé pour arrêter l'extension aux extrémités.

texte. Certains index, tels que le FM-index associé à la transformée de Burrows-Wheeler, permettent en outre de compresser le texte.

Cette double approche, combinant graine et index, est exploitée par BLAST [1], le logiciel d'alignement de séquences le plus populaire encore actuellement (cité 65 939 fois sur Google Scholar). BLAST implémente des algorithmes offrant un bon compromis entre la sensibilité et la vitesse pour l'alignement d'un nombre raisonnable (quelques milliers) de courtes séquences ($< 10\,000$ bp) sur un génome.

BLAST est en fait une suite de modules spécialisés, tels que BLASTN pour les alignements nucléotidiques, ou BLASTP pour les alignements protéiques. BLASTN propose différentes configurations en fonction de la sensibilité souhaitée, notamment en jouant sur le calcul du score, sur la taille des graines (7 nucléotides pour le mode le plus sensible, et jusqu'à 28 nucléotides pour la recherche de séquences très similaires). De plus, pour chaque alignement BLAST calcule une *e-valeur* (*e-value*) indiquant la significativité de l'alignement.

Lectures de séquençage contre génome Avec le développement du séquençage à haut débit, il est devenu nécessaire de pouvoir aligner de très grandes quantités de petites lectures de séquençage (jusqu'à plusieurs milliards de lectures de quelques centaines de nucléotides) sur un génome de référence. Dans ce contexte de *re séquençage*, on s'attend à peu de différences entre les lectures et le génome (soient les erreurs de séquençage et les polymorphismes), ce qui permet l'utilisation d'heuristiques plus performantes et mieux adaptées que BLAST.

De nombreuses méthodes ont ainsi été développées [46], parmi lesquelles Bowtie [41] et BWA [45], qui font appel à un index du génome basé sur une transformée de Burrows-Wheeler et l'utilisation d'un FM-index. Ces structures combinées à des algorithmes de *backtracking* novateurs et une implémentation parallèle permettent l'alignement rapide d'un très grand nombre de lectures sur ce génome, avec une faible empreinte mémoire.

Plus récemment, des approches telles que SortMeRNA [38, 39] (2.3.4.2) font appel à des structures de données et des algorithmes permettant d'augmenter la sensibilité et d'aligner des lectures plus distantes du génome de référence (jusqu'à 85% d'identité entre les lectures et le génome).

2.1.4 Les formats d'alignements : SAM/BLAST

De nombreux formats de fichier existent pour représenter des alignements de séquences. Nous présentons ici deux formats très largement utilisés et complémentaires.

Le format SAM est une représentation au format texte qui permet de stocker un ensemble d'alignements de séquences contre une base de données. Le cas d'école consiste à stocker les alignements d'une grande quantité de lectures de séquençage contre un génome de référence ou une base de génomes de référence. Un fichier SAM permet de stocker de très nombreux attributs pour chaque alignement. Les plus importants sont : l'identifiant de la séquence sujet (*subject*), l'identifiant de la séquence requête (*query*), la position du début de l'alignement sur la séquence requête, la structure de l'alignement (CIGAR) et la séquence sujet alignée (Figure 2.4). Afin de réduire la taille d'un fichier SAM, il est possible de le compresser sous la forme d'un fichier binaire au format BAM. Une description complète des spécifications du format SAM peut être trouvée sur le site officiel des SAMTOOLS. [3]

Le format BLAST tabulé est une représentation plus focalisée sur l'information de similarité entre la séquence sujet et la séquence requête. Les attributs les plus importants sont : les identifiants des séquences sujet et requête, la taille de l'alignement, le

3. SAM : https://samtools.github.io/hts-specs/SAMv1.pdf

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA     *
r003    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC        *
```

FIGURE 2.4 – Exemple d'alignements au format SAM

pourcentage d'identité entre les deux séquences sur l'alignement, les positions de l'alignement sur les deux séquences, le score de l'alignement, et une e-valeur. Pour comparer les alignements de deux lectures sur une même base de données, on privilégiera la e-valeur. Pour comparer les alignements de la même lecture sur deux séquences requêtes différentes, on privilégiera le score. Une description complète du format BLAST tabulé peut être trouvée dans le manuel de Blast+ sur le site du NCBI. [4]

2.2 La reconstruction de séquences

Comme précédemment évoqué, le séquençage à haut débit fournit un ensemble de courtes lectures de séquençage. Il est souvent nécessaire de pouvoir reconstruire la séquence initiale à partir de ces fragments en l'absence d'un génome de référence. C'est ce qu'on appelle le problème de l'*assemblage*, qui est central en bio-informatique, au même titre que l'alignement de séquences. Nous présentons tout d'abord l'assemblage dans le contexte classique de la génomique, avec la reconstruction d'un génome. Puis nous abordons l'assemblage en métagénomique.

Vous pouvez retrouver une introduction à l'assemblage sur la page Wikipédia Fr [a], dont je suis l'auteur principal sous le pseudonyme NG0S.

a. https://fr.wikipedia.org/wiki/Assemblage_(bio-informatique)

2.2.1 Définitions

Un *assemblage* consiste à fusionner des petites séquences issues d'une plus longue séquence afin de reconstruire cette séquence originale. Le problème de l'assemblage peut être comparé de manière imagée à celui de la reconstruction du texte d'un livre à partir de plusieurs copies de celui-ci, préalablement déchiquetées en petits morceaux. Son application principale en bio-informatique est la reconstruction de génomes, ou de gènes, à partir d'un ensemble de lectures de séquençage.

4. BLAST : https://www.ncbi.nlm.nih.gov/books/NBK279675/

Les principales difficultés de l'assemblage, qui rendent ce problème non trivial, sont :

— les erreurs de séquençage (1.2.1). Le taux et le type d'erreur varient en fonction des technologies de séquençage utilisées (1.2.3.3). De plus, le type d'erreur, substitution ou *indels*, doit être pris en compte par les algorithmes d'assemblage ;

— les longues régions répétées ou très similaires au sein d'un même génome ;

— l'hétérozygotie des génomes dans le cadre d'espèces polyploïdes, pouvant présenter plusieurs versions quasiment identiques de la même séquence ;

— les régions de faibles complexités comme les successions d'homopolymères de mêmes nucléotides, dinucléotides, tri nucléotides, etc. (ex. `TATATATATATATATA`).

Les assembleurs génomiques peuvent être classés suivant trois paradigmes distincts [63] :

— les assembleurs gloutons ;

— les assembleurs OLC (*Overlap-Layout-Consensus*, soit Chevauchement-Organisation-Consensus) ;

— les assembleurs basés sur un graphe de De Bruijn.

Historiquement, les paradigmes gloutons et OLC ont été développés pour l'assemblage de lectures issues du séquençage « bas débit » Sanger. C'est l'apparition des technologies de séquençage à haut débit qui a entraîné des contraintes supplémentaires en ce qui concerne la quantité de données à traiter et la mémoire utilisée, et a nécessité le développement des méthodes basées sur les graphes de De Bruijn.

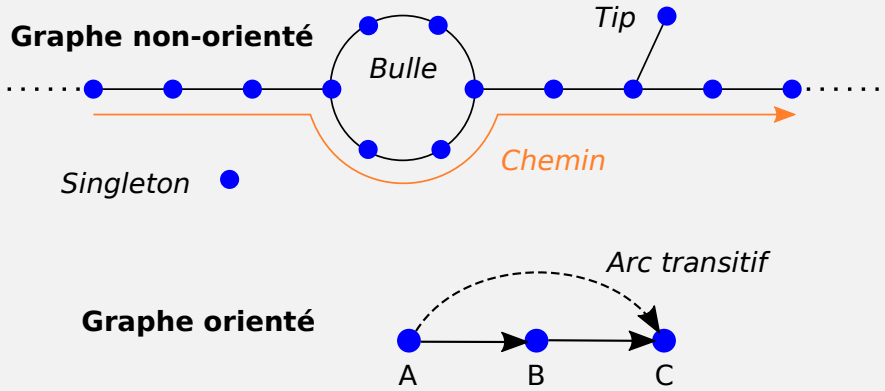
2.2.2 Le paradigme glouton

Il correspond historiquement à la première stratégie d'assemblage. Le principe général, schématiquement, est de chercher une super-séquence commune aux lectures de séquençage. Pour cela, tous les chevauchements deux à deux entre les lectures sont calculés, et la séquence initiale est reconstruite en regroupant les lectures dans l'ordre décroissant de leur score de chevauchement. Cette stratégie ne prend pas en compte la relation globale entre les séquences et peut donc mener à des optimums locaux. Elle n'est pas non plus adaptée à l'assemblage de génomes, qui peuvent contenir des régions répétées. La plupart des premiers assembleurs tels que phrap [5] ou CAP3 [33], ainsi que des outils plus récents tels que VCAKE [35], reposent sur ce principe.

En pratique, les assembleurs modernes se répartissent désormais entre les deux paradigmes d'assemblage suivants : les assembleurs OLC ou De Bruijn.

5. http://www.phrap.org/phredphrap/phrap.html

Rappels des termes généraux pour les graphes d'assemblage



graphe non orienté ensemble de nœuds et d'arêtes. Une arête est un couple de nœuds non ordonné

graphe orienté ensemble de nœuds et d'arcs. Un arc est un couple de nœuds dont le premier élément est le nœud de départ et le second élément est le nœud d'arrivée de l'arc

chemin/chaîne un chemin est une séquence de nœuds qui passe par des arêtes du graphe non orienté. Respectivement, une chaîne est une séquence de nœuds qui passe par des arcs du graphe orienté

degré/arité dans un graphe non orienté, le degré d'un nœud correspond au nombre de ses arêtes adjacentes. Dans un graphe orienté, l'arité d'un nœud correspond au nombre d'arcs sortant du nœud

singleton un singleton est un nœud d'arité ou de degré nul, connecté à aucun autre nœud

composante connexe un ensemble de nœuds forme une composante connexe dans un graphe non orienté s'il existe une suite d'arêtes permettant de relier n'importe quelle paire de nœuds de cet ensemble.

bulle une bulle apparaît lorsqu'il existe deux chemins alternatifs courts qui relient deux nœuds donnés

tip un *tip* est défini comme une courte « impasse » dans le graphe

arc transitif un arc $A \rightarrow C$ est transitif s'il existe un autre chemin entre A et C, par exemple s'il existe un arc $A \rightarrow B$ et un arc $B \rightarrow C$

2.2.3 Le paradigme OLC

Comme les approches gloutonnes, le paradigme OLC (Overlap-Layout-Consensus) cherche à reconstruire une super-séquence commune à partir de tous les chevauchements deux à deux des lectures (*Overlap*). La différence est que ces chevauchements sont trai-

tés de manière globale, par la construction d'un graphe de chevauchements. À partir de ce graphe sont identifiées des composantes, appelées *contigs* (*Layout*), et la séquence la plus probable pour chacun des contigs est enfin retenue (*Consensus*).

Définitions Un graphe de chevauchement est un graphe orienté dans lequel chaque *nœud* est une lecture de séquençage et chaque *arc* correspond à un chevauchement entre le suffixe du nœud de départ et le préfixe du nœud d'arrivée. Les chevauchements peuvent être estimés sans erreur, ou bien, lorsque des erreurs sont acceptées, en fixant un seuil de distance entre suffixes et préfixes (typiquement un pourcentage d'identité minimum). L'approche la plus naïve pour construire un graphe de chevauchement à partir d'un ensemble de lectures consiste à calculer un alignement semi-global (2.1.2) pour chaque paire de lectures possible. Si le chevauchement entre les deux séquences est suffisamment long et similaire, une arête est ajoutée entre les deux lectures. Ce type d'approche naïve a une complexité quadratique par rapport au nombre de lectures, et devient rapidement peu pratique lorsque le nombre de lectures augmente, comme c'est le cas avec l'apparition des technologies de séquençage à haut débit (1.2.3), et plus encore avec le séquençage d'échantillons métagénomiques (1.3).

Les heuristiques de construction Un ensemble d'heuristiques permet d'accélérer grandement la comparaison des paires de lectures. Pour la recherche de chevauchements sans erreur, l'utilisation de structures d'index permet de calculer très rapidement l'ensemble des chevauchements entre les lectures. L'empreinte mémoire dépend de la structure choisie, telle que la table des suffixes/préfixes [53], la table des suffixes compressés [27] ou le FM-index [88]. Si on souhaite autoriser des erreurs au sein des chevauchements, il est nécessaire de recourir à de véritables alignements. De manière analogue aux heuristiques de *graines* pour les alignements (2.1.3), il est possible d'indexer préalablement tous les k-mers présents dans l'ensemble des lectures. Cet index est utilisé pour ne calculer que les chevauchements des paires de séquences présentant des k-mers communs, à partir desquels un alignement est étendu par programmation dynamique.

L'exploitation du graphe Identifier la plus courte super-séquence commune dans le graphe de chevauchement revient, de manière schématique, à trouver un chemin minimal qui visite tous les nœuds du graphe exactement une fois. Il s'agit du problème bien connu de la recherche d'un chemin hamiltonien, qui est NP-complet. Ce qui veut dire que trouver un chemin optimal dans un graphe de chevauchement est NP-difficile. L'exploitation d'un graphe de chevauchement nécessite donc l'utilisation d'heuristiques pour pouvoir être réalisée dans des temps raisonnables.

Les méthodes d'assemblage selon le paradigme glouton (2.2.2) peuvent être ainsi vues comme la recherche gloutonne d'un chemin dans le graphe de chevauchement sans jamais le construire explicitement.

Dans une approche OLC, l'exploitation du graphe de chevauchement consiste à identifier des composantes (sous-graphes appelés *contigs*) regroupant des lectures issues de la même séquence. Ensuite, par une succession d'étapes de nettoyage, de simplification du graphe et de calcul de probabilités, la séquence la plus probable est déterminée pour chaque contig. L'identification des contigs comprend généralement :

— une étape de suppression des arcs transitifs

— un élagage des *tips* (généralement générés par une erreur de séquençage)

— un éclatement des bulles, au cours duquel les chemins alternatifs de faibles poids sont généralement supprimés

Les contigs correspondent alors aux chemins non branchants du graphe. Pour finir, les lectures de chaque contig sont réalignées, et une séquence consensus est déterminée en choisissant pour chaque position le nucléotide le plus fréquent parmi les lectures couvrant cette position.

Les méthodes Le paradigme OLC a été rendu populaire par les travaux de Gene Myers, et plus particulièrement l'assembleur Celera [62] qui a dominé le domaine de l'assemblage jusqu'à l'émergence des technologies de séquençage haut débit (1.2.3.3). Les meilleures méthodes d'assemblage de lectures de séquençage de type Sanger faisaient ainsi appel au paradigme OCL. Ces méthodes se sont toutefois montrées incapables d'assembler les trop grandes quantités de données générées par les technologies de séquençage haut débit, notamment à cause de la quantité de mémoire nécessaire pour stocker les graphes de chevauchement. Ces limitations ont ainsi nécessité l'apparition d'un nouveau paradigme d'assemblage, faisant appel à des graphes de De Bruijn (2.2.4). Plus récemment, l'utilisation de structures d'indexation performantes a permis au paradigme OLC de réémerger. Et c'est notamment grâce à l'utilisation d'un FM-index que l'assembleur SGA [89] est aujourd'hui capable d'assembler de grands génomes eucaryotes avec une empreinte mémoire raisonnable ($\sim$ 50 GB RAM).

2.2.4 Graphe de De Bruijn

Cette approche, dont l'application à l'assemblage a été proposée par Pevzner et al. en 2001 [72], est basée sur la construction d'un graphe de De Bruijn, et a été développée avec l'objectif de réduire les ressources nécessaires à l'assemblage de très gros jeux de données de séquençage (1.2.3).

Définitions et construction Un *graphe de De Bruijn* est le graphe de chevauchement de tous les k-mers de taille fixée pour l'ensemble des lectures. Un graphe de De Bruijn est un graphe orienté dans lequel les nœuds sont des mots de taille k (k-mers), et les arêtes sont les chevauchements de taille $k-1$ entre les k-mers. Par exemple les 5-mers ACTAG et CTAGT partagent exactement 4 lettres et sont reliés par une

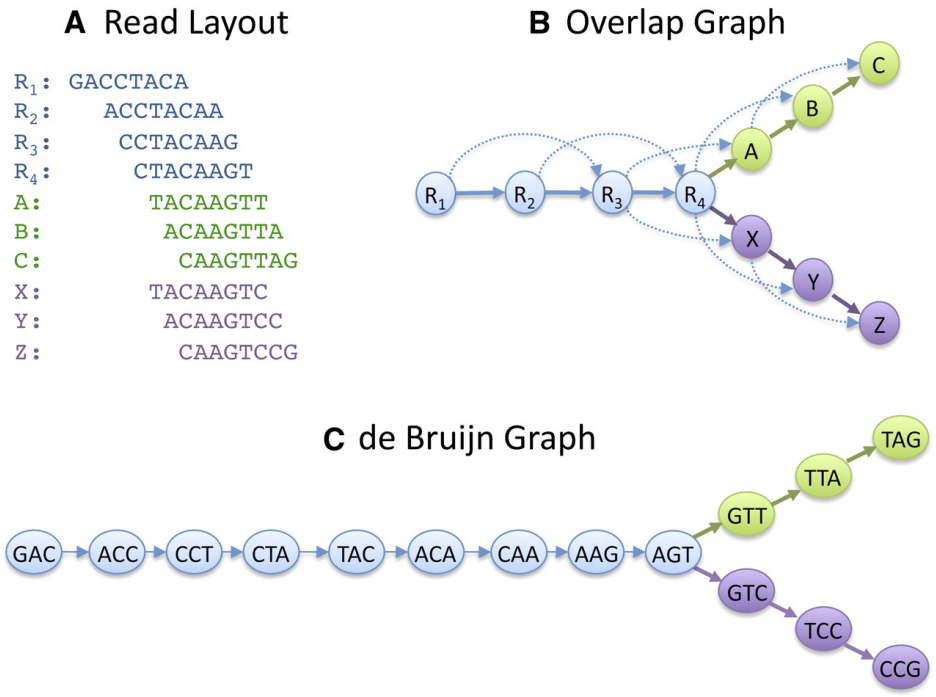


FIGURE 2.5 – Différences entre un graphe de chevauchement et un graphe de De Bruijn pour l'assemblage [80]. À partir d'un ensemble de 10 lectures de 8bp (A), un graphe de chevauchement peut être construit (B) dans lequel chaque lecture est un nœud, et les chevauchements supérieurs à 5bp sont représentés par des arcs. Les arcs transitifs sont représentés par des flèches en pointillé. Dans le graphe de De Bruijn (C), un nœud est créé pour chaque k-mer dans les lectures ; avec ici une taille de k-mer de 3 nucléotides. Des arcs sont tirés entre chaque paire de k-mers successifs dans les lectures, si ces k-mers se chevauchent sur $k-1$ nucléotides. La présence d'une différence entre les lectures vertes et violettes (où un A est remplacé par un C) entraîne la création d'un embranchement dans les deux graphes.

arête dans le graphe de De Bruijn. On définit aussi un *unitig* comme un chemin sans embranchement dans le graphe de De Bruijn.

Un graphe de De Bruijn est une représentation compressée (avec perte) de l'information de séquence portée par les lectures. En pratique, il est possible de stocker un graphe de De Bruijn dans très peu d'espace, par exemple en stockant seulement les k-mers dans une table de hachage, les chevauchements pouvant être retrouvés facilement. On peut aussi stocker le nombre d'apparitions de chaque k-mer dans le jeu de données, ce qui permet de pondérer les arêtes du graphe de De Bruijn. Contrairement à la construction d'un graphe de chevauchement de lectures, la construction du graphe de De Bruijn peut être réalisée en temps linéaire.

Exploitation du graphe En théorie, et toujours dans le cas simplifié sans erreurs de séquençage, sans répétitions génomiques, et avec une couverture de séquençage homogène, trouver la super séquence commune la plus courte revient à trouver un chemin qui passe par toutes les arêtes du graphe une seule fois, soit un chemin Eulérien [72]. Or, il existe des algorithmes qui permettent de trouver un chemin eulérien dans un graphe de manière efficace. Cette solution est donc plus efficace que l'approche utilisant un graphe de chevauchement de lectures. Toutefois, la présence d'erreurs de séquençage et de répétitions complexifie la structure du graphe de De Bruijn et rend son exploitation non triviale. Parmi les heuristiques standards pour nettoyer et simplifier un graphe de De Bruijn on trouve :

— la suppression des *tips* ;

— l'éclatement des bulles, en supprimant les arêtes de faible poids ;

— la suppression de chemins chimériques, à nouveau en supprimant les arêtes de faible poids.

Le résultat de l'assemblage, les *contigs*, sont alors les *unitigs* du graphe de De Bruijn nettoyé.

Critiques de l'approche Cette approche a été développée lors de l'apparition des premières technologies de séquençage nouvelle génération. Parmi les logiciels les plus populaires on peut notamment citer Velvet [105] et SOAPdenovo [47]. À cette époque, les lectures générées étaient de l'ordre de quelques dizaines de nucléotides, et dépassaient rarement la centaine de nucléotides. Il était donc sensé de découper les lectures en k-mers et de compresser ainsi l'information, sans trop de perte.

Toutefois, le principal défaut de cette approche est la perte de l'information de cohérence de lecture. Au moment de l'assemblage, des chemins peuvent être choisis dans le graphe de De Bruijn qui ne sont supportés par aucune des lectures. Cette perte d'information est d'autant plus dommageable que la longueur moyenne des lectures a augmenté avec le temps, accompagnée d'une diminution importante des taux d'erreur. Le développement récent des approches d'assemblage multi k-mers [69], qui consiste à construire itérativement des graphes de De Bruijn avec des tailles de k-mers croissantes, permet de réintégrer l'information issue des lectures complètes à chaque itération. Les assembleurs qui implémentent cette approche, comme SPAdes [4] ou IDBA [69], sont aujourd'hui considérés comme les plus performants pour l'assemblage génomique.

2.2.5 Le *scaffolding* / échafaudage

La plupart des assembleurs modernes réalisent une étape supplémentaire post-assemblage, appelée *scaffolding*. Cette étape consiste à utiliser des informations autres que celles portées par les séquences des lectures pour organiser et ordonner au mieux

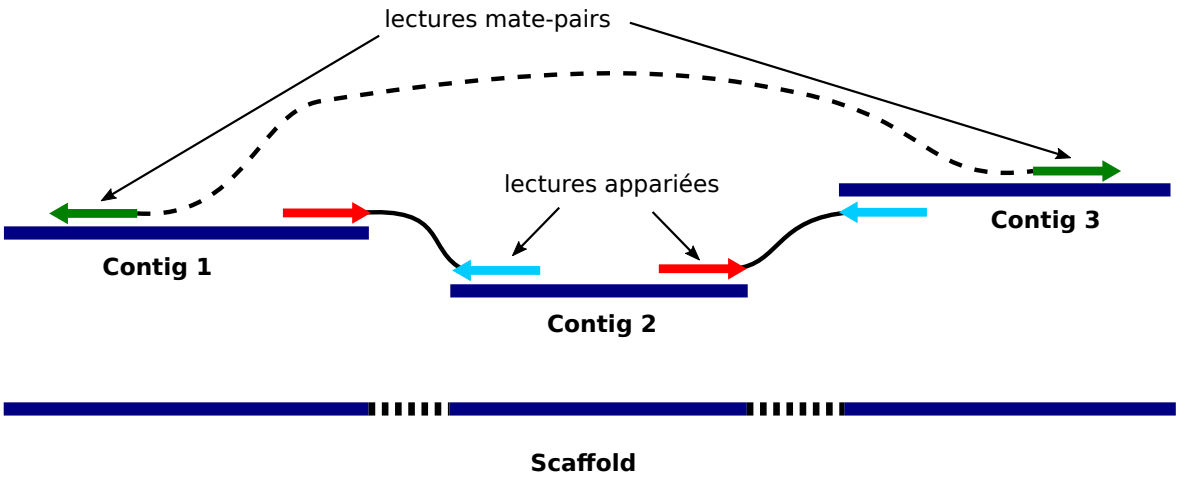


FIGURE 2.6 – Principe du scaffolding. Les lectures appariées et *mate-pairs* sont utilisées pour orienter et organiser les contigs entre eux. La taille des fragments sur lesquels sont séquencées les lectures permet d'estimer la distance entre les différents contigs. Finalement, un *scaffold* est généré en concaténant les contigs ordonnés, et en rajoutant des nucléotides inconnus (notés « N ») entre les contigs.

les contigs les uns par rapport aux autres. Le scaffolding exploite principalement l'information d'appariement des lectures de séquençage pour les librairies appariées ou mate-pairs. Les lectures sont donc réalignées sur les contigs, et lorsque suffisamment de lectures appariées sont alignées sur un contig et un autre, la taille moyenne des fragments est utilisée pour estimer la distance entre ces contigs. Les séquences consensus ainsi créées, appelées *scaffolds*, contiennent alors des séquences de nucléotides inconnus (notés « N ») pour indiquer l'écartement estimé entre les séquences des contigs.

2.2.6 Un mot sur le nettoyage des jeux de données de séquençage

Les pipelines d'analyse de données de séquençage intègrent généralement une étape de nettoyage des données, préalablement aux alignements ou assemblages. En fonction de la qualité du séquençage et de la robustesse des méthodes d'analyse aux erreurs et contaminations, ce nettoyage peut comprendre plusieurs phases :

- un *nettoyage sur la base de la qualité*. En fonction de la technologie de séquençage utilisée, un seuil de qualité minimale peut être utilisé pour couper un morceau de mauvaise qualité d'une lecture, ou bien pour supprimer une lecture entière de qualité moyenne trop faible ;
- la *suppression des nucléotides indéterminés*. Certains outils ne prenant pas en charge les nucléotides inconnus, il est possible de couper les lectures pour éliminer une portion en contenant, ou bien de supprimer une lecture qui en contient trop ;
- la *suppression des adaptateurs*. La chimie de certaines technologies de séquençage impose la ligation d'adaptateurs aux fragments d'ADN séquencés, et ces adaptateurs se retrouvent parfois aux extrémités des lectures. Il est important d'identifier et supprimer ces séquences artificielles, qui peuvent être confondues avec des régions répétées si elles restent présentes dans les données ;
- la *suppression des lectures trop courtes* après nettoyage. Les étapes précédentes de nettoyage peuvent amputer certaines lectures d'une grande part de leur séquence. Ces lectures trop courtes peuvent alors être supprimées du jeu de données.
- la *suppression des contaminants*. Malgré les précautions généralement employées, il est parfois possible (voire inévitable dans certains cas) de prélever des échantillons biologiques contaminés par des organismes que l'on ne souhaite pas séquencer. Les lectures correspondant à ces contaminants peuvent parfois être identifiées et éliminées *a posteriori* en utilisant par exemple un seuil basé sur le taux de GC (qui discrimine généralement efficacement entre procaryotes et eucaryotes).

La plupart des méthodes d'analyse de données de séquençage attendent des lectures de bonne qualité en entrée, et avec le moins de contamination possible. C'est pourquoi une bonne pratique consiste à réaliser le nettoyage des données de séquençage préliminairement, avec des outils tels que Prinseq [82] et Cutadapt [54], afin de fournir aux outils des lectures de la meilleure qualité possible.

2.3 Analyse de données de métagénomique

Nous arrivons maintenant à notre sujet d'intérêt principal : l'analyse des données de métagénomiques. Dans ce contexte, les méthodes bio-informatiques actuelles d'analyse taxonomique poursuivent deux objectifs majeurs complémentaires : identifier les espèces de micro-organismes présentes dans un échantillon en réalisant une *analyse taxonomique* directe des lectures, ou bien reconstruire les génomes (ou les régions d'intérêt) pour toutes les espèces présentes dans l'échantillon en réalisant un *assemblage* des lectures.

Cela se décline de plusieurs manières. Historiquement, la disponibilité de données de séquençage ciblé, et en particulier d'amplicons pour l'ARNr SSU (1.3.1), a permis le développement de méthodes pour l'analyse taxonomique à gros grains des échantillons métagénomiques. L'apparition plus récente de données de séquençage métagénomique complet (1.3.2) a permis de commencer à développer de nouvelles approches, que ce soit pour l'assemblage des jeux de données complets, ou bien l'analyse taxonomique sans assemblage. D'autres approches se focalisent sur l'analyse de marqueurs conservés dans ces jeux de données de séquençage métagénomique complet, d'abord par l'identification des lectures correspondant aux marqueurs, puis par la reconstruction « pleine taille » de ces marqueurs, qui peut ensuite être utilisée pour réaliser l'analyse taxonomique d'un échantillon de manière plus fine et sans *a priori* (1.2.3.1). Nous présentons, dans cet ordre, l'analyse de données de séquençage ciblé de type amplicon, l'assemblage de données de séquençage métagénomique complet, l'analyse taxonomique directe sans assemblage, et enfin la reconstruction de marqueurs conservés pour l'analyse taxonomique.

2.3.1 Analyse de séquençage ciblé de type amplicon

Comme nous l'avons vu au chapitre I, section 1.3.1, le séquençage de type amplicon d'un échantillon métagénomique consiste à réaliser le séquençage ciblé d'une ou plusieurs régions hypervariables de l'ARNr SSU (1.1.5.1). La définition *a priori* d'amorces « universelles » permet, en théorie, l'amplification de ces régions pour l'ensemble des espèces présentes dans l'échantillon. Comme nous l'avons vu, cette hypothèse peut être mise en défaut (1.3.1). Des séquences chimériques peuvent être créées, et les séquences d'espèces divergées inconnues peuvent ne pas être amplifiées.

2.3.1.1 Les données d'amplicons

Les régions hypervariables de l'ARNr SSU varient généralement en taille entre 30 et 100 bp (1.1.5.1). Puisque la combinaison de plusieurs régions hypervariables permet d'augmenter le *signal phylogénétique* (soit le nombre de sites disponibles pour prédire la taxonomie de la séquence) utile pour l'assignation taxonomique, les approches

amplicons ciblent généralement des zones de 200 à 400 bp couvrant plusieurs régions hypervariables contiguës. Les données issues d'un séquençage ciblé de type amplicon consistent alors en un ensemble de lectures de tailles *quasi* identiques, qui correspondent à la taille de la région amplifiée. Chaque lecture est aussi accompagnée de qualités estimées par le séquenceur, qui associe une probabilité d'erreur à chaque nucléotide (1.2.1).

Types d'erreurs En fonction de la technologie de séquençage utilisée, on connaît aussi le type d'erreurs attendues. Historiquement, la technologie Roche-454 a été majoritairement employée pour le séquençage d'amplicons, et les lectures s'accompagnent donc principalement d'erreurs de type *indels* dans les homopolymères. L'amélioration des longueurs de séquençage pour la technologie Illumina permet maintenant d'obtenir des lectures d'amplicons, principalement séquencées sur la plateforme MiSeq, avec un faible taux d'erreur, les erreurs étant limitées à des substitutions.

2.3.1.2 Les pipelines d'analyse

Plusieurs pipelines d'analyse taxonomique d'amplicon existent, parmi lesquels QIIME [11] et mothur [81] sont les plus populaires. La plupart de ces approches appliquent toutefois un principe général commun consistant à nettoyer et dé-répliquer les séquences, construire des OTUs par partitionnement, rejeter les séquences chimériques, et finalement assigner taxonomiquement les séquences représentatives de chaque OTU.

Nettoyage et dé-réplication des lectures La première étape d'une analyse amplicon consiste en une dé-réplication des lectures, parfois précédée d'une étape de nettoyage basée sur la qualité (2.2.5). La dé-réplication des lectures consiste à regrouper ensemble les lectures identiques, et à associer à chaque lecture une *abondance*, qui correspond au nombre de copies identiques de cette lecture.

On s'attend à ce qu'une espèce abondante dans l'échantillon soit représentée par de nombreuses copies identiques de la région amplifiée. À l'opposé, les espèces rares, et les lectures avec erreurs de séquençage ne possèdent généralement que peu de copies identiques dans le jeu de données.

Définition des unités taxonomiques opérationnelles (OTU) par partitionnement Après nettoyage et dé-réplication, les lectures sont regroupées dans des classes (appelées *OTU*) sur la base de leur similarité. L'objectif est de constituer des groupes de séquences similaires provenant probablement d'organismes proches.

Ces OTU sont construits par partitionnement des lectures, avec un seuil de partitionnement qui correspond généralement au seuil utilisé pour la définition des espèces procaryotes, c'est-à-dire une identité supérieure à 97% de l'ARNr SSU (1.1.6.1).

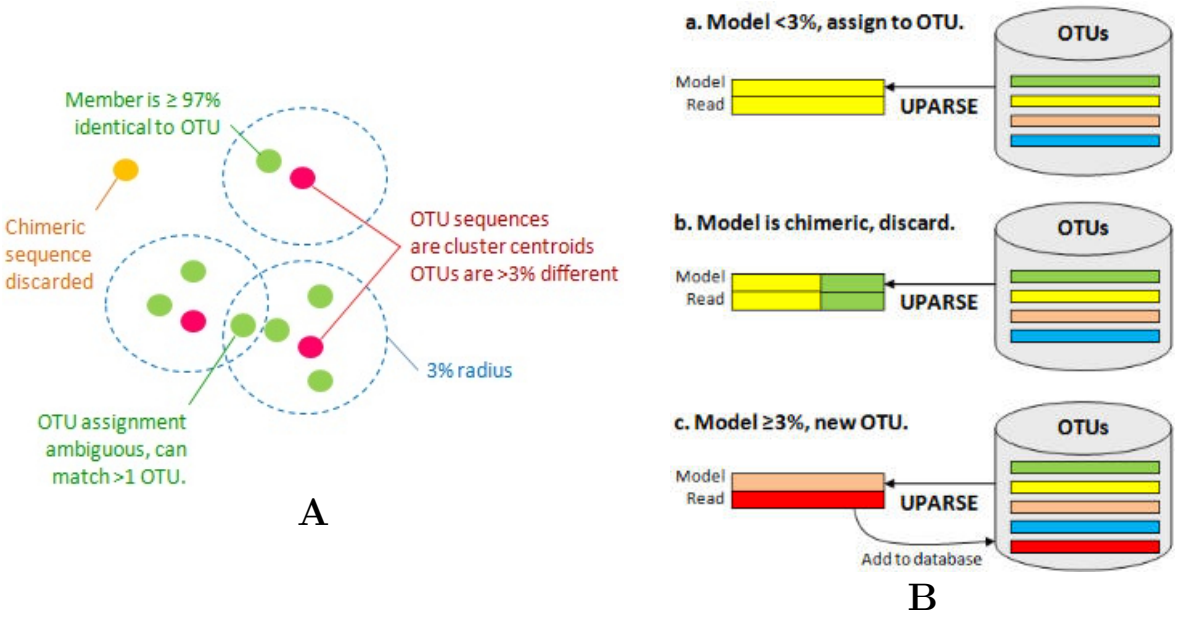


FIGURE 2.7 – Exemple d'un algorithme de partitionnement, tel qu'implémenté dans UPARSE [23]. A) Définition des OTU. Les séquences représentatives sont indiquées en rose, et le rayon de 97% d'identité est indiqué pour chaque OTU. B) Cas d'attribution dans un OTU. Dans cet algorithme, les lectures chimériques sont détectées *de novo* suite à l'alignement des lectures sur les séquences représentatives des OTU.

En pratique, des outils de partitionnement comme UCLUST [22] (maintenant supplanté par UPARSE [23]) ou CD-HIT [48] font appel à des heuristiques (Figure 2.7). La plus courante consiste à trier d'abord les lectures par abondance décroissante, considérant ainsi les lectures les plus fiables en premier. Chaque lecture est ensuite attribuée à la première partition dont la séquence représentative est similaire avec plus de 97% d'identité. Cette identité est estimée selon un alignement global calculé entre les deux séquences (2.1.2). Si aucune partition suffisamment similaire n'existe, la lecture devient la séquence représentative d'une nouvelle partition.

Identification des séquences chimériques La détection des séquences chimériques (1.3.1) peut être réalisée *de novo* ou bien à l'aide d'une séquence de référence :

— la détection *de novo* est généralement intégrée à l'étape de partitionnement. Sont alors identifiées comme chimériques les séquences dont une partie s'aligne bien sur la séquence représentative d'un OTU, et l'autre partie sur celle d'un autre OTU.

— si l'on dispose d'une base de séquences de référence de qualité, il est possible d'identifier les séquences chimériques *a posteriori* en alignant chaque lecture sur la base de référence avec un outil dédié comme UCHIME [24]. Sont alors considérées

comme chimériques les lectures dont une partie s'aligne préférentiellement sur une espèce, et l'autre partie sur une autre espèce.

Assignation taxonomique des séquences représentatives d'OTU Finalement, les séquences représentatives de chaque OTU sont assignées taxonomiquement par alignement des séquences sur des bases de références de haute qualité telles que SILVA [75] ou Greengenes [19], ou grâce à un classifier bayésien tel que RDP [98].

Les assignations taxonomiques, combinées à l'abondance de chaque OTU, donnent alors une vision de la composition taxonomique des organismes présents dans l'échantillon biologique de départ.

2.3.1.3 Les limitations de l'approche

La principale faiblesse de l'approche amplicon pour l'analyse taxonomique d'un échantillon concerne sa faible résolution. En effet, à cause des erreurs de séquençage, de la faible longueur des régions ciblées (∼400bp) et du seuil d'assignation des OTU à 97% d'identité, cette approche ne permet pas d'assigner taxonomiquement les séquences au niveau de l'espèce, et la plupart des analyses décrivent la composition taxonomique de l'échantillon au niveau du genre, voire de la famille.

De plus, les résultats de ces analyses taxonomiques dépendent du pipeline utilisé, et la variabilité des assignations estimées augmente d'autant plus que le taux d'erreur est élevé [87].

Ce manque de qualité du signal phylogénétique exploitable est une limitation intrinsèque des données de type amplicon. C'est pourquoi de nouvelles approches proposent d'utiliser l'information contenue dans les données de séquençage métagénomique complet pour assigner plus finement une taxonomie aux séquences, jusqu'au niveau de l'espèce, ou même de la souche.

2.3.2 Assemblage de données de séquençage métagénomique complet

De manière analogue aux projets d'assemblage génomique, on souhaite pouvoir reconstruire les génomes de chaque espèce présente dans un échantillon biologique.

2.3.2.1 Un nouveau problème

Dans le cadre de l'assemblage génomique, nous avons vu que le problème consiste à trouver un chemin dans un graphe de chevauchement ou un graphe de De Bruijn, afin de reconstruire le génome d'une seule espèce, séquencée de manière relativement homogène (2.2). L'assemblage de données de séquençage métagénomique complet doit répondre à des contraintes supplémentaires, qui rendent cette tâche particulièrement difficile :

— le nombre d'espèces est inconnu, et peut varier de quelques espèces à plusieurs milliers ;

— les distances évolutives entre les espèces sont inconnues. Assembler deux espèces très proches sans créer de séquences chimériques ou consensus sera forcément plus difficile qu'assembler deux espèces distantes ;

— l'abondance de chaque espèce est inconnue. Les lectures d'une espèce rare pourront alors être confondues avec des lectures avec erreurs de séquençage.

2.3.2.2 Les méthodes d'assemblage métagénomique

À l'heure actuelle, il n'existe pas de méthode d'assemblage métagénomique capable de reconstruire les génomes des espèces d'un échantillon métagénomique avec la qualité d'assemblage des projets génomiques. Toutefois, ces dernières années ont vu l'émergence de méthodes performantes qui prennent en compte les nouvelles contraintes de l'assemblage métagénomique.

En 2015, l'initiative CAMI [84] (*Critical Assessment of Metagenomic Interpretation*) a organisé le premier concours d'assemblage métagénomique, avec pour objectif d'évaluer les méthodes existantes et de déterminer les axes futurs de recherche méthodologique. Les résultats de ce concours ont permis d'identifier les méthodes d'assemblage multi k-mers telles que metaSPAdes [67], MEGAHIT [43], Minia [12] et IdbaUD [70] comme les plus performantes pour l'assemblage métagénomique. Toutefois, cette étude a aussi montré la difficulté pour tous ces assembleurs de distinguer des séquences proches, partageant plus de 95% d'identité.

2.3.3 Analyse taxonomique directe, sans assemblage

Dans l'objectif de réaliser l'analyse taxonomique d'un échantillon environnemental, de nouvelles approches, telles que Kraken [101], proposent d'assigner directement chacune des lectures des jeux de données de séquençage métagénomique complet.

Les approches traditionnelles pour l'analyse de données d'amplicon commencent par partitionner les lectures en OTU avant d'assigner taxonomiquement la séquence représentative de chaque OTU. Ces nouvelles approches proposent d'inverser ce paradigme en commençant par assigner directement chaque lecture avant de créer des partitions sur la base de ces assignations.

Kraken est le premier logiciel publié à s'inscrire dans cette lignée. Il utilise des séquences de référence connues pour construire une base de données de k-mers discriminants pour chaque niveau taxonomique. Une lecture est alors assignée sur la base de sa composition en k-mers et, plus précisément, elle est assignée au plus récent ancêtre commun des taxons désignés par les k-mers de cette lecture.

Une autre approche, représentée par MetaPhlan [85] et MetaPhyler [50], consiste à identifier des gènes discriminants pour chaque niveau taxonomique de l'arbre du

vivant. Ainsi, une lecture correspondant à un gène présent seulement dans un genre précis pourra être assignée à ce genre.

Dans tous les cas, une fois l'assignation des lectures réalisée, ces lectures sont ensuite regroupées dans des partitions correspondant aux différents taxons, ce qui permet de quantifier l'abondance de chacun de ces taxons.

Dans le cadre des données de séquençage métagénomique complet, le principal défaut de ces approches visant à assigner directement les lectures brutes concerne le manque de signal phylogénétique porté par ces lectures, qui ne dépassent généralement pas les quelques centaines de nucléotides.

2.3.4 Reconstruction de marqueurs conservés pour l'analyse taxonomique

Dans toutes les approches décrites précédemment, les obstacles à une assignation taxonomique plus précise des échantillons biologiques sont les taux d'erreur des lectures et le manque de signal phylogénétique exploitable. En effet, si on disposait de séquences reconstruites longues (pleine taille), avec très peu d'erreurs, et sans fusionner les séquences proches, les méthodes d'assignation taxonomique actuelles seraient capables d'assigner ces séquences au niveau de l'espèce, voire de la souche lorsque des polymorphismes du gène ciblé existent dans les populations. Une solution pour obtenir de telles séquences serait d'*assembler* les lectures issues d'un séquençage métagénomique complet et correspondant au marqueur ciblé. C'est le sujet de cette thèse et c'est ce que tente de faire la méthode que nous avons développée.

Nous montrons ici comment la stratégie d'assemblage des lectures d'un marqueur issues d'un séquençage métagénomique complet a déjà été abordée par deux outils existants. Nous décrirons leurs points forts, leurs points faibles, et les voies potentielles d'amélioration.

2.3.4.1 Le problème

Le problème de la reconstruction de séquences de marqueurs conservés, comme l'ARNr SSU, est différent de celui de la reconstruction des génomes en raison de la nature des données. Comme nous l'avons vu au chapitre I, section 1.1.5, l'ARNr SSU est présent et conservé dans tous les organismes cellulaires. Deux séquences d'espèces proches peuvent donc ne différer que de quelques nucléotides. En outre, à cause de la pression de sélection sur certaines régions, deux séquences d'espèces relativement distantes peuvent partager des régions parfaitement identiques, alors que le reste de leur séquence diverge. Il y a donc un risque élevé de créer des séquences chimériques. Un autre écueil est l'hétérogénéité de l'abondance. Enfin, contrairement à l'assemblage génomique pour lequel on cherche à assembler un génome séquencé avec une couverture « relativement » homogène, l'abondance des espèces d'un échantillon métagénomique

peut varier fortement. Dans ces conditions, lorsque l'abondance d'une espèce devient trop faible, il n'est plus possible de faire la différence entre un site divergé et une erreur de séquençage.

Pour s'attaquer au problème de la reconstruction de marqueurs conservés dans des données de métagénomique, des méthodes ont été développées récemment et implémentées dans des outils tels que EMIRGE et REAGO.

2.3.4.2 Identification des lectures de marqueurs conservés

Afin de pouvoir assembler ou assigner les séquences d'un marqueur conservé donné, la première étape consiste couramment à identifier et isoler les lectures correspondant à ce marqueur dans un jeu de séquençage métagénomique complet. Par exemple, l'ARN ribosomique correspond à environ 0,1% des lectures génomiques, et jusqu'à 80-90% des lectures d'ARN total. Isoler les lectures d'ARN ribosomique n'est d'ailleurs pas un problème spécifique de l'assemblage de marqueurs dans des données de métagénomique. Il s'agit d'une étape classique préalable à toute analyse RNA-seq, qui est intégrée par défaut dans la majorité des pipelines d'analyse.

SortMeRNA, filtrer les lectures d'ARNr Le logiciel SortMeRNA a été développé au sein de l'équipe BONSAI par Evguenia Kopylova, et publié en 2012 [38]. Dans sa première version, il s'agit d'un outil rapide et sensible qui permet de filtrer les lectures correspondant à un marqueur conservé dans un jeu de données de séquençage. Il a notamment été évalué sur l'ARN ribosomique, et surpasse tous les autres outils dédiés sur le plan du temps de calcul et de la sensibilité. Ses performances sont le résultat de la combinaison d'une structure de données spécialisée de type burst-trie pour stocker l'information de séquence issue d'une base de séquences de références, et d'un automate de Levenshtein universel et déterministe, capable de chercher des graines avec erreur dans les lectures de manière très efficace. Une lecture est alors considérée comme appartenant au marqueur si au moins deux graines de 18 nucléotides (avec une erreur maximum, de type substitution ou indel) de la base de séquences de référence sont trouvées dans la lecture.

SortMeRNA 2, aligner les lectures d'ARNr Dans la deuxième version de SortMeRNA (encore non publiée, mais disponible sur GitHub [6]), une étape supplémentaire permet de construire les alignements et de calculer leurs e-valeur. L'utilisateur peut donc disposer de l'ensemble des lectures correspondant au marqueur ainsi que des alignements locaux de ces lectures sur la base de données de séquences de référence. Afin de limiter le nombre d'extensions des alignements à calculer, SortMeRNA implémente une heuristique de sélection des séquences candidates basée sur le nombre de graines identifiées dans la lecture. Les séquences de références sont triées par ordre décroissant

6. https://github.com/biocore/sortmerna

du nombre de graines communes avec la lecture, et l'heuristique permet de sélectionner seulement les meilleures références potentielles, pour lesquelles les extensions d'alignement sont calculées. Cette stratégie est particulièrement efficace pour l'alignement de nombreuses lectures courtes (de type Illumina par exemple) et impacte peu la sensibilité de la recherche. Pour des lectures plus longues (> 300-400 bp), cette heuristique peut limiter la découverte d'alignements optimaux et doit être atténuée ou désactivée dans le cadre d'une recherche sensible.

2.3.4.3 EMIRGE

En 2011, EMIRGE [61] est la première méthode proposant une solution au problème de la reconstruction d'un marqueur conservé à partir de données de séquençage métagénomique complet. C'est une méthode robuste et capable de passer à l'échelle. En outre, depuis sa publication, EMIRGE est continuellement enrichi et son support est assuré par une équipe réactive. Cela en fait un outil populaire auprès des biologistes et des bio-analystes qui souhaitent reconstruire de l'ARNr SSU à partir de données métagénomiques afin de réaliser une analyse taxonomique (cité 159 fois sur Google Scholar).

Son principe repose sur une modélisation bayésienne et un algorithme de « maximisation de l'espérance » (*expectation maximization* [17]) permettant d'ajuster itérativement des modèles des séquences ciblées, sachant les données. Brièvement, l'algorithme procède ainsi :

0. initialisation des séquences modèles à partir des séquences de la base de référence SILVA SSURef partitionnée à 97% ;
1. alignement des lectures sur les séquences modèles ;
2. modification des séquences modèles à l'aide de l'information apportée par les lectures alignées. De manière à maximiser la vraisemblance de ces alignements sachant les séquences modèles, et en fonction des paramètres utilisés, l'algorithme peut modifier, diviser ou supprimer une séquence modèle. De plus, si deux séquences modèles sont trop similaires, elles sont fusionnées en une unique séquence modèle ;
3. jusqu'à un nombre fixé d'itérations, réaliser une nouvelle itération depuis l'étape 1 avec les séquences modèles obtenues à l'itération courante.

Si ce principe est simple, la mise en œuvre demande toutefois de fixer un ensemble de paramètres critiques :

— la probabilité minimale d'un nucléotide alternatif pour être considéré comme un variant (0,1 par défaut) ;

— la fraction minimale de variants par rapport à une séquence modèle nécessaire pour diviser cette séquence modèle en deux pour la prochaine itération (0,04 par défaut) ;

— le seuil d'identité minimum entre deux séquences modèles pour les fusionner (97% par défaut) ;
— la couverture moyenne minimale par des lectures pour garder une séquence modèle (3 par défaut).

Bien que ces paramètres puissent être modifiés pour augmenter la sensibilité d'EMIRGE (nous le montrerons dans le chapitre IV, section 4.1.1.2), la méthode semble optimisée pour reconstruire des séquences avec un seuil de 97% d'identité entre deux séquences proches. De même, EMIRGE peine à reconstruire des séquences d'espèces de faible abondance ($\lesssim$ 10x).

2.3.4.4 REAGO

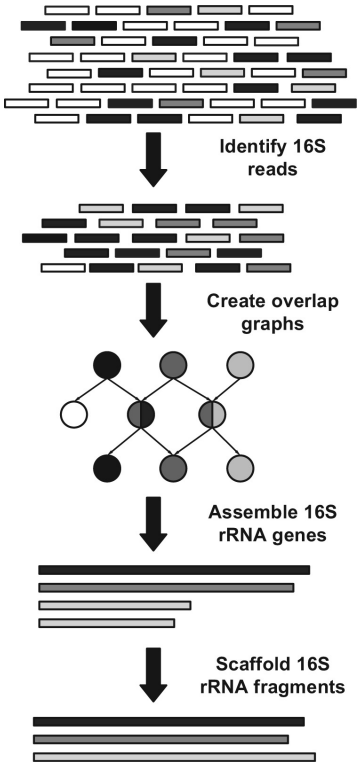


FIGURE 2.8 – Principe du pipeline de REAGO.

Publié en 2015, REAGO [104] propose une nouvelle approche basée sur la construction d'un graphe de chevauchement, à l'instar de l'assemblage génomique classique (2.2). La méthode se déroule en plusieurs étapes (Figure 2.8) :

1. identification des lectures appartenant au marqueur ARNr 16S par leur alignement par Infernal [64] sur un profil probabiliste qui prend en compte la séquence du marqueur et sa structure secondaire ;
2. génération du graphe de chevauchement. Une table de préfixes/suffixes est construite pour l'ensemble des lectures, ce qui permet d'identifier tous les chevauchements

sans erreur d'une taille suffisante (≥80% de la taille des lectures par défaut) entre toutes les paires de lectures ;

3. nettoyage et simplification itérative du graphe de chevauchement. Les chaînes linéaires de nœuds sont compressées, les chemins alternatifs de faible poids sont supprimés, les bulles et les *tips* sont supprimés ;
4. suppression des arêtes chimériques identifiées par l'assignation des nœuds du graphe nettoyé au niveau du genre ;
5. extraction des contigs. À cette étape, l'information d'appariement des lectures est utilisée pour sélectionner des chemins dans le graphe ;
6. scaffolding des petits contigs. L'information de positions des lectures appariées sur le profil est finalement utilisée pour positionner les petits contigs sur ce profil et joindre les contigs connectés par des paires de lectures.

Les auteurs ont évalué REAGO contre EMIRGE et une sélection d'assembleurs métagénomique (IDBA, MetaVelvet), et décrivent de meilleurs résultats pour ce qui est de la reconstruction de séquences à 98% d'identité près.

2.3.4.5 Comparaison expérimentale d'EMIRGE et REAGO

Dans le cadre de cette thèse, nous avons évalué différents assembleurs, dont EMIRGE et REAGO, sur plusieurs jeux de données de test, et les protocoles et résultats de ces évaluations sont décrits en détails au chapitre IV. Nous présentons ici un extrait de ces résultats focalisé sur EMIRGE et REAGO et l'assemblage d'un jeu de données de séquençage d'une communauté synthétique, dont les organismes connus ont été cultivés puis rassemblés en laboratoire. Nous espérons ainsi donner au lecteur une idée plus précise de l'état de l'art sur la reconstruction de marqueurs conservés, ainsi qu'une intuition des pistes d'amélioration possibles.

Le jeu de données Le jeu de données d'évaluation est issu du séquençage métagénomique complet Illumina d'une communauté synthétique, composée de 16 espèces d'archées et de 48 espèces de bactéries, et publiée par Shakya et al. en 2013. Au total, cette communauté comprend 106 séquences distinctes d'ARNr 16S, avec des similarités deux à deux qui varient de 59,64% à 99,93% d'identité.

Le jeu de données de séquençage Illumina comprend 109 millions de lectures appariées de 101 bp, et la profondeur de séquençage moyenne des différentes espèces varie entre 6x et 318x. La description complète du jeu de données peut être trouvée au chapitre IV, section 4.3.

Comparaison des résultats EMIRGE et REAGO ont été paramétrés pour maximiser leur sensibilité (4.1.1) et les assemblages ont été évalués avec MetaQUAST [60], lequel a réaligné les contigs de chaque assemblage sur les 106 séquences d'origine (4.1.3).

TABLE 2.1 – Statistiques d'assemblage avec EMIRGE et REAGO pour la communauté synthétique.

	Nb contigs	FR (%)	TE (%)	Ns (%)	Temps	Charge CPU (16 cœurs)	Mémoire (RAM)
EMIRGE	82	50,7	0,17	1,12	21 h 17	1550%	3,3 Go
REAGO	59	42,8	0,06	0	53 min 36 s	1267%	27,54 Go

Les résultats sont donnés en Table 2.1. En regardant la première colonne (le nombre de contigs) et les trois dernières colonnes (temps, charge CPU et mémoire), il apparait d'emblée que les deux programmes produisent des résultats différents avec une utilisation des ressources également très différente. EMIRGE trouve 82 contigs, contre seulement 59 pour REAGO. EMIRGE reconstruit donc plus de séquences que REAGO, ce qui est en contradiction avec les résultats annoncés par les auteurs de REAGO. REAGO prend près de 20 fois moins de temps qu'EMIRGE pour réaliser la reconstruction, mais sa consommation mémoire est presque 10 fois plus importante.

Pour mieux apprécier la qualité des résultats, nous avons introduit trois métriques :

— *FR* : la fraction reconstruite, qui correspond au nombre de nucléotides des séquences d'origine couverts par des contigs, divisé par la taille totale des séquences d'origine ;

— *TE* : le taux d'erreur, qui correspond au pourcentage de substitutions et d'*indels* observés par rapport à la séquence la plus proche dans l'échantillon d'origine ;

— *Ns* : le taux de nucléotides inconnus.

Ce sont les colonnes 2 à 4 de la table. Le taux d'erreur des séquences reconstruites par REAGO est bien plus faible que celui d'EMIRGE. On remarque toutefois qu'aucune des deux méthodes n'arrive à reconstruire plus de 50% des séquences originales. Une exploration plus précise des résultats montre que lorsque plusieurs séquences originales proches ($>$ 97-98% d'identité) existent dans le jeu de données, les deux méthodes ont tendance à les fusionner et à ne reconstruire qu'une seule séquence consensus.

Cette brève illustration des résultats d'EMIRGE et REAGO sur l'assemblage d'un jeu de données de séquençage métagénomique complet montre clairement que, bien que ces méthodes permettent déjà de reconstruire des séquences de marqueurs conservés avec une qualité appréciable, il existe des voies d'amélioration possibles. Notamment, on aimerait être capable de différencier les séquences d'espèces proches à plus de 97% d'identité de manière à pouvoir réaliser une analyse taxonomique plus fine et mieux résolue des échantillons. Cette amélioration de la résolution de la reconstruction doit aussi s'accompagner du maintien de très faibles taux d'erreurs, ainsi que de l'utilisation d'algorithmes performants pour analyser de gros jeux de données de séquençage

métagénomique complet.

2.3.5 Conclusion sur les méthodes d'analyse de données métagénomique

Il existe de nombreuses approches différentes et complémentaires pour l'analyse d'échantillons environnementaux, chacune avec ses avantages et ses limites.

La métagénomique ciblée permet aujourd'hui de facilement réaliser l'analyse taxonomique d'un échantillon. C'est un processus bien maîtrisé, avec des méthodes de séquençage et d'analyse performantes et rodées. Mais cette approche possède également des biais qui limitent de manière intrinsèque sa capacité à caractériser finement la composition taxonomique d'un échantillon, avec une précision allant jusqu'au niveau de l'espèce.

A contrario, les données de séquençage métagénomique complet sont peu biaisées et contiennent l'ensemble du signal taxonomique, avec l'intégralité des génomes des organismes du microbiote. Toutefois, l'analyse de ces données reste difficile et les méthodes actuelles font face à des limitations certaines, que ce soit par l'assemblage de l'ensemble des lectures ou bien par l'assignation taxonomique directe de ces lectures.

Face à cette situation, une voie intermédiaire propose de se focaliser sur l'analyse de marqueurs conservés dans des données de séquençage métagénomique complet. Il s'agit d'un problème moins ambitieux, pour lequel des méthodes telles qu'EMIRGE ou REAGO établissent la faisabilité. Ces deux programmes permettent d'obtenir des résultats encourageants avec la reconstruction de marqueurs pleines tailles, même s'il reste encore une marge d'amélioration visible, que ce soit en termes de qualité de résultats ou en termes de charge computationelle. C'est dans cette voie que s'inscrivent nos contributions, présentées au chapitre III.

Chapitre 3

MATAM, la méthode

Nous avons vu en conclusion du chapitre II que la reconstruction de marqueurs ciblés dans les jeux de données de séquençage métagénomique complet était une approche pertinente, qui ouvrait la voie à de possibles améliorations pour l'assignation taxonomique. C'est dans ce contexte que nous avons développé l'outil MATAM, dont la présentation fait l'objet de ce chapitre. Nous décrivons tout d'abord le schéma général de la méthode. Puis, nous expliquons le détail de la méthode d'une part, et son implémentation d'autre part. Finalement, nous illustrons le déroulement de la méthode sur un jeu de données de test, construit pour être pédagogique et représentatif. Dans le chapitre suivant, le chapitre IV, nous analyserons le comportement de MATAM en situation réelle, sur des jeux de données réalistes.

3.1 Schéma général de la méthode

MATAM, pour *Mapping-Assisted Targeted Assembly for Metagenomics* (« assemblage ciblé assisté par alignement pour la métagénomique »), assemble les séquences d'un marqueur phylogénétique conservé à partir d'un jeu de données de séquençage métagénomique complet de type *shotgun*.

3.1.1 Le choix du marqueur conservé

Nous avons développé MATAM comme un outil généraliste capable d'assembler un marqueur phylogénétique quelconque. Cependant, dans le reste de ce manuscrit, nous illustrerons son application à la reconstruction de l'ARNr SSU, c'est-à-dire l'ARNr 16S chez les procaryotes et 18S chez les eucaryotes (1.1.5.1). L'ARNr SSU est en effet le candidat idéal pour le développement d'une telle méthode, et cela pour trois raisons. Il s'agit du marqueur de référence pour l'analyse taxonomique, conservé universellement par les espèces cellulaires de l'ensemble des domaines du vivant (2.3.1.1). D'un point de vue algorithmique, il est constitué d'une succession de régions conservées et de ré-

gions hypervariables. L'assemblage d'un tel marqueur est donc un défi, et nous pensons qu'une méthode performante dans ce contexte pourra traiter tout autre type de marqueurs. Enfin, de grandes bases de données de séquences de qualité sont disponibles, telles que SILVA [75] ou Greengenes [19].

3.1.2 Données en entrée

La méthode prend en entrée les lectures d'un jeu de données de séquençage métagénomique complet et une base de séquences de référence pour un marqueur donné.

Les lectures de séquençage La technologie de séquençage Illumina est actuellement la plus populaire dans les projets de métagénomique (1.3.3), en raison de son coût, de son débit et du faible taux d'erreur (1.2.3.3). De plus, les erreurs générées par la technologie Illumina sont exclusivement du type substitution, ce qui simplifie le problème d'assemblage. Nous avons donc fait le choix de nous focaliser sur l'assemblage de jeux de données de séquençage de type Illumina, c'est-à-dire pouvant contenir des millions de lectures, de tailles comprises entre 100bp et 250bp, avec un taux d'erreur largement inférieur à 1%, et sans erreur de type *indel*. De plus, nous recommandons aux utilisateurs de fournir des jeux de données de séquençage de la meilleure qualité possible et préalablement nettoyés (2.2.5).

La base de référence MATAM prend en entrée une base de données de séquences de référence pour le marqueur. Dans l'idéal, cette base doit représenter une diversité d'organismes la plus large possible, et contenir des séquences de qualité. Pour notre application à l'assemblage de l'ARNr SSU, nous avons choisi d'utiliser la base *SILVA 128 SSURef NR99* (désignée *SSURef complète* dans la suite du manuscrit), qui contient 645 151 séquences de référence de haute qualité. Cette base de référence a été dérépliquée par un partitionnement des séquences avec un critère de 99% d'identité, à l'exception des séquences des espèces cultivées qui ont été toutes préservées dans la base. Le détail des étapes de nettoyage et de contrôle qualité de cette base peut être trouvé sur le site du projet : https://www.arb-silva.de/projects/ssu-ref-nr/

3.1.3 Résultats en sortie

Les deux sorties principales de MATAM sont :

— des séquences assemblées correspondant au marqueur taxonomique ciblé, dont chacune peut idéalement être affectée à une espèce présente dans l'échantillon environnemental ;

— une visualisation de la composition taxonomique de l'échantillon, calculée à partir de l'abondance estimée et de l'assignation taxonomique de chaque séquence reconstruite.

3.1.4 Les étapes de MATAM

La méthode se compose de quatre étapes majeures, illustrées dans la Figure 3.1 :

i identification des lectures d'ARNr SSU et alignement sur des séquences de référence ;

ii construction d'un graphe de chevauchement de lectures ;

iii compression du graphe, identification et assemblage des composantes ;

iv reconstruction des séquences pleines tailles.

L'enchaînement de ces quatre étapes est semblable au schéma général de REAGO (Figure 2.8 en section 2.3.4.4). La mise en œuvre de chacune des étapes diffère toutefois de manière significative :

— l'étape de filtrage des lectures ne nécessite pas la construction d'un profil probabiliste ;

— MATAM utilise l'information d'alignement des lectures, résultant de l'étape de filtrage, pour réduire l'espace de recherche des paires de lectures à comparer, alors que REAGO n'utilise l'alignement contre un profil probabiliste que pour identifier les lectures correspondant au marqueur. Pour REAGO, les chevauchements sont ensuite calculés *de novo* comme dans les approches classiques à base de graphe OLC ;

— dans REAGO, l'assemblage des contigs consiste à chercher les chemins les plus probables dans le graphe de chevauchement complet. Notre approche consiste à réduire le problème d'assemblage métagénomique en un ensemble de sous-problèmes d'assemblage, plus facile à résoudre, par partitionnement préalable des données sous forme de composantes du graphe complet.

— la reconstruction des séquences pleines tailles des marqueurs utilise un scaffolding original, informé par une base de référence.

3.2 Détail de la méthode

3.2.1 Identification des lectures d'ARNr et alignement sur une base de référence

La première étape de MATAM consiste à aligner les lectures d'un jeu de séquençage métagénomique complet de type *shotgun* sur une base de séquences de référence partitionnée. Ceci permet d'obtenir un partitionnement des lectures en les distribuant à gros-grain dans l'arbre du vivant, tout en positionnant chaque lecture dans la longueur du marqueur.

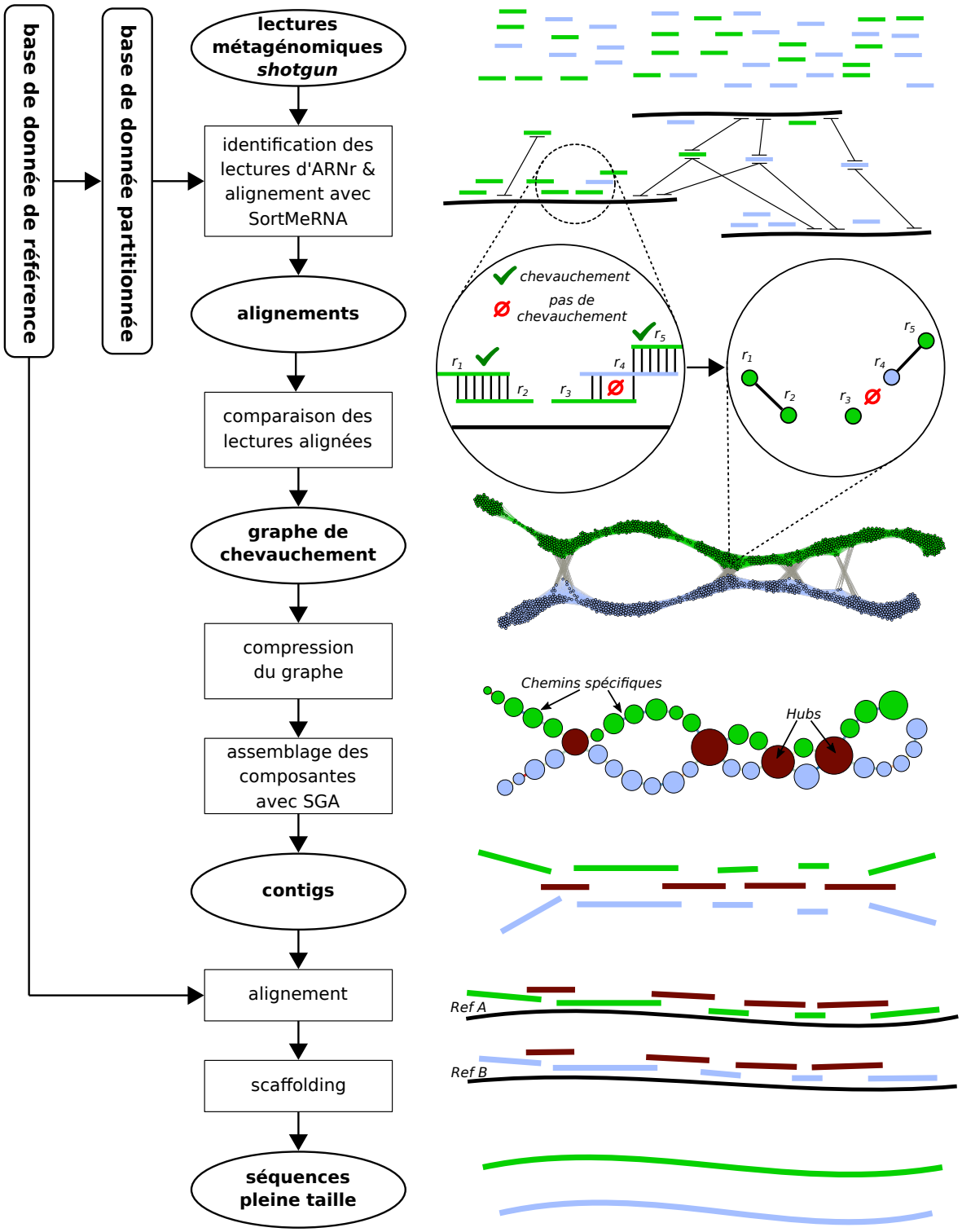


FIGURE 3.1 – Schéma général de MATAM. Les étapes sont décrites à gauche et illustrées à droite pour le cas de deux espèces (en bleu et vert).

3.2.1.1 Construction de la base de référence partitionnée

Préalablement à l'alignement des lectures, nous construisons une base de séquences de référence partitionnée avec SUMACLUST [58] (jusqu'à la version 0.9.9 de MATAM, ensuite avec l'algorithme de UCHIME [24] implémenté dans VSEARCH [78]).

1. les séquences de référence sont triées par taille décroissante ;
2. les partitions sont construites itérativement avec un seuil d'identité de 95%, de manière analogue au partitionnement d'amplicons (2.3.1.2) ;
3. la base de référence partitionnée est constituée des séquences représentatives des partitions.

Pour notre application à l'ARNr SSU, nous avons ainsi généré la base de référence partitionnée *SILVA 128 SSURef NR95* (désignée *NR95* dans la suite du manuscrit).

3.2.1.2 Alignement des lectures sur la base de référence partitionnée

Nous utilisons SortMeRNA v2.1 pour identifier les lectures correspondant au marqueur et les aligner sur la base de référence partitionnée (voir 2.3.4.2). En pratique, pour un séquençage métagénomique complet, on s'attend à ce que les lectures d'ARNr correspondent à une proportion de l'ordre de 0,1% à 0,01% de l'ensemble des lectures.

SortMeRNA 2 fonctionne en deux temps :

1. identification des lectures correspondant au marqueur et proposition d'alignements candidats sur la base de la présence d'un minimum de graines partagées : deux graines de 18 nucléotides avec une erreur par défaut ;
2. extension de chaque alignement par programmation dynamique et calcul d'une e-valeur. Les alignements avec une e-valeur suffisante ($< 10^{-5}$ dans notre cas) sont gardés.

En pratique, une lecture provenant d'une région (très) conservée peut potentiellement s'aligner correctement sur un (très) grand nombre de séquences de référence. À l'opposé, une lecture provenant d'une région spécifique (c'est à dire propre à un clade récent) ne s'alignera que sur une ou deux séquences de références. Des heuristiques supplémentaires de SortMeRNA permettent de limiter le nombre d'alignements candidats ainsi que de sélectionner seulement les meilleurs alignements finaux (c'est-à-dire ceux avec les scores les plus élevés). Pour MATAM, nous gardons jusqu'aux 10 meilleurs alignements de chaque lecture, et nous les sauvegardons au format SAM (2.1.4).

```
$ sortmerna --ref NR95.fa,NR95 --reads lectures.fq --aligned alignements_lectures --
    fastx --sam --best 10 --min_lis 10 -e 1e-5
```

3.2.1.3 Sélection des alignements informatifs

Pour une lecture donnée, les alignements générés ne sont pas forcément tous utiles, bien que de e-valeur significative. Nous souhaitons garder seulement les alignements

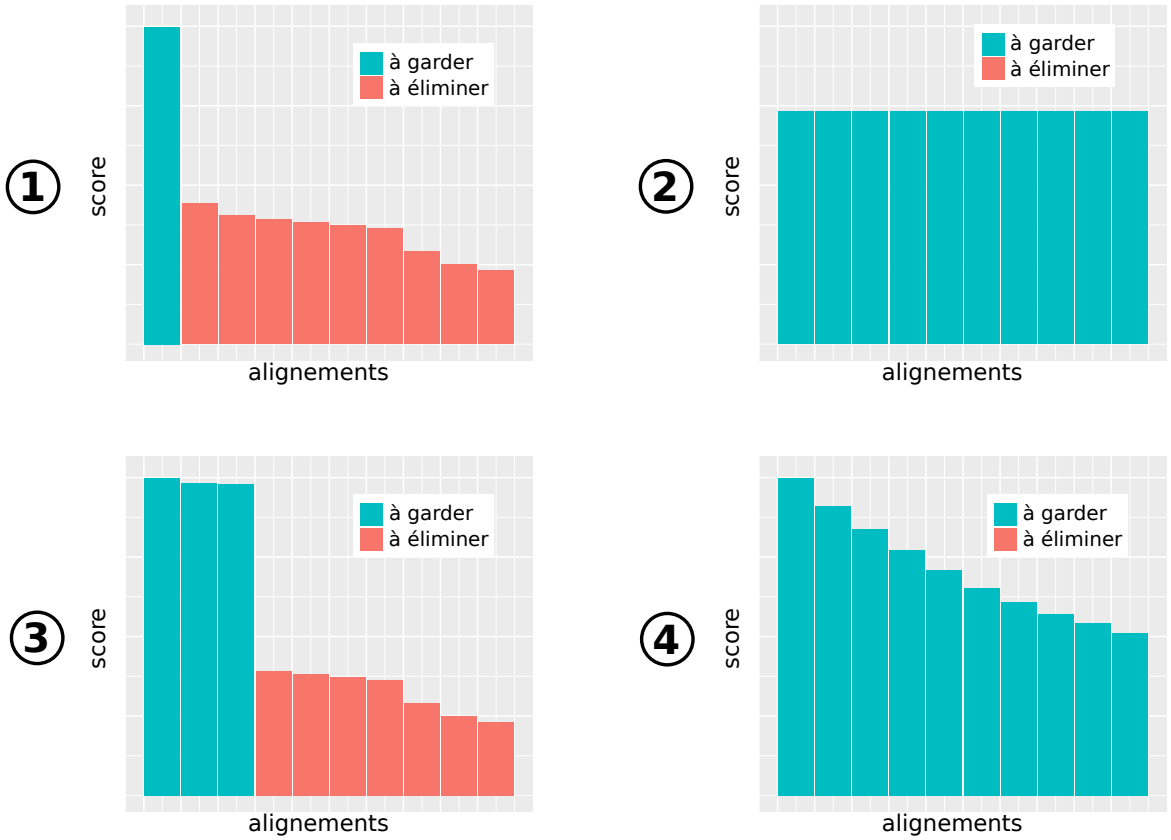


FIGURE 3.2 – Différents cas d'alignement d'une lecture contre plusieurs références. Les 10 alignements sont distribués sur l'axe des abscisses par scores décroissants, et leurs scores respectifs sont indiqués sur l'axe des ordonnées. 1) lecture spécifique, qui ne s'aligne correctement que sur une seule séquence de référence. 2) lecture conservée universellement, qui s'aligne avec le même score sur toutes les références. 3) lecture conservée au niveau d'un clade intermédiaire, qui s'aligne avec des scores comparables sur quelques séquences de références. 4) cas théorique problématique, dans lequel on observe une décroissance continue des scores des alignements.

informatifs pour chaque lecture. Plusieurs cas sont possibles, décrits Figure 3.2 :

— pour une lecture spécifique, ou conservée dans une partie seulement des références (cas 1. et 3., respectivement), nous ne gardons que les meilleurs alignements ;

— pour une lecture conservée universellement par toutes les références (cas 2.), nous gardons tous les alignements ;

— le seul cas, théorique (cas 4.), pour lequel nous préférons ne pas éliminer d'alignement consisterait en une lecture dont les scores d'alignement ne permettent pas de distinguer un seuil clair entre les références proches et celles plus distantes.

Nous employons donc une approche conservatrice qui consiste à identifier des

« ruptures » dans la distribution des meilleurs scores, ce qui permet de n'éliminer les alignements que s'il est clair qu'ils ne sont pas informatifs :

1. les alignements de chaque lecture sont triés par score décroissant ;
2. les alignements sont gardés itérativement au moyen d'un seuil géométrique. L'alignement A_{i+1} est gardé si $score(A_{i+1}) \geq score(A_i) * facteur$, avec un facteur égal à 0,9 par défaut.

3.2.2 Construction d'un graphe de chevauchement de lectures

La deuxième étape de MATAM consiste en la construction d'un graphe non-orienté de chevauchements des lectures. Cette étape est traditionnellement coûteuse en temps et nécessite généralement l'emploi d'heuristiques (2.2.3). Nous proposons ici de tirer parti des alignements des lectures sur les séquences de référence pour accélerer la comparaison des paires de lectures. Le principe repose sur l'idée que deux lectures chevauchantes provenant de la même espèce devraient s'aligner sur les mêmes séquences de références, et à des positions proches. Les séquences de références sont alors utilisées comme des guides qui permettent de positionner les lectures dans la longueur du marqueur. Deux lectures sont dites *compatibles* si elles s'alignent sur la même séquence de référence et qu'il existe un chevauchement entre les deux lectures suffisamment grand et similaire. Cette stratégie permet notamment de restreindre la comparaison aux paires de lectures qui proviennent potentiellement de la même région du marqueur, contrairement aux approches traditionnelles (2.2.3) qui ne tiennent compte que de l'information de séquence.

3.2.2.1 Détail de l'algorithme de construction du graphe de chevauchement

Nous disposons des lectures correspondant au marqueur, ainsi que des alignements informatifs de ces lectures au format SAM. Pour comparer les paires de lectures, nous faisons appel à une série de filtres successifs. Les premiers filtres sont peu couteux et capables d'éliminer rapidement une grande proportion des paires candidates alors que les derniers filtres, plus couteux, permettent de confirmer définitivement la compatibilité de deux lectures.

1. on ne considère que les paires de lectures qui s'alignent sur une même séquence de référence ;
2. on élimine les paires de lectures dont le chevauchement des alignements est inférieur à une taille minimale m. Cette information est déduite du CIGAR ;
3. sachant l'alignement des paires de lectures chevauchantes, on élimine les paires de lectures dont l'alignement est inférieur à un seuil d'identité i. L'alignement des deux lectures se calcule en temps linéaire (de l'ordre du nombre de bases dans le chevauchement) en parcourant conjointement les CIGAR des alignements et les séquences des lectures ;

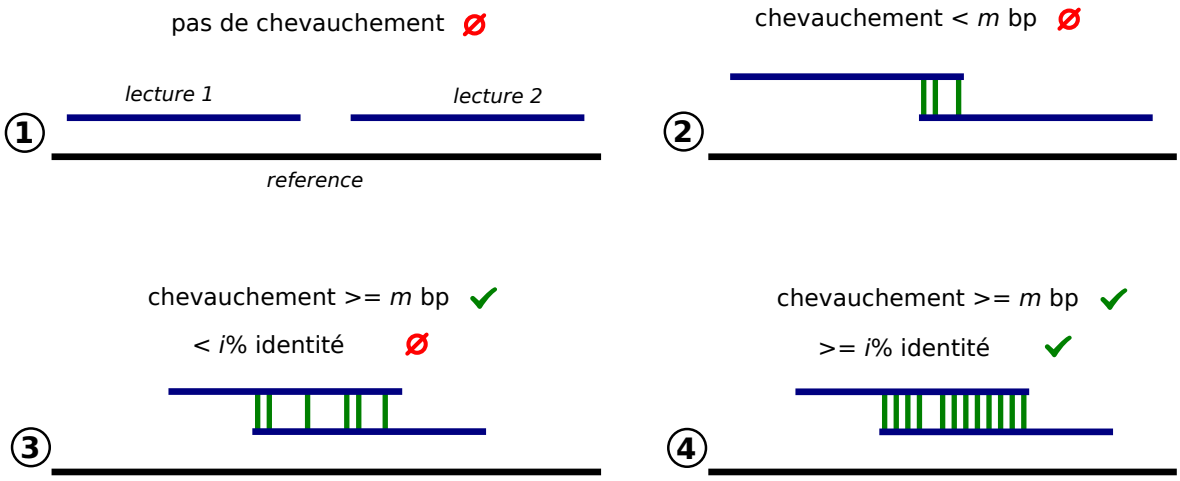
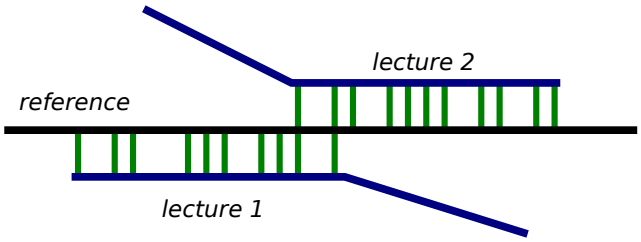


FIGURE 3.3 – Filtres successifs pour la comparaison des alignements de deux lectures. Ces filtres sont paramétrés par la taille minimale m et le seuil mimimum d'identité i du chevauchement des deux lectures.

A l'issue de ce filtrage, on considère donc comme *compatibles* les paires de lectures alignées sur une même référence, avec un chevauchement de taille supérieure à m et d'identité supérieure à i. Ces paires de lectures sont liées dans le graphe de chevauchement. Typiquement, la taille minimale du chevauchement m est de l'ordre de plusieurs dizaines de nucléotides et le seuil d'identité minimum i dépend de la méthode de séquençage utilisé, et peut varier entre 90 et 100% d'identité.

3.2.2.2 Implémentation et pièges liés aux alignements locaux



(a) Chevauchement trop court de deux lectures partiellement alignées sur la séquence de référence.

```
Lecture_1   ...TTT..AAAGGG...
Lecture_2   ...TTTAAA..GGG...
Reference   ACGTTTAAAAGGGACT
```

(b) Positions différentes du même *indel* dans deux alignements, entraînant la création d'*indels* faussement spécifiques à chacune des deux lectures.

FIGURE 3.4 – Pièges de comparaison des lectures liés aux alignements locaux

En pratique, les alignements dont nous disposons à cette étape sont les alignements locaux calculés par SortMeRNA. Cela se traduit par des contraintes supplémentaires à prendre en compte au moment de la comparaison des alignements (Figure 3.4) :

— les lectures peuvent n'être que partiellement alignées sur la référence. La comparaison des alignements de deux lectures partiellement alignées peut donc sous-estimer la taille du chevauchement entre ces deux lectures (Figure 3.4a) ;

— si une lecture s'aligne sur la référence avec un *indel* dans un homopolymère, cet *indel* peut être introduit à plusieurs positions possibles. La comparaison des alignements de deux lectures possédant un tel *indel* peut entraîner à tord la création d'*indels* spécifiques à chaque lecture (Figure 3.4b).

L'implémentation de notre algorithme de comparaison nécessite donc l'adaptation de certains filtres afin de ne pas omettre ces cas problématiques de chevauchement compatibles :

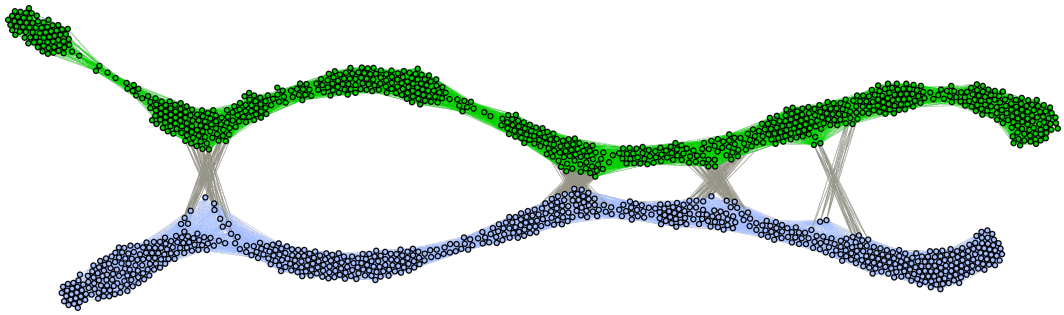
1. les deux lectures doivent s'aligner sur la même séquence de référence ;
2. la paire de lectures est gardée si les alignements des deux lectures partagent au moins une position commune sur la référence. C'est-à-dire que le chevauchement des alignements doit être de taille supérieure ou égale à 1 bp ;
3. sur ce chevauchement, les deux lectures doivent être similaires, avec une identité de l'ordre du seuil minimal i, mais un peu plus de tolérance peut être accordée aux différences de type *indel* ;
4. pour toutes les paires de lectures satisfaisant les conditions précédentes, l'alignement semi-global des deux lectures est calculé. Le chevauchement ainsi obtenu doit être de taille supérieure ou égale à m, avec une identité supérieure ou égale à i.

La taille minimum de chevauchement m et le seuil d'identité i sont des paramètres, qui peuvent être modifiés par l'utilisateur.

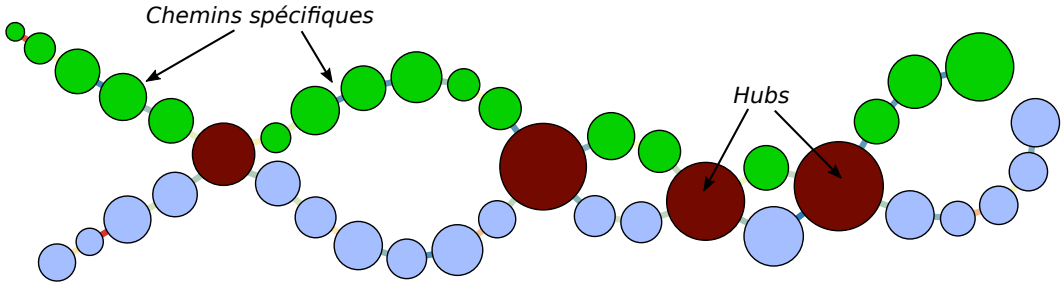
3.2.3 Compression du graphe de chevauchement, identification et assemblage des composantes

3.2.3.1 Propriétés du graphe de chevauchement

Dans le cas de l'application à l'ARNr SSU, les graphes de chevauchement obtenus présentent une topologie caractéristique. Nous présentons dans la Figure 3.5a l'exemple d'un graphe de chevauchement obtenu à partir des lectures de deux espèces (en vert et bleu) et représenté à l'aide d'un algorithme de visualisation basé sur les forces (*Force Atlas 2*) implémenté dans Gephi [6]. Ces graphes de chevauchement possèdent des structures très marquées. On définit ainsi trois types de *régions* :



(a) Graphe de chevauchement



(b) Graphe compressé

FIGURE 3.5 – Graphe de chevauchement et graphe compressé pour 2 espèces (ici en vert et bleu)

— les régions spécifiques de chaque espèce sont représentées sous la forme de sous-graphes denses qui possèdent une méta-structure linéaire et forment des « *faisceaux* » ;

— les régions conservées correspondent à des sous-graphes plus denses, qui relient les régions spécifiques au seins de « *hubs* » où les lectures issues de plusieurs espèces sont identiques et connectées de manière indifférenciée ;

— une petite proportion des lectures correspond à des nœuds d'arité nulle, des *singletons*. Ces nœuds ne sont pas représentés sur la Figure 3.5a.

Formellement, de par le mode de construction du graphe de chevauchement, chaque sous-graphe est un graphe d'intervalle de la séquence de référence sous-jacente (spécifique ou conservée), avec des intervalles de la taille des lectures. La densité locale de ces sous-graphes est directement proportionnelle à la couverture par les lectures de séquençage de la région du marqueur correspondante.

3.2.3.2 Compression du graphe de chevauchement

Afin d'exploiter efficacement le graphe de chevauchement construit à l'étape précédente, nous souhaitons le simplifier et isoler ses régions spécifiques et conservées. Dans ce but, nous calculons un graphe de chevauchement *compressé* dont un exemple

est montré Figure 3.5b.

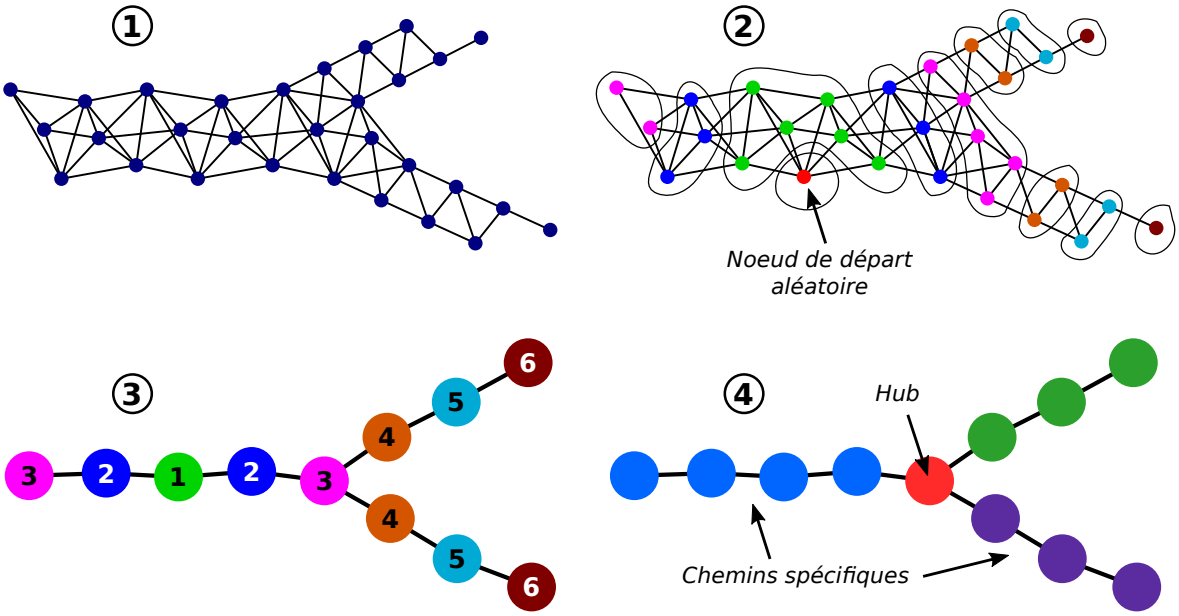


FIGURE 3.6 – Etapes de la compression du graphe de chevauchement. 1) Soit un graphe de chevauchement de lectures. 2) Un parcours en largeur à partir d'un nœud de départ aléatoire permet d'annoter chaque nœud avec leur profondeur. 3) Les nœuds à même profondeur et appartenant à la même composante connexe sont regroupés au sein d'un unique nœud compressé, et les arêtes sortantes sont fusionnées. 4) Le graphe compressé est partitionné en 3 types de sous-graphes (*chemins spécifiques*, *hubs*, *singletons*).

Exploitation du graphe, parcours en largeur Nous compressons le graphe de chevauchement au moyen d'un parcours en largeur (*BFS* pour *Breadth First Search*), illustré dans la Figure 3.6. Ce parcours en largeur part d'un nœud de départ choisi aléatoirement et permet d'annoter tous les nœuds du graphe avec une profondeur calculée comme la distance à partir de ce nœud initial. Les nœuds à même profondeur et appartenant à la même composante connexe sont ensuite regroupés au sein d'un unique nœud compressé, et les arêtes sortantes sont fusionnées. Un filtre additionnel permet de supprimer du graphe compressé les nœuds et arêtes de trop faible poids. Par défaut, nous supprimons les nœuds ne contenant qu'une seule lecture, et les arêtes ne représentant qu'un seul chevauchement.

Propriétés du graphe compressé Nous présentons dans la Figure 3.5b l'exemple du graphe compressé obtenu à partir du graphe de chevauchement des lectures de deux espèces (en vert et bleu). Dans ce graphe compressé, les nœuds colorés en vert ou bleu ne contiennent que des lectures de l'espèce correspondante, alors que les nœuds

colorés en brun contiennent des lectures des deux espèces. Le graphe compressé ainsi obtenu est quasi-planaire et de plusieurs ordres de grandeur plus petit que le graphe de chevauchement initial. De plus, la structure du graphe compressé correspond bien à l'organisation en régions du graphe de chevauchement. Il semble donc que l'approche retenue permette de capturer certaines propriétés du graphe de chevauchement tout en le simplifiant.

3.2.3.3 Identification des composantes

Le graphe compressé peut être partitionné en trois types de composantes :

— les *chemins spécifiques*, qui sont les chaines de nœuds de degré inférieur ou égal à 2, et correspondent aux régions spécifiques ;

— les *hubs*, qui sont composés de nœuds de degré supérieur à 2, et correspondent à des régions conservées ;

— les *singletons*, qui sont les nœuds de degré nul.

A chaque composante est associé un ensemble de lectures, et une lecture du graphe de chevauchement ne peut être que dans une seule composante du graphe compressé.

3.2.3.4 Assemblage des composantes

A ce stade, nous faisons l'hypothèse que chaque composante contient principalement des lectures représentant une seule séquence sous-jacente, spécifique d'une espèce ou d'un groupe d'espèces proches pour les chemins spécifiques, ou bien partagée entre des espèces distantes pour les *hubs*. Par conséquent, l'assemblage des lectures d'une composante peut être réduit à un problème d'assemblage génomique classique, et peut être résolu grâce à l'utilisation d'un assembleur génomique standard (2.2). De plus, l'ARNr SSU ne possède pas de régions répétées, ce qui rend l'assemblage d'autant plus facile.

Choix de l'assembleur Il existe de très nombreux assembleurs génomiques (2.2), et bien que de nouvelles méthodes soient encore développées régulièrement il est possible de choisir parmi un ensemble assez large d'assembleurs très performants. Nous avons sélectionné un assembleur qui correspond à l'état de l'art actuel pour les assembleurs OLC : SGA [90] (2.2.3). Il faut noter que, en fonction des évolutions méthodologiques, il est tout à fait envisageable de remplacer cet assembleur par un assembleur plus performant. La construction modulaire de notre implémentation permettra de le faire facilement.

Paramétrage de SGA SGA est un assembleur OLC modulaire, constitué de multiples étapes organisées en trois phases :

1. Correction d'erreur : un FM-index est construit à partir de l'ensemble des lectures et les erreurs de séquençage sont identifiées à partir des k-mers peu fréquents ;
2. Assemblage des contigs : un graphe de chevauchement de lectures est construit après la suppression des lectures dupliquées ou de faible qualité, et les contigs sont identifiés à partir du graphe ;
3. Scaffolding : les lectures appariées sont ré-alignées sur les contigs, ce qui permet d'orienter les contigs, de calculer les distances entre eux, et de générer des scaffolds.

Pour notre utilisation dans MATAM, nous faisons appel à la phase d'assemblage des contigs (étape 2), et optionnellement, à la phase de correction d'erreur pour les composantes suffisamment couvertes par les lectures de séquençage. Nous n'utilisons pas la phase de scaffolding.

Pour chaque composante du graphe compressé, nous procédons comme suit :

0. extraction des lectures appartenant à la composante ;
1. assemblage des contigs (étape 2 de SGA) ;
2. estimation de la couverture de la composante à partir de la taille totale des lectures et de la taille totale des contigs. $couverture \simeq \frac{tailleLectures}{tailleContigs}$.

   Si la couverture de la composante est suffisante ($> 20x$ par défaut) :
3. correction d'erreur (étape 1 de SGA) sur l'ensemble des lectures de la composante ;
4. assemblage des contigs (étape 2 de SGA).

Cette stratégie en deux temps permet d'éliminer des erreurs de séquençage si la couverture de la composante est suffisante, tout en étant capable d'assembler des composantes faiblement couvertes, dont les lectures seraient écartées par la correction d'erreur si celle-ci était activée dans ce cas.

Propriétés de l'assemblage L'assemblage des lectures des composantes avec SGA fournit un ou plusieurs contigs par composante. Chacun de ces contigs correspond à une séquence spécifique ou au contraire conservée parmi les espèces présentes dans notre échantillon. Toutefois, un contig spécifique ou conservé ne l'est pas forcément dans l'absolu, par rapport à toutes les références connues. Le degré de conservation d'une séquence d'une espèce est déterminé relativement aux autres espèces présentes dans l'échantillon. Par exemple, une séquence conservée entre toutes les bactéries du phylum *Firmicutes* peut être spécifique dans notre assemblage si une seule représentante des *Firmicutes* est présente dans l'échantillon. De même, une séquence spécifique d'une espèce particulière peut être conservée dans notre assemblage si de nombreuse souches de cette espèce sont présentes dans l'échantillon. Par conséquent, s'il existe dans notre échantillon une espèce isolée, distante de toutes les autres espèces de l'échantillon, on

peut trouver un contig unique correspondant qui couvre dans sa totalité ou presque la séquence initiale. A l'opposé, si on retrouve un groupe d'espèces proches dans l'échantillon, l'assemblage des séquences de ces espèces donnera un ensemble de petits contigs correspondants aux régions spécifiques et conservées. Globalement, plus des espèces proches sont présentes dans l'échantillon, et plus l'assemblage des contigs sera fragmenté. Dans ce cas, on a donc encore besoin d'une étape de traitement supplémentaire pour regrouper ces contigs.

3.2.4 Reconstruction des séquences en pleine longueur

L'étape suivante de MATAM vise à reconstruire des séquences plus longues que les contigs, idéalement des ARNr SSU pleine taille, tout en minimisant le risque de création de chimères. C'est ce que nous appelons l'étape de *scaffolding*. Pour cela, nous commençons par réaligner les contigs sur la base de référence SSURef complète, avant de regrouper prudemment ces contigs. Notre approche est conservatrice et préfèrera reconstruire des séquences courtes, fragmentées, mais sans erreur plutôt que de risquer de construire des chimères.

3.2.4.1 Alignement des contigs

Les contigs générés à l'étape précédente sont ré-alignés sur la base de référence SSURef complète, ce qui permet de les positionner dans la longueur du marqueur et d'estimer leur degré de conservation. Cette étape d'alignement est faite en deux temps :

1. alignement exhaustif des contigs sur la base SSURef avec SortMeRNA ;
2. sélection des *meilleurs alignements équivalents* pour chaque contig.

Alignement avec SortMeRNA Les contigs sont alignés sur la base SSURef complète avec SortMeRNA, de la manière la plus sensible possible. L'objectif est d'éviter à tout prix de passer à côté de très bons alignements pour un contig donné. Par conséquent, l'heuristique de sélection des séquences candidates est désactivée (2.3.4.2) à cette étape pour maximiser la sensibilité de la recherche d'alignements. SortMeRNA est donc paramétré pour calculer tous les alignements possibles (option `-num_alignments 0`) contre la base de référence qui passent le filtre sur la e-valeur ($\leq 10^{-5}$). L'utilisation de cette recherche exhaustive est possible à cette étape car le nombre de contig à aligner est extrêmement réduit comparé au nombre initial de lectures.

```
$ sortmerna --ref SSURef.fa,SSURef --reads contigs.fa --aligned alignements_contigs
    --sam --blast "1" --num_alignments 0 -e 1e-5
```

Sélection des meilleurs alignements équivalents Un filtre est appliqué *a posteriori* pour sélectionner les *meilleurs alignements équivalents* parmi l'ensemble des

alignements de chaque contig sur la base de référence. Les *meilleurs alignements équivalents* sont définis comme tous les alignements de scores supérieurs ou égaux à 99% du score du meilleur alignement de ce contig sur la base de référence. Suite à ce filtre, un contig ne possédant qu'un seul alignement sur la base de référence est donc spécifique de l'espèce ou de la souche sur laquelle il s'est aligné. À l'opposé, un contig s'alignant de manière équivalente sur des milliers de séquences de référence correspond à une séquence conservée pour une large partie de l'arbre du vivant.

3.2.4.2 Sélection des alignements pour scaffolding

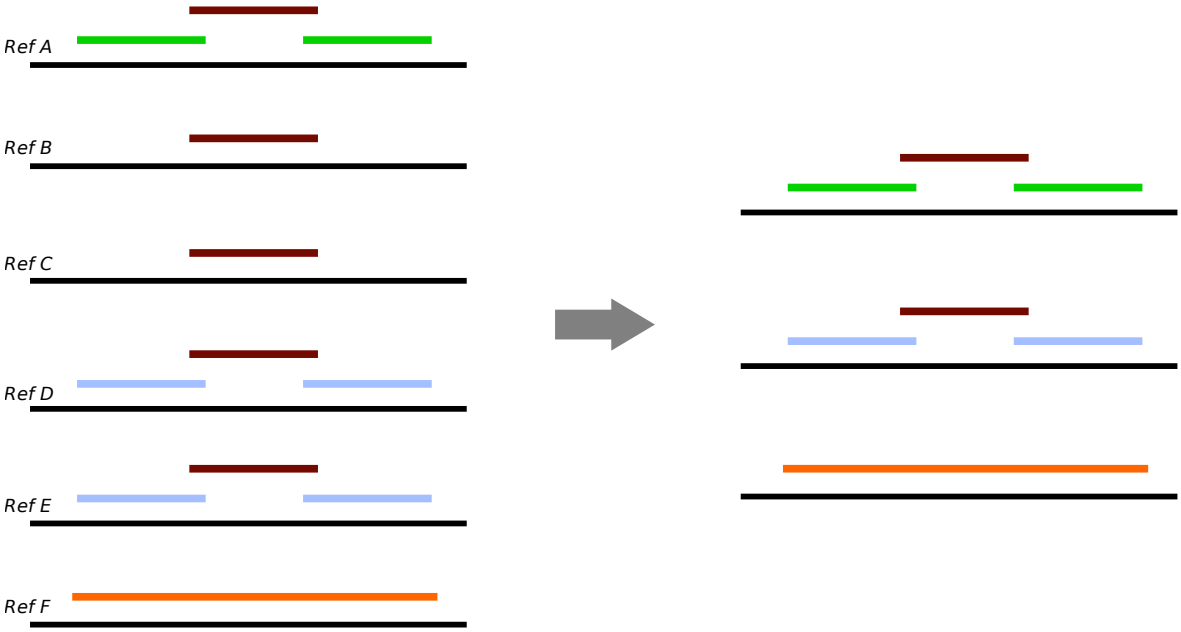


FIGURE 3.7 – Sélection des alignements pour le scaffolding des contigs. Cette étape vise à supprimer la redondance de certains alignements (contigs spécifiques bleu clair), tout en éliminant les alignements non informatifs des contigs conservés (ici, ceux du contig brun sur les références B et C).

À la fin de l'étape précédente, l'information portée par les alignements est encore redondante (Figure 3.7). En effet, comme précédemment évoqué, les contigs très conservés s'alignent sur un grand nombre de références, dont la plupart ne correspondent donc pas à des espèces présentes dans l'échantillon environnemental étudié. De même, les contigs spécifiques d'une espèce du microbiote peuvent s'aligner sur plusieurs séquences de références de cette espèce, si plusieurs séquences de cette espèce (par exemple plusieurs souches, paralogues, allèles, etc.) sont présentes dans la base de référence. Nous procédons donc à une sélection des alignements pour constituer des groupes de contigs représentant probablement la même séquence initiale et que l'on souhaiterait *scaffolder* pour reconstruire des séquences pleine taille. Pour cela, nous proposons une heuristique qui commence par sélectionner les longs contigs spécifiques

qui s'alignent de manière non ambiguë, et qui utilise ensuite les petits contigs conservés pour compléter les *vides*.

Tri des contigs Les contigs sont triés par nombre de meilleurs alignements équivalents croissant, qui est une estimation de la spécificité de ces contigs. Pour un degré de spécificité identique (même nombre d'alignements), et considérant qu'un long contig contient plus de signal phylogénétique qu'un petit, les contigs sont ensuite triés par taille décroissante.

Distribution itérative des alignements À l'aide d'un algorithme glouton itératif, les contigs spécifiques sont assignés sur leur référence la plus proche. Si un contig spécifique s'aligne sur plusieurs références, c'est la référence sur laquelle s'aligne le plus grand nombre de contigs qui sera choisie. Les contigs conservés sont ensuite utilisés pour remplir les espaces entre les contigs spécifiques. Cette approche permet de garantir que deux contigs spécifiques ne seront assignés sur la même référence que s'il s'agit bien de la plus proche dans la base de référence pour ces deux contigs. Par conséquent, plus la base de référence est complète vis-à-vis de l'échantillon analysé, plus le risque de créer des chimères diminue.

3.2.4.3 Génération des scaffolds

Finalement, les scaffolds sont générés par le consensus des contigs alignés sur une même référence.

Représentation des alignements À la suite de l'étape de distribution des alignements, nous disposons d'un fichier d'alignement des contigs sur un ensemble réduit de références. Les alignements sont alors synthétisés sous la forme d'un fichier d'empilement de type *pileup*[1]. Ce format stocke les nucléotides alignés sur chaque position des références et est idéal pour calculer une séquence consensus à partir d'un ensemble d'alignements sur la même référence. Cependant, l'information de continuité des contigs est perdue : on ne peut plus faire le lien entre un nucléotide aligné sur une position de la référence et un autre nucléotide du même contig aligné sur une position différente.

Consensus Dans notre cas, si la base de référence est suffisamment complète, on s'attend à ce que la plupart du temps une seule espèce s'aligne sur une référence. Réaliser une séquence consensus dans cette situation est trivial puisque l'ensemble des séquences alignées portent la même information. On peut aussi envisager le cas de deux souches très proches, qui ne diffèrent que de quelques nucléotides et qui ne sont représentées que par une seule séquence dans la base de référence. Dans ce cas, nous

1. http://samtools.sourceforge.net/pileup.shtml

résolvons les ambiguïtés en choisissant les polymorphismes majoritaires, ou bien un polymorphisme arbitraire en cas de couverture identique.

Dans tous les cas, les séquences des références sur lesquelles les contigs sont alignés ne sont pas considérées. Comme dans le reste de la méthode, les références ne servent ici que de pivots qui permettent de partitionner et positionner les contigs. Un biais résulte toutefois de la complétion de la base de données : lorsqu'une seule référence représente la diversité d'un clade, les différentes séquences de ce clade présentes dans les données produiront un scaffold unique. L'information ainsi perdue reste toutefois disponible pour l'utilisateur dans l'ensemble des contigs ayant servi à produire le scaffold.

Séquences finales Finalement, les scaffolds redondants sont supprimés et les grands scaffolds supérieurs à 500 bp sont sélectionnés.

3.2.5 Analyse taxonomique de l'échantillon

La dernière étape de MATAM a pour but d'estimer l'abondance de chaque scaffold, ainsi que de proposer une analyse taxonomique de l'échantillon dans le cas de l'ARNr SSU.

3.2.5.1 Estimation des abondances

Pour estimer les abondances des scaffolds, nous reprenons l'intégralité des lectures initiales obtenues après l'étape de filtrage (3.2.1.2), pour les aligner sur les scaffolds. Cela est réalisé avec SortMeRNA, paramétré pour identifier jusqu'à 10 bons alignements par lecture.

```
$ sortmerna --ref scaffolds.fa,scaffolds --reads lectures_marqueur.fq --aligned
    alignements_lectures --sam --best 10 --min_lis 10 -e 1e-5
```

Les alignements sont ensuite filtrés pour ne garder que les meilleurs alignements équivalents pour chaque lecture (score alignement $\geq 99\%$ meilleur score). L'abondance A_S d'un scaffold est estimée par la somme des poids des lectures alignées sur ce scaffold. Le poids p_i d'une lecture i est défini comme l'inverse du nombre de meilleurs alignements équivalents n_i de cette lecture sur les scaffolds

$$A_S = \sum p_i,\ p_i = 1/n_i$$

Un des fichiers principaux de sortie de MATAM est donc le fichier des scaffolds, annotés par leurs abondances estimées.

3.2.5.2 Assignation taxonomique des scaffolds

Dans le cadre de l'analyse du marqueur ARNr SSU, nous assignons taxonomiquement les séquences reconstruites en utilisant le classifier RDP [98], le modèle d'en-

traînement « 16srrna » et un seuil de confiance supérieur à 90%. Cet outil fournit pour chaque scaffold une assignation dont la précision va au maximum jusqu'au niveau du genre.

3.2.5.3 Représentation de la composition taxonomique avec Krona

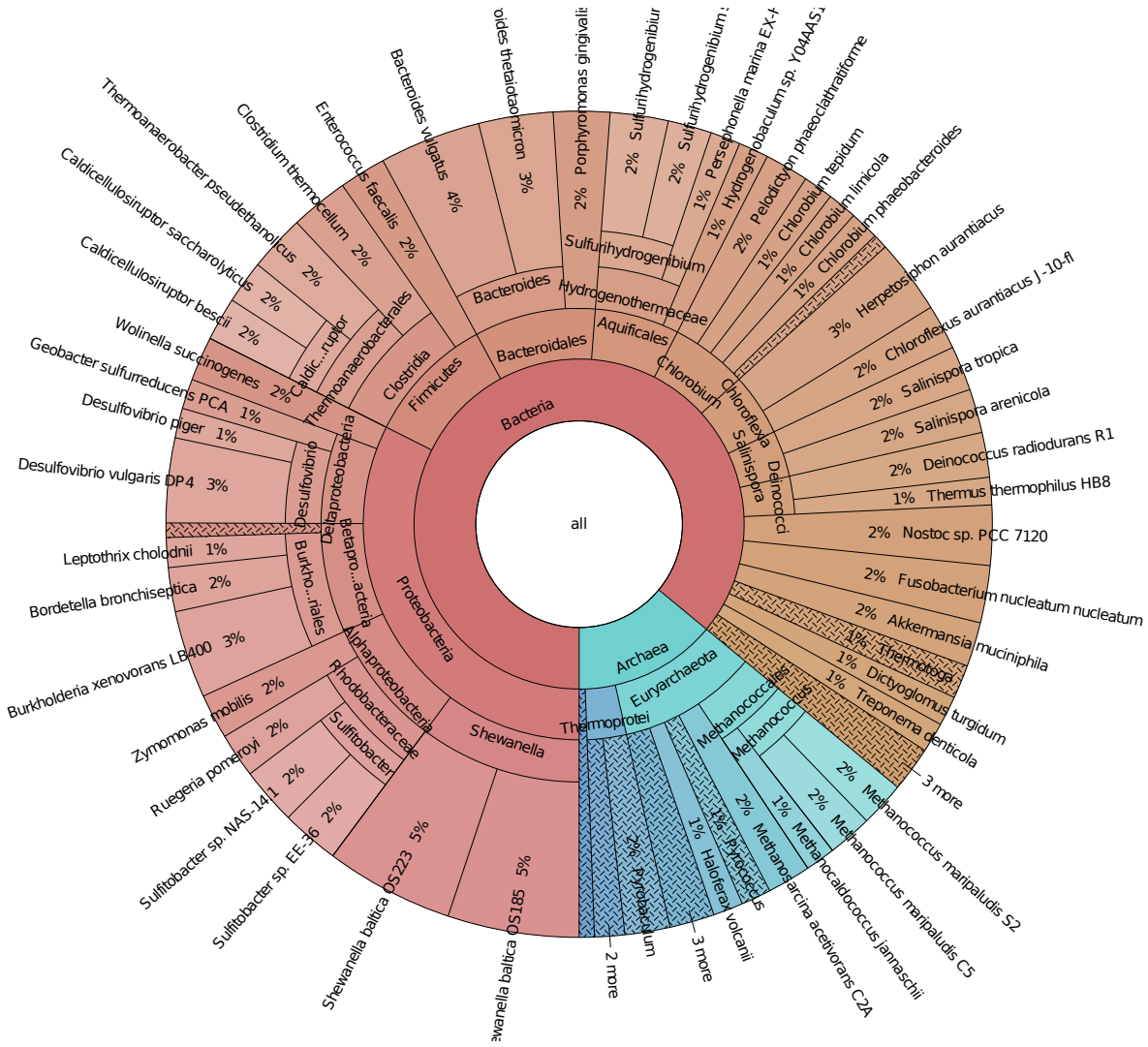


FIGURE 3.8 – Exemple d'une représentation Krona pour la communauté synthétique décrite au chapitre IV (4.3). Cette figure se lit de l'intérieur (qui correspond à la racine de l'arbre du vivant) vers l'extérieur (qui correspond aux taxons les plus spécifiques), et l'angle de chaque secteur est proportionnel à l'abondance des organismes qui correspondent à ce taxon.

Les informations d'abondance et d'assignation taxonomique des scaffolds sont synthétisées sous la forme d'une représentation Krona [68] (Figure 3.8), une visualisation interactive au format HTML qui permet de lire facilement la composition

taxonomique de l'échantillon.

3.3 Illustration de la méthode sur un jeu de 16 ARNr SSU bactériens

Dans cette section, nous reprenons pas à pas les différentes étapes de MATAM sur un exemple test, spécifiquement construit pour permettre l'évaluation de l'outil et de son paramétrage.

3.3.1 Génération du jeu de données

3.3.1.1 Propriétés du jeu de données

Afin de pouvoir être utilisé dans un cadre de validation de la méthode, le jeu de données a été conçu de manière à satisfaire les caractéristiques suivantes :

— être rapide à assembler. L'exécution de notre programme sur ce jeu de données doit se terminer en quelques minutes. Cela permet d'évaluer facilement et rapidement les nombreux paramètres de chacune de nos étapes ;

— contenir à la fois des espèces isolées et des groupes d'espèces proches, voire très proches (ARNr SSU > 99% d'identité). Idéalement, on souhaite trouver une gamme d'identité la plus continue possible afin de tester l'impact de la distance entre les séquences sur l'assemblage final ;

— permettre de retracer facilement l'origine des lectures. Chaque lecture correspond à une espèce et une position sur une séquence de référence, et cette information est facilement accessible.

— être contrôlé. Tous les paramètres de séquençage (distribution de longueur des lectures, taux d'erreur, profondeur de séquençage) sont connus et modifiables.

3.3.1.2 Sélection des séquences initiales

TABLE 3.1 – Statistiques descriptives pour les séquences représentatives des 16 espèces

Nb séquences	16
Longueur min	1 393 bp
Longueur max	1 555 bp
Longueur moy	1 504,12 bp
Nb total nt	24 066 bp

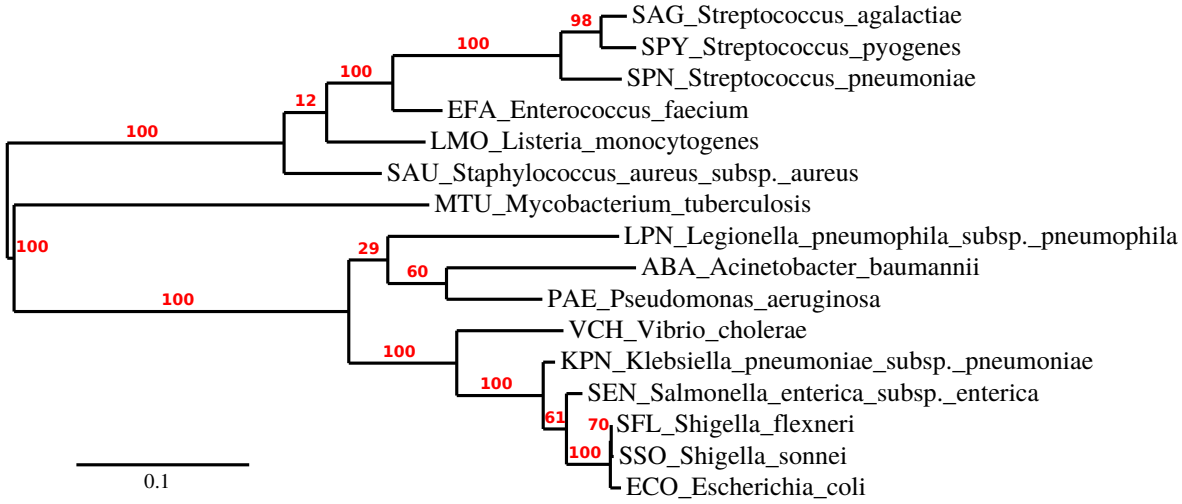


FIGURE 3.9 – Arbre phylogénétique des séquences d'ARN 16S des 16 espèces sélectionnées. Les feuilles de l'arbre correspondent aux espèces, qui sont représentées par un code à 3 lettres suivi de leur nom binominal. Les supports des branches sont indiqués en rouge et en pourcentage.

Nous avons sélectionné 16 espèces bactériennes bien étudiées sur la base de leur positionnement dans l'arbre du vivant. On retrouve :

— neuf espèces de l'embranchement *Gammaproteobacteria*, dont trois espèces très proches du genre *Escherichia-Shigella* ;

— six espèces de l'embranchement *Bacilli*, dont trois espèces proches du genre *Streptococcus* ;

— et une espèce isolée de l'embranchement *Actinobacteria*

Pour chacune de ces 16 espèces, une séquence d'ARN 16S de bonne qualité a été sélectionnée dans la base SILVA LTPs 119 SSU [2], validée manuellement [102]. Les statistiques descriptives pour ces séquences sont décrites Table 3.1. Les 16 séquences ont été alignées sur le site `www.phylogeny.fr` (mode « One Click » : alignement par MUSCLE, pas de curation par Gblocks et phylogénie par PhyML). L'arbre phylogénétique résultant est affiché Figure 3.9, et la matrice des similarités entre paires de séquences Table 3.2.

2. https://www.arb-silva.de/projects/living-tree/

TABLE 3.2 – Matrice d'identité entre les séquences représentatives des 16 espèces

LPN	100															
VCH	85,3	100														
ECO	83,1	90,1	100													
SFL	83,5	90,6	99,4	100												
SSO	83,6	90,4	99,2	99,9	100											
KPN	83,6	91,1	95,7	96,2	96,2	100										
SEN	83,6	90,8	96,5	97,0	96,9	97,5	100									
ABA	85,4	83,3	84,3	84,4	84,3	84,8	85,0	100								
PAE	86,4	85,5	85,0	85,4	85,5	86,6	85,9	88,3	100							
MTU	78,4	77,3	77,5	77,4	77,5	78,1	77,8	76,3	77,7	100						
SPN	76,3	76,6	77,6	77,7	77,5	77,6	78,0	76,9	77,8	78,2	100					
SAG	75,2	75,6	76,2	76,5	76,3	76,5	77,3	76,9	76,9	77,8	94,1	100				
SPY	77,3	78,1	77,6	77,9	77,9	78,4	79,3	78,4	78,5	78,5	95,0	97,2	100			
SAU	77,0	77,4	77,1	77,2	77,3	77,8	77,7	76,1	78,4	79,0	84,5	84,9	85,9	100		
LMO	76,0	76,5	77,1	77,3	77,4	77,1	77,7	76,4	78,9	78,1	86,3	85,9	86,7	89,7	100	
EFA	76,3	76,8	77,3	77,7	77,7	78,0	78,4	76,3	78,1	79,8	89,1	89,1	89,4	89,5	90,6	100

3.3.1.3 Simulation des lectures

Choix du simulateur Il existe un grand nombre de simulateurs de lectures de séquençage, dont notamment Grinder [2], ART [31], Mason [30] et MetaSim [77]. Nous avons choisi ART pour sa facilité d'utilisation, et la présence de modèles d'erreur précalculés estimés sur des données réelles. ART ne permet pas de simuler une distribution d'abondance pour une communauté métagénomique. Dans notre cas, nous avons simulé un séquençage sur un ensemble de séquences de référence avec la même couverture de séquençage et ART se prête bien à cet exercice

Paramètres de simulation Nous avons simulé un jeu de données de séquençage de 11650 lectures avec ART et les paramètres suivants :

- modèle d'erreur : Illumina HiSeq2500 ;
- taille des lectures : 100 bp ;
- séquençage apparié ;
- taille du fragment : 150 bp, écart-type 30 ;
- profondeur de séquençage : 50x ;

3.3.2 Illustration des étapes de MATAM

3.3.2.1 Alignement des lectures

Comme décrit à la section 3.2.1, les 11 650 lectures ont été alignées sur la base de référence partitionnée NR95 avec SortMeRNA, et un maximum de 10 alignements par

TABLE 3.3 – Statistiques alignements

Nb lectures	Nb lectures alignées	Nb alignements	Nb moyen d'alignements/lecture
11 650	11 650	116 144	9,97

lecture ont été identifiés (Table 3.3). En pratique, la très grande majorité des lectures (11 572) s'alignent sur 10 références différentes et seulement 10 lectures s'alignent sur 3 références ou moins.

3.3.2.2 Sélection des alignements informatifs

À la suite de l'étape de sélection des alignements basée sur la distribution des scores de ces alignements (3.2.1.3), le nombre de lectures qui s'alignent sur 10 références tombe à 10 535 tandis que le nombre de lectures spécifiques qui s'alignent sur 3 références ou moins est de deux ordres de grandeur supérieurs (827 lectures après sélection des alignements) (Figure 3.10).

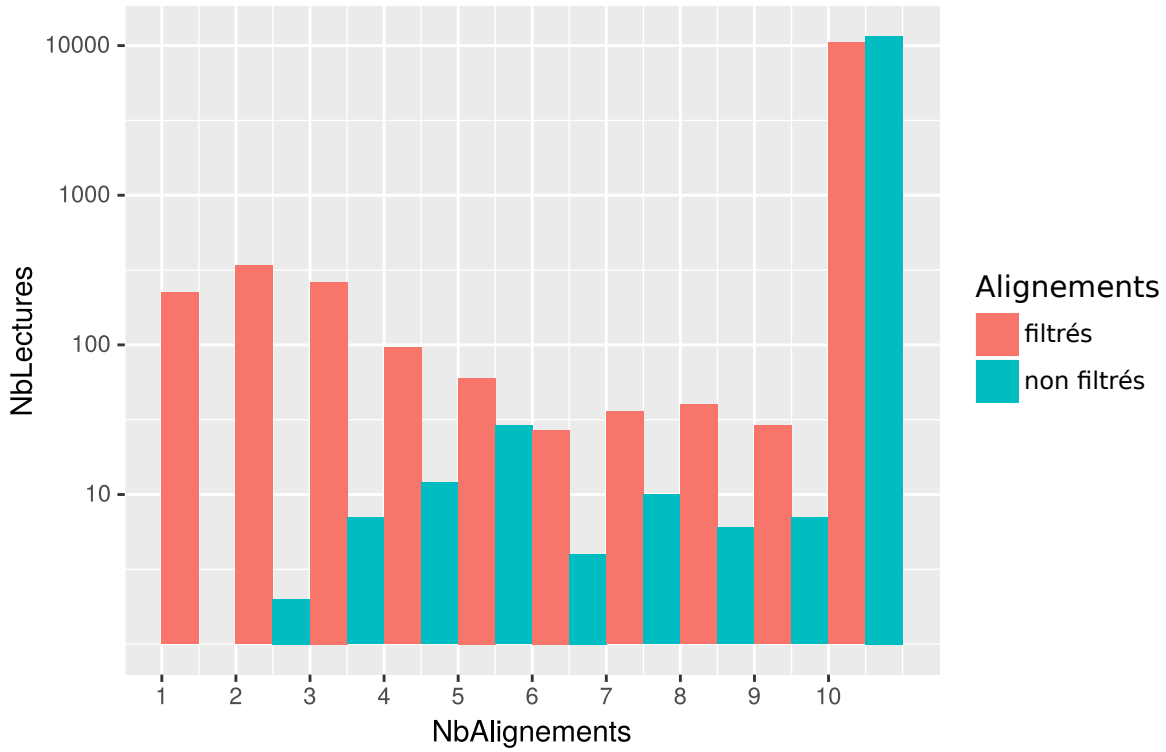


FIGURE 3.10 – Effet de la sélection sur le nombre d'alignements par lecture.

TABLE 3.4 – Statistiques descriptives des graphes de chevauchement et graphes compressés pour deux paramétrages différents

		Paramétrage permissif	Paramétrage strict
Graphe de chevauchement	Nb lectures	11 650	11 650
	- dont singletons	3 (0,03%)	212 (1,8%)
	Nb chevauchements	612 243	336 976
	- dont interspécifiques (faux positifs)	251 448 (41,1%)	120 075 (35,6%)
Graphe compressé	Nb nœuds compressés	203	311
	Nb arêtes	222	332
	Nb lectures représentées	11 639	11 323
	- dans des nœuds mono spécifiques	5 227 (44,9%)	5 917 (52,3%)
	Nb composantes	76	100
	- dont singletons	2	4
	- dont hubs	32	39
	- dont chemins spécifiques	42	57

3.3.2.3 Construction du graphe de chevauchement

Nous avons évalué l'influence des paramètres de l'étape de construction du graphe de chevauchement sur les résultats finaux, et en particulier la taille minimale m et le seuil minimum d'identité i des chevauchements. Nous présentons ici deux exemples caractéristiques de paramétrages pour le jeu de données de 16 espèces : un paramétrage permissif avec $m = 30$ bp et $i = 97\%$ d'identité, et un paramétrage plus strict avec $m = 50$ bp et $i = 100\%$ d'identité.

La visualisation des graphes de chevauchements respectifs pour les deux paramétrages (Figures 3.11a et 3.11b) permet d'observer que pour le paramétrage permissif, les 16 espèces semblent difficiles à distinguer les unes des autres et peu de sous-graphes spécifiques semblent se détacher. *A contrario*, avec le paramétrage strict, les différentes espèces se distinguent plus facilement au sein du graphe de chevauchement. On observe même des sous-graphes monospécifiques, dont toutes les lectures proviennent d'une même espèce. Il semble aussi que le graphe ne soit pas trop fragmenté malgré le paramétrage particulièrement strict qui entraîne une diminution de près de moitié du nombre de chevauchements identifiés (Table 3.4). De plus, le taux de chevauchements interspécifiques (qui correspondent à des faux positifs) diminue aussi avec le

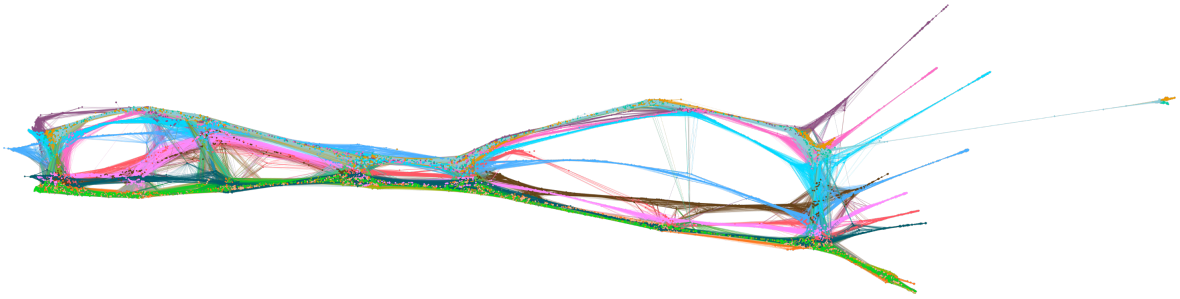
paramétrage strict.

3.3.2.4 Compression du graphe de chevauchement

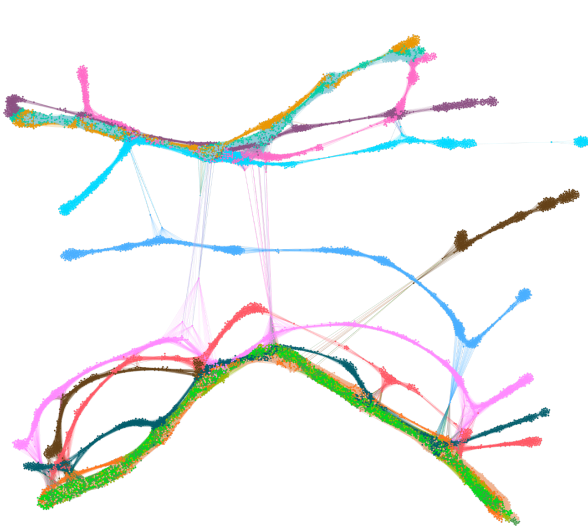
La compression des deux graphes de chevauchement construits avec les deux jeux de paramètres permet de diminuer de plusieurs ordres de grandeur la taille de ces graphes (Table 3.4). Nous avons notamment évalué la qualité des graphes compressés en identifiant les nœuds mono-spécifiques du graphe compressé, c'est-à-dire les nœuds qui ne représentent que des lectures de la même espèce. Par opposition, les nœuds compressés qui contiennent des lectures d'espèces différentes sont considérés comme hybrides.

On observe sur la Figure 3.11c que la majorité des nœuds du graphe compressé sont des nœuds mono-spécifiques ce qui semble valider notre approche. On peut aussi observer que, de manière attendue, on retrouve la plupart des nœuds hybrides (en noir) au sein des groupes d'espèces proches, comme celui des *Enterobacteriales*, ou celui des *Lactobaciliales*.

De plus, le nombre de lectures dans des nœuds mono-spécifiques augmente clairement dans le graphe compressé issu du paramétrage strict (Table 3.4), ce qui semble confirmer que ce paramétrage est le plus adapté pour ce jeu de données.



(a) Graphe de chevauchement pour un paramétrage permissif ($m = 30$ bp et $i = 97\%$ d'identité)



(b) Graphe de chevauchement pour un paramétrage strict ($m = 50$ bp et $i = 100\%$ d'identité)

(c) Graphe compressé pour un paramétrage strict ($m = 50$ bp et $i = 100\%$ d'identité)

FIGURE 3.11 – Graphes de chevauchement et graphe compressé obtenus avec les lectures des 16 espèces et deux paramétrages différents. Dans les graphes de chevauchement, les lectures de chaque espèce sont représentées de la même couleur. Pour le graphe compressé, sont colorés les nœuds ne contenant que des lectures de la même espèce et sont représentés en noir les nœuds possédant au moins deux lectures issues d'espèces différentes.

3.3.2.5 Assemblage des composantes

TABLE 3.5 – Comparaison des contigs et scaffolds pour deux paramétrages différents

		Paramétrage permissif	Paramétrage strict
Contigs	Nb contigs	326	308
	- min	100 bp	100 bp
	- max	1 092 bp	1 092 bp
	- moy	156,0 bp	159,3 bp
	FR	98,7%	98,5%
	ER	0%	0,013%
Scaffolds	Nb scaffolds	23	22
	- min	665 bp	537 bp
	- max	1 527 bp	1 526 bp
	- moy	1 284,3 bp	1 241,5 bp
	FR	90,9%	92,3%
	ER	0,078%	0,050%

Suite à la compression des graphes de chevauchements, et au partitionnement des graphes compressés en trois types de composantes, les lectures de chaque composante sont assemblées avec SGA. Par définition, le nombre de contigs résultants ne peut être inférieur au nombre de composantes du graphe compressé. En pratique, les assemblages obtenus pour les paramétrages permissif et strict sont encore relativement fragmentés (Table 3.5). Dans les deux cas, plus de 300 contigs ont été reconstruits (pour une cible de 16 séquences) et la taille moyenne des contigs est de l'ordre de 160 bp, pour des lectures initiales de longueur 100bp. La fragmentation de l'assemblage n'est cependant pas homogène. Pour l'espèce la plus isolée du jeu de données (*Mycobacterim tuberculosis*), un contig correspondant quasiment à la séquence pleine taille de l'ARNr SSU a été reconstruit. Par contre, les groupes d'espèces proches appartenant au même genre sont principalement représentés par des petits contigs de quelques centaines de nucléotides.

Malgré la fragmentation, ces assemblages semblent avoir retenu la quasi-totalité de l'information initiale. En effet, les 16 séquences originales sont couvertes à plus de 98%, et avec un taux d'erreur très bas (de l'ordre de $0,01\%$). On remarque aussi que le paramétrage strict semble fournir un assemblage un peu moins fragmenté que le paramétrage permissif.

3.3.2.6 Scaffolding

La dernière étape de scaffolding vise à reconstruire des séquences pleines tailles à partir des contigs assemblés par SGA. Pour ce jeu de données, le résultat idéal consisterait à obtenir 16 scaffolds (un pour chaque espèce) de longueur moyenne de 1 500 bp et avec un taux d'erreur le plus faible possible.

En pratique, nous n'en sommes pas si loin (Table 3.5). MATAM reconstruit un peu plus de 20 contigs de longueur moyenne d'environ 1 250 bp, représentant plus de 90% des séquences d'origine, et avec un taux d'erreur inférieur à 0,1%, soit moins d'une erreur tous les 1 000 nucléotides. En y regardant de plus près, on s'aperçoit que la majorité des espèces possède un représentant pleine taille, mais que quelques espèces appartenant à un groupe d'espèces proches peuvent être représentées par deux ou trois plus petits contigs.

Là encore, les différences de résultats entre les paramétrages sont modérées, mais on observe toutefois que le paramétrage strict permet d'obtenir un peu moins de contigs, tout en couvrant plus de séquences d'origine, et avec un taux d'erreur plus bas que pour le paramétrage permissif. Ces résultats semblent confirmer que le paramétrage strict est donc le plus adapté pour ce jeu de données, et par extension, pour les jeux de données de séquençage Illumina avec des caractéristiques similaires.

3.3.2.7 Estimation de la composition taxonomique

La dernière étape de MATAM consiste à estimer l'abondance de chaque scaffold ainsi qu'à les assigner taxonomiquement grâce au classifieur RDP. Dans le cas du jeu de données 16 espèces, puisque les lectures ont été simulées avec une abondance homogène pour toutes les espèces, nous nous attendons à ce que l'abondance relative de chaque espèce soit égale à 6,25% de l'abondance totale.

En pratique, la composition taxonomique estimée par MATAM est très proche de celle attendue (Figure 3.12). Les 12 genres du jeu de données sont bien représentés et l'abondance relative de chaque espèce est comprise entre 5 et 7%. Pour ce jeu de données de test, il semble donc bien que MATAM a été capable d'évaluer correctement la composition taxonomique du jeu de données à partir des séquences reconstruites.

3.4 Implémentation et disponibilité

L'ensemble du pipeline MATAM est implémenté en Python 3, et l'algorithme de construction et compression du graphe de chevauchement est écrit en C++11 en utilisant la librairie SeqAn [20].

MATAM est versionné dans un dépôt git et la plupart des dépendances (Sumaclust, SeqAn, SGA, RDP, Krona) sont incluses sous la forme de sous-modules git. MATAM est distribué sous licence GNU Affero GPL v3.0 et le code source est dispo-

FIGURE 3.12 – Représentation Krona de l'assemblage avec MATAM du jeu 16 espèces.

nible librement à cette adresse : https://github.com/bonsai-team/matam

Chapitre 4

MATAM, validation sur données expérimentales

Dans ce chapitre, nous présentons une évaluation approfondie des résultats de MATAM sur des données de séquençage : un ensemble de cinq jeux de données simulés avec une variation de la profondeur de séquençage, une communauté synthétique, et deux jeux de données réels issus du projet *Human Microbiote Project*. Pour chacun de ces cas d'étude, nous nous concentrons sur le marqueur ARNr 16S, et faisons une analyse comparative avec deux assembleurs métagénomiques génériques, SPAdes [4, 67] et MEGAHIT [43] (2.3.2), ainsi qu'avec deux outils dédiés à l'assemblage d'ARNr 16S, EMIRGE [61] (2.3.4.3) et REAGO [104] (2.3.4.4).

4.1 Protocole d'évaluation

Nous présentons dans cette partie le protocole d'évaluation des résultats que nous avons mis en place, avec le paramétrage de chaque outil, les étapes de nettoyage appliquées à chaque assemblage, et les métriques utilisées pour comparer les assemblages.

4.1.1 Paramétrage des logiciels

Notre approche vise à reconstruire des séquences d'un marqueur conservé avec peu d'erreurs tout en étant capable de distinguer entre deux séquences proches. Afin d'évaluer tous les outils sur un pied d'égalité, nous avons donc fait des choix de paramètres qui visent à maximiser leur sensibilité. Le paramétrage tient également compte de la nature des lectures de séquençage en matière de longueur et de taux d'erreur : il s'agit de lectures *Illumina* appariées.

4.1.1.1 MATAM

Nous avons utilisé MATAM v0.9.9. Les paramètres sont décrits en détail dans la section 3.2. Ils sont au nombre de quatre :

— le nombre maximum d'alignements de chaque lecture contre la base partitionnée : 10 ;

— la taille minimum du chevauchement entre deux lectures : $m = 50$ bp ;

— l'identité minimale entre les séquences des deux lectures sur leur chevauchement : $i = 100\%$ d'identité ;

— les filtres de la compression du graphe de chevauchement : supprimer les nœuds et les arêtes du graphe compressé de poids inférieur ou égal à 1.

4.1.1.2 EMIRGE

Par défaut, EMIRGE v0.61.1 est configuré pour assembler des séquences d'ARNr 16S avec un seuil d'identité de 97%. Cela correspond au seuil arbitraire classique de distinction des espèces bactériennes (1.1.6.1). En pratique, si deux espèces proches sont présentes dans l'échantillon, et que les séquences de leurs ARNr 16S divergent de moins de 3% de différence, EMIRGE ne reconstruira qu'une seule séquence représentative. Comme nous souhaitons évaluer la capacité des outils à reconstruire les séquences d'ARNr 16S avec le moins d'erreurs possible et avec la résolution la plus précise possible, nous avons paramétré EMIRGE afin de maximiser sa sensibilité à de faibles taux de différence. De plus, en suivant les recommandations des auteurs d'EMIRGE, nous avons augmenté le nombre d'itérations de manière à améliorer la précision de l'assemblage. Enfin, puisque tous nos jeux de données contiennent des lectures appariées, nous avons utilisé EMIRGE en mode apparié.

Les paramètres utilisés sont ainsi les suivants :

— nombre d'itérations : 100, au lieu de 40 par défaut ;

— seuil de fusion : 100% d'identité, au lieu de 97% par défaut ;

— seuil de fraction de SNP : 0,001 (0,1%), au lieu de 0,04 (4%) par défaut ;

— fraction minimale pour identifier un variant à une position donnée : nous avons gardé la valeur par défaut de 0,1 ;

— profondeur moyenne minimale : nous avons gardé la valeur par défaut de 3. Une diminution de ce paramètre mène à des sur assemblages.

4.1.1.3 REAGO

Nous avons utilisé REAGO v1.1, avec les paramètres par défaut. Il faut noter que REAGO n'accepte que des lectures appariées et de tailles identiques.

4.1.1.4 Sélection des lectures d'ARNr 16S avec SortMeRNA

Afin de pouvoir les assembler avec SPAdes et MEGAHIT, les lectures d'ARNr 16S de chaque jeu de données ont été initialement identifiées et filtrées avec SortMeRNA v2.1 et la base de séquences de référence partitionnée (SILVA_128_SSURef_NR95). Cette étape n'est pas nécessaire pour EMIRGE et REAGO qui intègrent leur propre filtre.

Nous avons utilisé les paramètres par défaut de SortMeRNA, :

— taille des graines : 18 bp +/- une erreur ;

— nombre minimum de graines : 2 ;

— e-valeur (calculée sur l'ensemble des lectures) : 1.

4.1.1.5 SPAdes

Nous avons utilisé SPAdes v3.9.0, paramétré pour être particulièrement sensible aux polymorphismes, ce qui correspond au mode *careful* par défaut.

Les auteurs de SPAdes ont récemment implémenté un mode d'assemblage métagénomique, sous l'appellation metaSPAdes [67]. Cette version est optimisée pour l'assemblage des génomes d'un échantillon métagénomique, et a récemment fait ses preuves comme l'un des meilleurs assembleurs métagénomiques au concours d'assemblage CAMI [84] (2.3.2.2). Nous avons donc tenté d'assembler les ARNr 16S avec metaSPAdes, mais les assemblages obtenus étaient moins bons qu'en utilisant SPAdes en mode *careful*. Nous ne présentons donc pas ces résultats ici.

4.1.1.6 MEGAHIT

Nous avons utilisé MEGAHIT v1.1.1 et son préréglage *meta-large*, optimisé pour les métagénomes complexes. Des trois préréglages disponibles (défaut, *meta-large* et *meta-sensitive*), il s'agit de celui qui nous a permis d'obtenir les meilleurs résultats pour l'assemblage d'ARNr 16S.

4.1.2 Post-traitement des assemblages

Attention ! À partir d'ici, nous appellerons *contigs* les séquences finales de chaque assemblage, et *séquences de référence* les séquences connues des ARNr 16S pour les espèces présentes dans l'échantillon initial.

Afin de pouvoir évaluer les différents outils sur un pied d'égalité, nous avons appliqué un post-traitement uniforme des résultats, conforme à l'usage de la communauté. L'objectif est de ne garder que les séquences reconstruites satisfaisant les mêmes critères.

Suppression des petits contigs Pour chaque assemblage, nous n'avons gardé que les contigs de longueur supérieure à 500 bp. Ce seuil est assez laxiste, puisque REAGO préconise de ne conserver que les contigs de longueur supérieure à 1 350 nucléotides dans le cas de l'ARNr 16S.

Suppression des chimères Les séquences reconstruites ont ensuite été filtrées pour éliminer de potentielles chimères, en utilisant l'algorithme UCHIME [24] implémenté dans VSEARCH [78]. Les auteurs de l'algorithme UCHIME recommandent d'utiliser la base de données de référence la plus grande possible [1]. Nous avons donc utilisé la base SILVA 128 SSURef NR99 complète.

4.1.3 Comparaison des assemblages finaux

Pour les jeux de données dont nous connaissions les séquences de référence initialement présentes dans les échantillons (jeux simulés en section 4.2 ou communauté synthétique en section 4.3), nous avons évalué les assemblages avec MetaQUAST [60]. MetaQUAST aligne les contigs de chaque assemblage sur les séquences de référence. Pour chaque contig, nous avons gardé les meilleurs alignements strictement équivalents (de même score) et supérieurs à 97% d'identité.

Pour chaque assemblage, nous avons pris en compte les métriques suivantes :

— *Nb contigs* : le nombre de contigs ;

— *LT* : la longueur totale, qui correspond au nombre total de nucléotides dans l'ensemble des contigs ;

— *LTA* : la longueur totale alignée, qui correspond au nombre de nucléotides des contigs alignés sur les séquences de référence ;

— *FR* : la fraction reconstruite, qui correspond au nombre de nucléotides des séquences de référence couverts par des contigs, divisé par la taille totale des séquences de référence ;

— *TE* : le taux d'erreur, qui correspond au pourcentage de substitutions et d'*indels* des alignements des contigs avec les séquences de référence les plus proches ;

— *Ns* : le nombre de nucléotides non résolus dans les contigs.

Toutes ces métriques s'appliquent aux contigs non chimériques de longueur supérieure à 500 nucléotides.

1. http://drive5.com/usearch/manual/cmd_uchime2_ref.html

4.2 Jeux simulés avec variation de la profondeur de séquençage

Dans cette première partie, nous évaluons notre méthode sur des jeux de données simulés, pour lesquels nous contrôlons toutes les caractéristiques du séquençage. Nous avons plus particulièrement analysé l'impact de la profondeur de séquençage sur les résultats. En particulier, l'une des difficultés des analyses métagénomiques est la capacité à détecter des espèces peu abondantes et séquencées avec de faibles profondeurs (1.1.2).

4.2.1 Génération des jeux de données

Les jeux de données sont simulés à partir d'une sélection de génomes initialement publiés par Mavromatis et al. en 2007 [56] et complétée par Pignatelli et Moya en 2011 [73]. La sélection comprend les génomes de 122 espèces, qui correspondent à la diversité taxonomique d'une communauté réaliste. L'ensemble des génomes contient 287 copies distinctes (non redondantes) d'ARNr 16S.

Nous avons simulé cinq jeux de données de séquençage haut débit avec des profondeurs de séquençage variables : 50x, 20x, 10x, 5x, 2x par génome respectivement. Des lectures de type Illumina ont été générées avec le simulateur ART [32] et les paramètres suivants :

— profil d'erreur : HiSeq2500 ;

— librairie : appariée ;

— taille du fragment : 250 bp (30 bp écart-type) ;

— longueur des lectures : 101 bp ;

Dans ces simulations, toutes les espèces ont la même abondance, ce qui correspond à la communauté de *haute complexité* introduite par Mavromatis et al.

Les cinq jeux de donnée simulés, les 122 génomes initiaux et les séquences d'ARNr 16S de référence sont disponibles à cette adresse : `http://bioinfo.cristal.univ-lille.fr/matam/material.php`.

4.2.2 Résultats

Les jeux de données simulés complets ont été assemblés par MATAM, EMIRGE et REAGO. SPAdes et MEGAHIT ont été utilisés pour assembler les lectures d'ARNr 16S préalablement identifiées par SortMeRNA. La Table 4.1 présente les résultats de la comparaison des assemblages réalisés par les cinq outils sur les jeux de données simulés. Les résultats des cinq profondeurs de séquençage sont moyennés (moy) et accompagnés de leur écart-type (ET).

TABLE 4.1 – Résultats pour les jeux de données simulés avec variation de la profondeur de séquençage. Les métriques pour les cinq profondeurs de séquençage sont moyennées. TMC est la taille moyenne des contigs.

	Chimères (%)		LTA/LT (%)		TE (%)		Ns (%)		TMC	
	moy	ET	moy	ET	moy	ET	moy	ET	moy	ET
MATAM	1,28	0,55	99,3	0,2	0,03	0,02	0,00	0,00	1 252	116,9
EMIRGE	36,89	9,42	79,9	11,6	0,62	0,16	0,55	0,36	1 436	15,4
REAGO	42,11	10,36	91,5	0,8	0,31	0,13	0,00	0,00	1 333	298,9
SPAdes	21,23	9,05	73,5	15,9	0,60	0,49	0,02	0,04	966	47,4
MEGAHIT	23,81	2,85	80,3	4,9	0,36	0,18	0,00	0,00	962	87,6

Tout d'abord, on peut remarquer que plus de 99% des nucléotides des contigs de MATAM ont été alignés par MetaQUAST [60] sur une des 287 séquences d'ARNr 16S de l'échantillon initial (moy LTA/LT). Cette proportion ne dépasse pas les 91% avec REAGO, et les 80% pour les autres méthodes. De plus, les contigs de MATAM ont le taux d'erreur le plus bas (TE=0,03%), ce qui représente un gain de précision d'un facteur 10 par rapport aux autres assembleurs, et un gain d'un facteur 20 par rapport à EMIRGE. Les contigs d'EMIRGE contiennent par ailleurs 0,5% de bases inconnues (Ns), ce qui amène son taux d'erreur effectif au-dessus des 1%. On remarquera aussi que MATAM a reconstruit environ 30 fois moins de séquences chimériques qu'EMIRGE ou REAGO.

Sur la Figure 4.1, nous avons représenté la fraction reconstruite du marqueur par rapport à la profondeur de séquençage. Alors que MATAM a reconstruit entre 76% et 85% des séquences de références pour des profondeurs de séquençage supérieures à 10x, EMIRGE a reconstruit moins de 55% des séquences de références, et la fraction reconstruite pour les autres méthodes est inférieure à 22%. Pour de faibles profondeurs de séquençage, MATAM obtient aussi les meilleurs résultats, avec une fraction reconstruite de 33% pour une profondeur de séquençage de 2x, alors que celle des autres assembleurs varie entre 5 et 10%.

Finalement, bien que les assemblages de MATAM semblent un peu plus fragmentés que ceux d'EMIRGE ou REAGO (TMC inférieur pour MATAM, Table 4.1), ils représentent une amélioration importante en matière de résolution de reconstruction et de taux d'erreur. Les séquences reconstruites par MATAM seraient ainsi tout à fait adaptées pour obtenir une assignation taxonomique au niveau de l'espèce.

Nous avons aussi évalué les estimations de la composition taxonomique de l'échantillon par MATAM et EMIRGE sur les jeux de données simulés 10x, 20x, et 50x. Nous avons ainsi représenté la différence entre l'abondance estimée par chacune des méthodes et l'abondance théorique connue (Figure 4.2). Cette différence a été évaluée au moyen

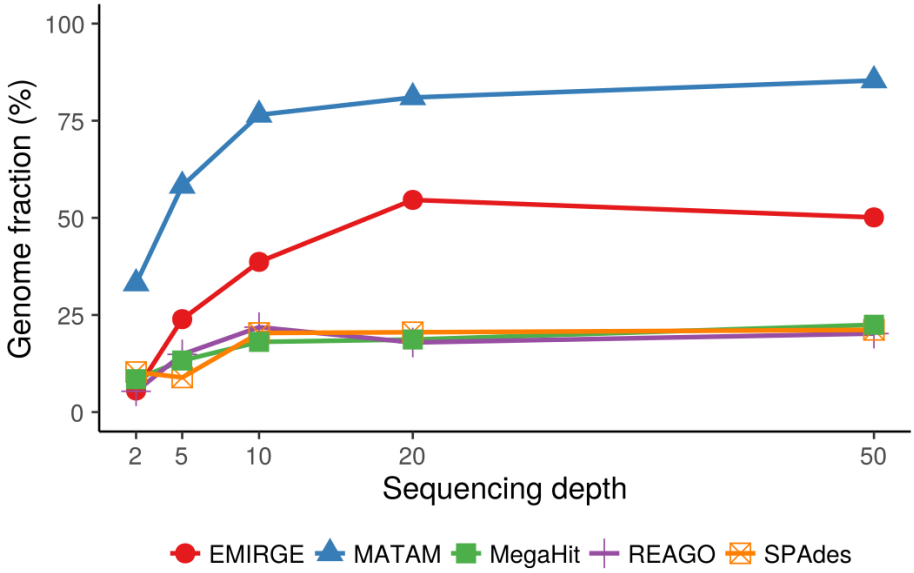


FIGURE 4.1 – Effet de la profondeur de séquençage (*sequencing depth*) sur la fraction reconstruite (*genome fraction*) des assemblages.

TABLE 4.2 – Statistiques de comparaison des abondances de genres estimées par MATAM et EMIRGE.

	MATAM			EMIRGE		
	10x	20x	50x	10x	20x	50x
Distance de Pearson avec la distribution théorique	0,968	0,966	0,967	0,920	0,944	0,918
Nb Faux Positifs	15	16	17	9	16	25
Nb Faux Négatifs	16	15	16	34	28	32

de trois métriques :

— la distance de Pearson entre la distribution observée et la distribution théorique. Cette distance est calculée pour l'ensemble des genres communs aux distributions de MATAM et EMIRGE et la distribution théorique ;

— le nombre de faux positifs, c'est-à-dire de genres identifiés comme présent par la méthode alors qu'ils ne sont pas présents dans la communauté initiale ;

— le nombre de faux négatifs, c'est-à-dire de genres présents dans la communauté initiale et non retrouvés par la méthode.

Nous observons donc que pour les jeux de données simulées, MATAM comme EMIRGE estiment correctement la composition taxonomique de la communauté (Table 4.2). Toutefois, MATAM est systématiquement meilleur dans son estimation, avec de plus

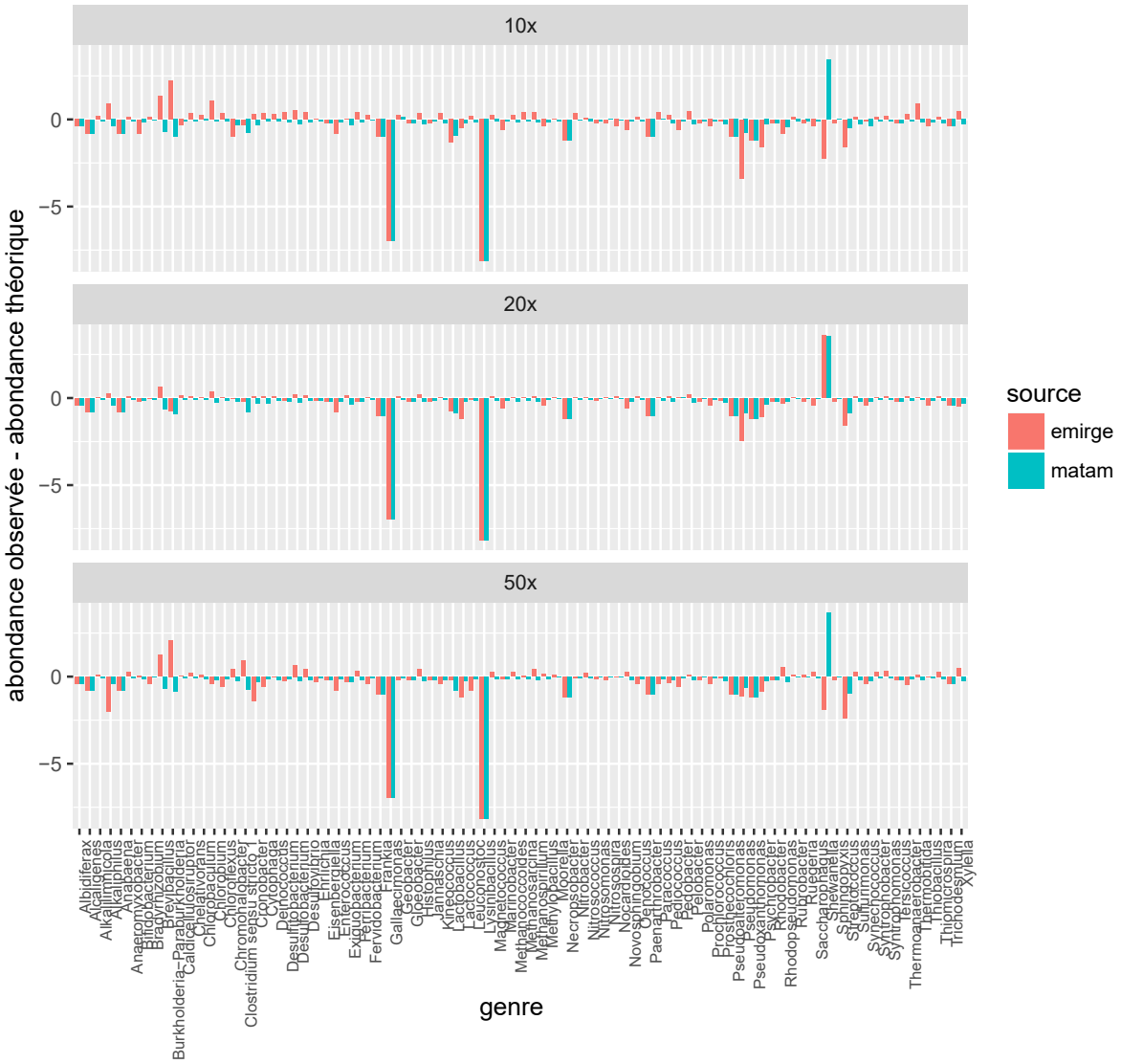


FIGURE 4.2 – Comparaison des abondances de genres estimées par MATAM et EMIRGE. Pour chacun des trois jeux de données simulés est représentée la différence entre l'abondance relative observée et l'abondance théorique pour chaque genre de la communauté initiale.

petites distances par rapport à la distribution théorique, ainsi qu'un taux de faux positifs et de faux négatifs inférieur à celui d'EMIRGE. L'observation plus précise de ces compositions taxonomiques estimées montre toutefois que la plupart des erreurs de MATAM comme d'EMIRGE correspondent à des séquences faussement identifiées comme appartenant à un genre proche d'un genre de la communauté théorique.

4.3 Communauté synthétique AB

Dans un deuxième temps, afin de nous rapprocher de conditions plus réalistes, nous avons utilisé un jeu de séquençage d'une communauté métagénomique synthétique. Un tel jeu de données est construit en séquençant un mélange d'espèces microbiennes en proportions connues, et dont on dispose d'une séquence génomique de référence, ou de la séquence de l'ARNr 16S par défaut. Dans notre cas, nous avons utilisé le jeu de données publié par Shakya et al. en 2013 [86]. C'est également le jeu de données avec lequel nous avions fait une courte comparaison d'EMIRGE et REAGO au chapitre II, section 2.3.4.5.

4.3.1 Description du jeu de données

4.3.1.1 Composition de la communauté synthétique

La communauté synthétique *Archaea-Bacteria* (AB) est composée de 16 espèces d'archées de 12 genres différents, ainsi que de 48 espèces de bactéries réparties parmi 36 genres (numéro d'accession SRS372410). Le génome de chaque espèce est connu (bien que non nécessairement publié), et comporte entre 1 et 10 opérons ribosomiques. Au total, cette communauté comprend 106 séquences distinctes d'ARNr 16S, avec des distances deux à deux qui varient de 59,64% à 99.93% d'identité.

Enfin, contrairement aux jeux de données simulés précédents (4.2), les cultures des différents organismes ont été mélangées avec des proportions variables, avec un rapport de plus de 200 entre l'espèce la plus abondante (8% de l'ADN total) et l'espèce la plus rare (0.05% de l'ADN total).

4.3.1.2 Caractéristiques du jeu de données de séquençage Illumina

Nous avons utilisé le jeu de données de séquençage complet Illumina associé à la publication, dont le numéro d'accession SRA est SRR606249 [2]. Les caractéristiques de ce jeu de données sont les suivantes :

— plateforme de séquençage : Illumina HiSeq2000 ;
— librairie : appariée ;
— taille moyenne de fragment : 250 bp (30 bp écart-type) ;
— taille des lectures : 101 bp ;
— nombre total de lectures : 109 millions.

Dans ce jeu de données Illumina, la profondeur de séquençage moyenne des différentes espèces varie entre 6x et 318x. La composition détaillée est disponible dans la Table S1 donnée en matériel supplémentaire de la publication [3].

2. https://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR606249
3. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3665634/bin/

4.3.1.3 Nettoyage du jeu de données

Puisque nous travaillons sur un jeu de données issu d'un séquençage réel, nous avons appliqué une étape préalable de nettoyage des lectures. Ce sont ces lectures nettoyées qui seront traitées par les différents logiciels, hormis REAGO qui ne peut pas prendre en charge des lectures de tailles variables et qui, par conséquent, n'a pu analyser que le jeu de données brut.

Nettoyage basé sur la qualité Nous avons tout d'abord nettoyé le jeu de données avec Prinseq Lite v0.20.4 [82] et les paramètres suivants :

- *trimming* des bases inconnues (Ns) en 5' et 3' puis suppression des lectures restantes contenant encore des « N » ;
- *trimming* en 5' et 3' sur la qualité (min Phred=20) ;
- suppression des lectures avec une qualité moyenne < 25 ;
- *trimming* des poly-A/T supérieurs à 5 bp ;
- suppression des lectures de faible complexité (entropie < 70) ;
- suppression des petites lectures < 50 bp.

Ce qui correspond à la ligne de commande bash suivante :

```
$ prinseq-lite.pl -fastq SRR606249.fastq -min_len 50 -min_qual_score 20 -
    min_qual_mean 25 -ns_max_n 0 -noniupac -lc_method entropy -lc_threshold 70 -
    trim_tail_left 5 -trim_tail_right 5 -trim_ns_left 1 -trim_ns_right 1 -
    trim_qual_left 20 -trim_qual_right 20
```

Suppression des adaptateurs Nous avons ensuite supprimé les séquences d'adaptateurs (*Illumina Universal Adapter*) avec Cutadapt [54], et supprimé les lectures de longueurs inférieures à 50 nucléotides.

```
$ cutadapt -b AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT -b
    AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT -m 50 -o SRR606249.
    prinseq_good.cutadapt.fastq SRR606249.prinseq_good.fastq
```

Identification des lectures appariées Finalement, nous avons utilisé un script *ad hoc* pour ne conserver que les lectures appariées et éliminer les lectures singletons résultant des nettoyages précédents.

Jeu de données nettoyé final Le jeu Illumina nettoyé comprend 67,6 millions de lectures appariées, dont 108 560 lectures d'ARNr 16S identifiées par SortMeRNA (4.1.1.4). Le jeu de données nettoyé final est disponible à cette adresse : http://bioinfo.cristal.univ-lille.fr/matam/material.php

NIHMS439897-supplement-Supp_Table_S1.xlsx

4.3.2 Résultats

Nous avons exécuté :

— REAGO sur le jeu brut non nettoyé ;

— MATAM et EMIRGE sur le jeu nettoyé ;

— SPAdes et MEGAHIT sur les lectures d'ARNr 16S extraites du jeu nettoyé avec SortMeRNA.

4.3.2.1 Analyse globale

TABLE 4.3 – Résultats pour la communauté synthétique.

	Chimères (%)	Nb contigs	LT	LTA	FR (%)	TE (%)	Ns (%)
MATAM	3.2	101	139220	130654	83.1	0.05	0
EMIRGE	17.4	82	117138	102856	50.7	0.17	1.12
REAGO	15.5	59	90269	81297	42.8	0.06	0
SPAdes	5.5	59	70229	59988	39.9	0.11	0.05
MegaHit	3.0	61	77251	68904	44.3	0.18	0

La Table 4.3 présente les résultats obtenus.

Tendance générale De manière similaire à ce qui a été observé sur les jeux de données simulés, MATAM est l'approche qui reconstruit le plus grand nombre de séquences (101 contigs obtenus pour 106 ARNr 16S attendus) tout en ayant la fraction reconstruite la plus importante (FR=83%). De plus, avec un taux d'erreur TE inférieur à toutes les autres méthodes, l'assemblage de MATAM apparaît particulièrement précis. Sur le plan des fractions reconstruites, la deuxième meilleure approche est EMIRGE. Mais celui-ci génère dans le même temps l'assemblage dont le taux d'erreur TE est le plus élevé, et qui présente la plus grande quantité de bases inconnues Ns et de séquences chimériques

Comparaison des profils taxonomiques Afin de confirmer la tendance observée à travers ces mesures, nous avons réalisé une classification taxonomique des contigs de MATAM et EMIRGE avec le classifier RDP [98], et comparé ces classifications avec la composition attendue de la communauté synthétique. Sur 48 genres attendus, MATAM en retrouve 47 alors qu'EMIRGE n'en reconstruit que 44.

Impact de la similarité de séquences En analysant plus précisément les alignements des contigs réalisés par MetaQUAST [60] sur les séquences de références, on

observe que toutes les méthodes reconstruisent correctement tous les gènes d'ARNr 16S qui partagent moins de 90% d'identité avec la séquence la plus proche dans la communauté synthétique. Par contre, les performances de la plupart des assembleurs diminuent fortement lorsqu'il s'agit d'assembler des séquences proches dans le jeu de données. C'est en particulier le cas pour les copies paralogues d'ARNr 16S qui, chez les espèces du jeu de données en possédant, partagent plus de 99% d'identité.

TABLE 4.4 – Utilisation des ressources.

	Temps	Charge CPU (16 cœurs)	Mémoire (RAM)
MATAM	2 h 33	160%	13,3 Go
EMIRGE	21 h 17	1 550%	3,3 Go
REAGO	53 min 36 s	1 267%	27,5 Go
SPAdes	1 min 03 s	192%	10,6 Go
MEGAHIT	34 s	168%	56 Mo

Utilisation des ressources La Table 4.4 donne une comparaison des performances des cinq méthodes sur le jeu de données. On constate que le temps de calcul et l'empreinte mémoire de MATAM sont comparables à ceux des assembleurs dédiés à la reconstruction de marqueurs conservés. Il faut aussi noter que le temps de calcul affiché pour les assembleurs généralistes (SPAdes et MEGAHIT) ne prend en compte que l'assemblage des lectures préalablement identifiées comme appartenant à l'ARNr 16S par SortMeRNA.

4.3.2.2 Analyse détaillée sur un sous-ensemble représentatif de séquences

Afin de mieux comprendre les différences observées entre les assemblages générés par les différentes méthodes, nous avons sélectionné un sous-ensemble de quatre espèces proches qui possèdent chacune d'une à trois copies paralogues d'ARNr 16S.

Construction du sous-ensemble Pour générer ce sous-ensemble représentatif, nous avons aligné les contigs reconstruits par MATAM, EMIRGE et REAGO, ainsi que les séquences de référence d'ARNr 16S de la communauté synthétique avec MUSCLE [21] et construit un arbre phylogénétique avec PhyML [28] et les paramètres par défaut appliqués sur le site web Phylogeny.fr [18]. Nous avons ensuite sélectionné toutes les séquences d'un clade représentatif de l'arbre complet, et regroupant quatre espèces distinctes.

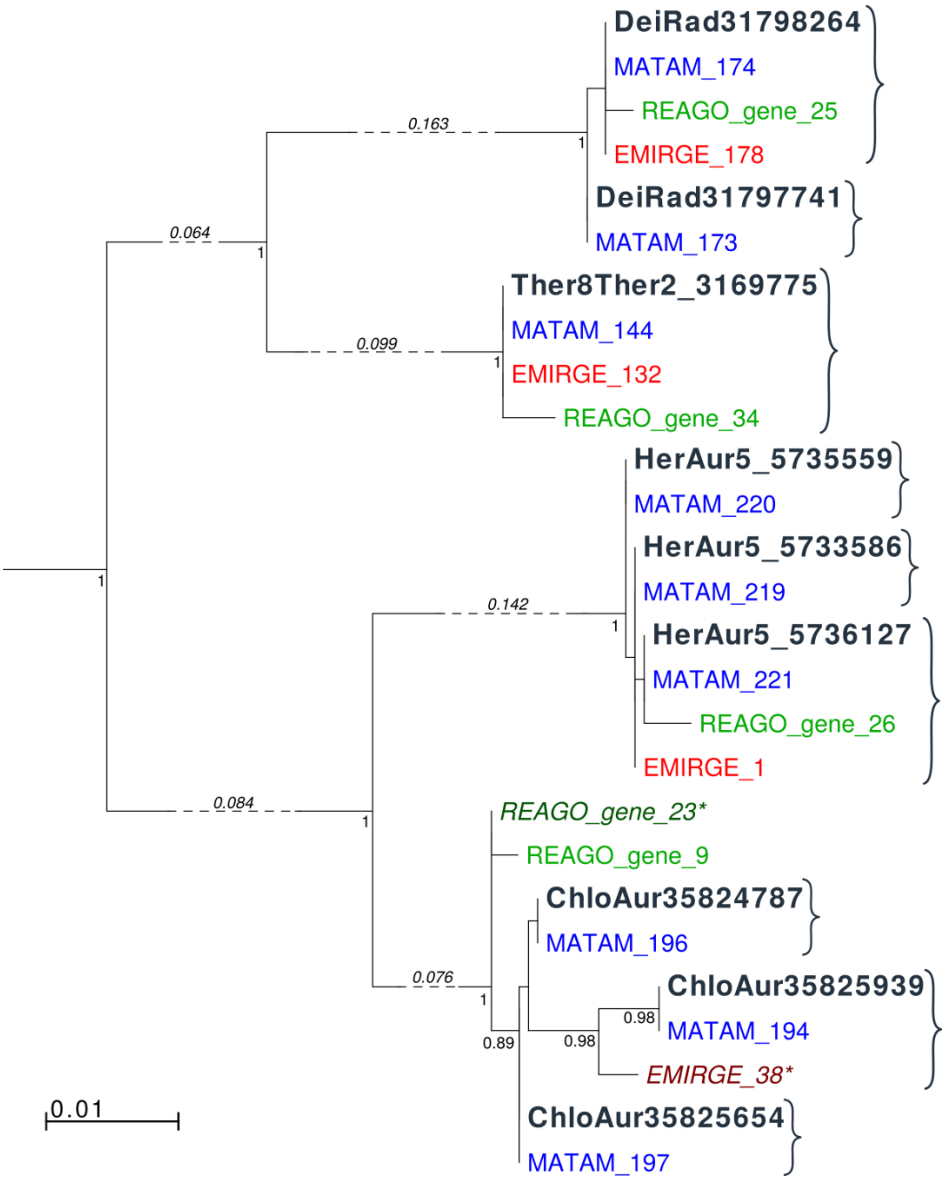


FIGURE 4.3 – L'alignement des séquences de référence avec les contigs reconstruits montre la capacité de MATAM à distinguer entre des séquences très similaires. Les contigs de MATAM, EMIRGE et REAGO sont représentés en bleu, rouge et vert, respectivement. Dans l'idéal, tous les assembleurs devraient produire un contig pour chaque séquence de référence (en noir). Les contigs suivis d'une étoile et représentés avec une couleur sombre sont considérés comme chimériques par VSEARCH. Dans l'arbre, la longueur d'un chemin séparant un contig donné de la séquence de référence la plus proche indique le taux d'erreur de la méthode ayant produit le contig. La barre d'échelle indique la longueur d'un chemin correspondant à 1% de sites divergents. Les branches les plus profondes ne sont pas représentées à l'échelle, et leurs longueurs sont indiquées numériquement en face de pointillés. Les supports aLRT des clades sont indiqués à la base des sous-arbres.

TABLE 4.5 – Matrice de similarité des séquences de référence pour le sous-ensemble représentatif.

DeiRad31797741	100,0								
DeiRad31798264	**99,87**	100,0							
Ther8Ther2_3169775	80,96	80,89	100,0						
ChloAur35824787	74,17	74,24	78,65	100,0					
ChloAur35825654	74,17	74,24	78,65	**99,86**	100,0				
ChloAur35825939	74,59	74,65	78,80	**98,85**	**98,85**	100,0			
HerAur5_5733586	74,40	74,47	75,41	82,02	82,02	81,75	100,0		
HerAur5_5735559	74,40	74,47	75,41	82,09	82,09	81,82	**99,93**	100,0	
HerAur5_5736127	74,47	74,53	75,35	81,95	81,95	81,68	**99,93**	**99,86**	100,0

Comparaison des assemblages pour le sous-ensemble représentatif de séquences La Figure 4.3 représente l'arbre phylogénétique obtenu pour le sous-ensemble représentatif sélectionné, et la Table 4.5 correspond à la matrice de similarité des séquences de référence d'ARNr 16S de ce sous-ensemble.

L'observation de l'arbre phylogénétique montre que MATAM a assemblé correctement tous les différents paralogues quasiment sans erreur (les longueurs des branches séparant un contig MATAM de la séquence de référence la plus proche sont pratiquement nulles). D'autre part, EMIRGE et REAGO n'ont reconstruit qu'une seule séquence candidate par espèce. Cela signifie qu'EMIRGE et REAGO ont assemblé en une seule séquence les lectures provenant de copies paralogues différentes, ce qui donne lieu à un assemblage erroné avec un fort taux d'erreur et une fraction reconstruite sous-estimée. De plus, on observe que toutes les séquences assemblées avec REAGO, ainsi qu'une séquence d'EMIRGE sur quatre, semblent se regrouper à une certaine distance du paralogue cible le plus proche. Ces distances correspondent aux erreurs de reconstruction des méthodes. De manière attendue, dans deux cas, les séquences candidates assemblées par EMIRGE et REAGO ont été identifiées comme chimériques par VSEARCH.

Ces observations sont représentatives de ce que l'on peut observer pour l'ensemble du jeu de données. Les contigs de MATAM sont systématiquement groupés plus près des séquences de référence que les contigs obtenus par les autres approches. Lorsque plusieurs paralogues sont présents, MATAM en distingue systématiquement davantage, avec un plus faible taux d'erreur. Ainsi, MATAM est ici capable de séparer des séquences qui diffèrent de moins de 0,1% d'identité, soit une unique substitution sur l'ARNr 16S complet, de taille environ 1 500 nucléotides. Nous obtenons donc une résolution suffisamment précise pour différencier des souches de la même espèce microbienne dès qu'un unique polymorphisme peut être observé.

4.4 Jeux métagénomiques réels : HMP

Dans ce dernier cas d'étude, nous travaillons sur deux jeux métagénomiques réels, séquencés à partir de véritables environnements. Pour ceux-ci, nous ne connaissons donc pas la composition microbienne exacte des échantillons, ce qui nous empêche d'appliquer les métriques basées sur MetaQUAST et utilisées précédemment pour l'évaluation des résultats. En revanche, nous disposons d'une analyse amplicon 16S, qui nous permet de mettre en place une stratégie d'évaluation alternative.

4.4.1 Description des jeux de données

Les jeux de données réels proviennent du *Human Microbiote Project* [14] (1.1.7.3 et 1.3.4.2). Nous avons sélectionné deux échantillons métagénomiques, l'un provenant du microbiote intestinal (accession : SRS011405) et l'autre du microbiote de la bouche (accession : SRS016002), pour lesquels nous avons récupéré le jeu de séquençage complet Illumina :

— intestin : http://downloads.hmpdacc.org/data/Illumina/stool/SRS011405.tar.bz2 ;

— bouche : http://downloads.hmpdacc.org/data/Illumina/tongue_dorsum/SRS016002.tar.bz2.

Les lectures de ces jeux de données sont déjà nettoyées pour la qualité et nous n'avons appliqué aucun traitement supplémentaire. Pour ces échantillons, des tables d'OTU (appelées *final OTU table*) sont également disponibles, qui ont été construites à partir des données d'amplicon V1-V3 avec QIIME [4]. Pour chaque OTU est également fournie une séquence représentative [5].

4.4.2 Résultats

Les jeux de données Illumina ont été assemblés par MATAM et EMIRGE. Les jeux de données Illumina contenant des lectures de tailles variables, nous n'avons pas pu utiliser REAGO pour les assembler. Quant à SPAdes et MEGAHIT, les résultats obtenus étaient très incorrects, et ne sont pas présentés.

Comparaison des profils taxonomiques Pour pouvoir réaliser une comparaison avec l'analyse taxonomique basée sur les amplicons 16S, nous avons utilisé le classifieur RDP [98]. Deux classifications ont été construites : la première pour les séquences représentatives des OTU obtenus pour la région V1-V3, et la seconde pour les séquences reconstruites à partir du jeu de métagénomique. Nous avons sélectionné les assignations taxonomiques avec un seuil de confiance supérieur à 50% et calculé leur intersection.

4. http://hmpdacc.org/HMQCP/
5. http://hmpdacc.org/doc/ppAll_V13_map.txt

TABLE 4.6 – Comparaison de l'assignation taxonomique des assemblages et de l'analyse amplicon pour les jeux de données HMP. La colonne *Nb classes* contient le nombre total de classes taxonomiques distinctes identifiées par le classifieur RDP, et le nombre de ces classes validées par l'approche amplicon (indiquées entre parenthèses). La colonne *Nb genres* donne la même information au niveau du genre.

		Chimères (%)	Nb contigs	LT	Nb classes	Nb genres
SRS011405	MATAM	3.37%	218	187710	5 (4)	21 (17)
	EMIRGE	43.04%	273	393152	2 (2)	12 (8)
SRS016002	MATAM	4.92%	353	320748	13 (13)	31 (28)
	EMIRGE	46.01%	282	394087	12 (12)	25 (23)

La Table 4.6 présente les résultats de la comparaison des assignations réalisées avec le classifieur RDP pour les assemblages. Pour les deux échantillons, MATAM a assemblé plus de classes et de genres distincts qu'EMIRGE. De plus, la grande majorité de ces taxons sont validés par l'approche amplicon.

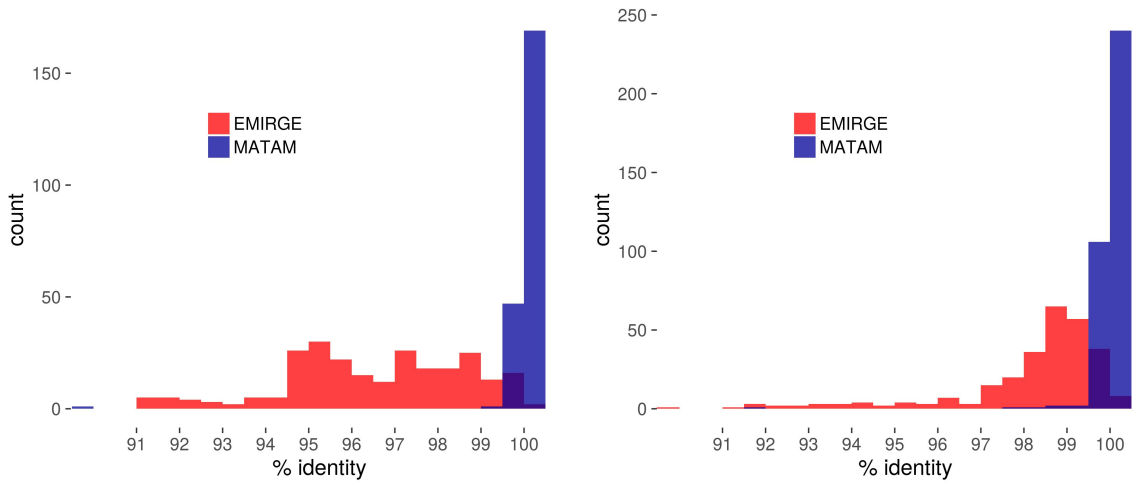
Cette comparaison sur les deux échantillons a aussi permis d'identifier trois genres assemblés à la fois par MATAM et EMIRGE, mais qui n'ont pas été identifiés par l'approche amplicon : *Odoribacter*, *Peptococcus* et *Bergeyella*. De nombreuses espèces microbiennes appartenant à ces genres sont connues pour appartenir au microbiote humain, et ont été identifiées par ailleurs dans d'autres échantillons similaires du projet HMP. Par conséquent, il semble envisageable que ces genres aient été manqués par l'approche amplicon tout en étant correctement reconstruits par MATAM et EMIRGE à partir du jeu de séquençage métagénomique complet.

Distribution des meilleurs matchs Une autre manière d'évaluer la qualité des assemblages de MATAM et EMIRGE est de comparer les séquences des contigs reconstruits avec les séquences de la base de référence SILVA complète. Comme ces échantillons de microbiotes proviennent d'environnements particulièrement bien étudiés, on s'attend à ce que la majorité des séquences d'ARNr 16S reconstruites appartiennent à des espèces déjà connues et dont un représentant proche est référencé dans la base SILVA.

Les contigs ont été alignés à l'aide de BLASTN [1] sur la base SSURef complète. Les paramètres de BLASTN ont été choisis pour maximiser sa sensibilité :

```
$ blastn -db SILVA_128_SSURef -query contigs.fa -out alignments.tab -task blastn -
    evalue 1e-5 -word_size 7 -outfmt '6 std qlen' -max_target_seqs 1 -max_hsps 1 -
    dust 'no' -xdrop_ungap 100 -xdrop_gap 150 -xdrop_gap_final 500
```

Pour chaque contig, nous avons sélectionné le meilleur alignement sur la base SILVA complète. Les Figures 4.4a et 4.4b représentent la distribution des identités



(a) Échantillon de microbiote intestinal humain SRS011405

(b) Échantillon de microbiote buccal humain SRS016002

FIGURE 4.4 – Distribution du pourcentage d'identité des meilleurs alignements sur SILVA 128 SSU Ref NR99 (SSURef complète).

des meilleurs alignements obtenus. On observe que la quasi-totalité des séquences de MATAM s'aligne sur une séquence d'ARNr 16S connue de SILVA avec plus de 99% d'identité, et qu'une grande majorité d'entre elles s'alignent à 100% d'identité. Ce résultat suggère que les séquences de MATAM pourraient être assignées au niveau de l'espèce, voire de la souche. Il semble également indiquer que seule une très faible proportion des deux microbiotes correspond à des espèces divergées inconnues.

D'un autre côté, les séquences d'EMIRGE offrent une image très discordante. Pour l'échantillon de microbiote buccal, la plupart des séquences s'alignent sur un ARNr 16S connu avec plus de 97%, mais seulement une toute petite partie d'entre elles s'aligne avec 100% d'identité (Figure 4.4a). Cette différence avec les séquences de MATAM est encore plus prononcée pour l'échantillon de microbiote intestinal. Pour ce dernier, seulement 47% des séquences d'EMIRGE s'alignent sur une séquence de SILVA avec plus de 97% d'identité (Figure 4.4b). Contrairement aux résultats de MATAM, ceux d'EMIRGE suggèrent donc que seulement une petite proportion de la diversité des deux microbiotes humains possède un représentant proche dans la base SILVA. Toutefois, en nous basant sur nos résultats précédents (4.2 et 4.3), nous pouvons faire l'hypothèse qu'au moins une partie de la diversité divergée apparemment inconnue qui peut être identifiée avec EMIRGE correspond en fait à des artefacts d'assemblage.

Cette discordance entre les deux méthodes, MATAM et EMIRGE, est loin d'être anodine. En effet, s'il s'agissait d'environnements moins bien étudiés que les microbiotes humains, on ne saurait dire si MATAM a un biais vers la reconstruction de séquences déjà connues et ne parvient pas à identifier de nouvelles espèces pourtant présentes, ou réciproquement, si EMIRGE tend à reconstruire des séquences n'ayant pas de réalité

biologique. De plus, on note que ni EMIRGE, ni MATAM ne sont parvenus à identifier la présence de certains genres inférés par les analyses amplicon. On peut suspecter ici un ensemble de biais affectant les deux approches, comme par exemple la robustesse aux faibles abondances des organismes recherchés. Ces observations suggèrent que, bien que de qualités appréciables, les assemblages de marqueurs produits par EMIRGE et MATAM peuvent et doivent encore être améliorés afin de permettre la production d'analyses taxonomiques fiables à partir de jeux métagénomiques complets.

4.5 Tests sur des jeux d'hybridation 16S

Au cours de cette thèse, nous avons eu l'occasion d'évaluer MATAM sur la reconstruction de séquences d'ARNr SSU dans des jeux de données d'hybridation 16S, grâce à une collaboration avec le laboratoire de Pierre Peyret à l'Université d'Auvergne.

L'hybridation, ou *capture de gènes*, est une nouvelle technique biomoléculaire en métagénomique, qui propose d'utiliser des sondes pour capturer physiquement des brins d'ADN portant un gène marqueur cible, avant la préparation de la librairie de séquençage [25]. L'application de cette technique à l'ARNr SSU permet ainsi d'enrichir fortement une librairie de séquençage métagénomique complet en séquences du marqueur. Cette approche a pour bénéfice ajouté de capturer également les régions flanquantes du marqueur, ce qui peut ainsi permettre leur assemblage postérieur.

Nous avons pu faire fonctionner MATAM sur deux jeux de données issus du même échantillon environnemental de sol :

— SRR3546814, correspondant à un séquençage métagénomique complet classique par un séquenceur Illumina de l'échantillon. Le jeu de données, nettoyé par nos soins, comporte 39 332 820 lectures appariées d'une longueur moyenne de 217 bp ;

— SRR3648004, un jeu d'hybridation 16S, fortement enrichi en séquences de l'ARNr 16S. Le jeu de données nettoyé comprend 7 473 580 lectures appariées d'une longueur moyenne de 224 bp.

4.5.1 Jeu de données de séquençage métagénomique complet

L'application de MATAM sur le jeu de séquençage métagénomique complet s'est bien déroulée. Sur les 39 millions de lectures initiales, 49 092 ont été identifiées comme appartenant à l'ARNr SSU, et un total de 326 scaffolds d'une taille moyenne de 718 bp ont été reconstruits. L'analyse taxonomique de ces scaffolds est cohérente avec le type d'environnement étudié, et compatible avec des expériences préalables réalisées avec EMIRGE par l'équipe de P. Peyret.

4.5.2 Jeux d'hybridation 16S

Nos tentatives d'assemblage du jeu d'hybridation 16S ont permis de toucher du doigt certaines limitations de la méthode et de son implémentation. Sur les 7,5 millions de lectures initiales, 4,4 millions (58,4%) ont été identifiées par SortMeRNA comme appartenant à l'ARNr SSU, soient plusieurs ordres de grandeur au-dessus de ce qui est attendu pour un jeu de séquençage métagénomique complet. À titre de comparaison, c'est le genre de proportions que l'on peut trouver dans un jeu de séquençage d'ARN total dans une étude de métatranscriptomique.

Sur ces 4,4 millions de lectures, la construction du graphe de chevauchement de MATAM devient problématique. Nous estimons que cette étape, si nous l'avions menée à son terme, se serait terminée en un temps déraisonnable (> plusieurs mois). Une exploration plus précise des alignements des lectures sur la base de référence partitionnée par SortMeRNA a permis d'identifier certaines références qui pouvaient être couvertes plusieurs centaines de milliers de fois. Or, notre algorithme de construction du graphe de chevauchement n'est pas prévu pour traiter une telle couverture. Il est de complexité quadratique en fonction de la couverture, ce qui explique ses difficultés pour traiter de telles quantités d'alignements.

Pour remédier à cette situation, nous avons cherché à forcer la reconstruction des séquences d'ARNr SSU pour ce jeu de données, en mettant en place une stratégie qui a nécessité plusieurs interventions manuelles sur les résultats intermédiaires de MATAM. Nous avons notamment filtré brutalement les séquences de références de la base de référence partitionnée sur lesquelles s'alignaient trop de lectures (couverture > 200x), ce qui a supprimé potentiellement l'information de séquences pour les espèces de l'échantillon les plus couvertes. Cette stratégie nous a toutefois permis de finir la reconstruction des séquences de l'ARNr SSU pour ce jeu de données, et MATAM a ainsi reconstruit 1 683 scaffolds de 700 bp en moyenne. L'analyse taxonomique de cet assemblage n'est à nouveau pas incohérente avec le type d'environnement étudié. L'exploration de ces résultats reste toutefois encore ouverte, et il serait intéressant de revenir étudier ce jeu de données si des modifications futures sont apportées à MATAM pour permettre à l'outil d'analyser de tels types de jeux de données.

4.6 MATAM, points forts et limitations

Les nombreux tests que nous avons réalisés avec MATAM sur des jeux de données variés ont montré la capacité de notre méthode à reconstruire correctement les séquences d'ARNr SSU dans la plupart des jeux de données de séquençage métagénomique complet. Toutefois, ces tests nous ont aussi permis d'identifier un certain nombre de limites intrinsèques à la méthode. Les deux principales sont le manque de robustesse vis-à-vis d'un taux de couverture trop élevé et la difficulté de reconstruire correctement des espèces inconnues.

Taux de couverture : : L'algorithme actuel pour la construction du graphe de chevauchements est inapplicable lorsqu'une ou plusieurs espèces de l'échantillon sont trop couvertes ($\gg$ 500x). C'est le genre de situation que l'on observe notamment dans des données de métatranscriptomique ou d'hybridation 16S, pour lesquelles certaines espèces peuvent être couvertes plusieurs centaines de milliers de fois. Pour y remédier, une première approche, sur laquelle nous sommes en train de travailler, consiste à réduire dynamiquement la couverture des références trop couvertes par échantillonnage. La principale difficulté est de trouver une stratégie de suppression des alignements qui impacte au minimum les liens entre les régions conservées d'espèces différentes. Une approche complémentaire pourrait viser à dé-répliquer les lectures de séquençage avant même de les aligner sur la base de référence partitionnée. Il faudrait alors pondérer les lectures par leur abondance respective et en tenir compte dans les étapes de construction et de compression du graphe de chevauchement. Enfin, dans le cas où on se limite aux chevauchements sans erreurs, on peut également imaginer d'indexer les lectures de séquençage.

Reconstruction de séquences inconnues absentes de la base de référence L'étape de scaffolding repose pour l'instant fortement sur l'ensemble des alignements obtenus entre les contigs et les séquences présentes dans la base de données de référence. Cela peut conduire à deux types d'erreurs, notamment lorsque la séquence à reconstruire est éloignée de toute séquence de la base de référence.

Le premier cas est celui où deux contigs de la même espèce sont alignés sur des références toutes différentes. Comme le regroupement des contigs est glouton, ces deux contigs ne seront pas regroupés et risquent d'être incorporés dans deux scaffolds distincts. En pratique, nous avons observé que cette situation ne se produisait que très rarement.

Le deuxième cas est celui où des contigs issus d'espèces distinctes sont alignés sur des références communes et regroupés ensuite au sein d'un même scaffold. Comme nous construisons actuellement une séquence consensus pour l'ensemble des espèces dont les contigs sont alignés sur la référence, il y a un risque de création de chimères. Pour résoudre ce défaut, il faudrait procéder plus finement en cherchant à distinguer les contigs appartenant à des espèces différentes. Cela pourrait se faire en tenant compte de l'information de chevauchement des contigs, ou bien de l'appariement des lectures.

Conclusions et perspectives

Dans ce manuscrit, j'ai présenté le résultat de mes recherches doctorales pour le développement d'une nouvelle méthode de reconstruction de séquences de marqueurs conservés à partir de données de séquençage métagénomique complet. Ce travail a pour objectif l'amélioration des analyses taxonomiques en métagénomique.

Notre méthode, MATAM, repose de manière originale sur un partitionnement rapide des lectures de séquençage sachant une base de séquences de références couvrant une large diversité taxonomique, préalablement à l'assemblage des partitions. Ceci permet de résoudre en partie un des principaux paradoxes de l'assemblage de jeux de données métagénomiques complets : assembler les *lectures similaires* provenant de génomes d'espèces proches, et ne pas assembler les *lectures similaires* provenant de génomes d'espèces distantes [6]. Le partitionnement préalable des lectures permet donc, dans ce contexte, de séparer les lectures issues de régions spécifiques du gène ciblé, ne provenant que d'une unique espèce, et les lectures issues de régions conservées entre de nombreuses espèces, donc impossibles à discriminer. Sur l'ensemble des jeux de données analysés, MATAM s'est montré plus performant que les autres méthodes testées, notamment EMIRGE et REAGO, en particulier sur deux points : sa capacité à différencier des séquences (très) proches, et à reconstruire des séquences avec un très faible taux d'erreur. En outre, MATAM est suffisamment efficace pour traiter sans encombre des jeux de métagénomique de taille standard. La méthode et ces résultats ont fait l'objet d'une publication acceptée dans la revue *Bioinformatics* [71] [7].

Ce travail soulève plusieurs questions et suggère plusieurs perspectives.

Ces différences peuvent être liées à des biais affectant chacune des deux approches, mais elles peuvent aussi résulter de différences réelles concernant l'information portée par les deux types de jeux de données. Un séquençage amplicon permet potentiellement de séquencer des espèces de très faibles abondances, mais au prix de nombreux biais. Un séquençage métagénomique complet, moins biaisé, comporte l'information de l'ensemble des génomes de l'échantillon, mais son analyse taxonomique peut être

6. *dixit* Mihai Pop, dans "First steps towards automated metagenomic assembly" : assemble similar sequences from related genomes together ; do not assemble similar sequences from unrelated genomes.

7. https://academic.oup.com/bioinformatics/article-abstract/doi/10.1093/bioinformatics/btx644/4457361/MATAM-reconstruction-of-phylogenetic-marker-genes

limitée par la faible couverture des espèces peu abondantes. En tout cas, il semble utile de mieux comprendre ce qui peut expliquer des conclusions divergentes entre les approches. Une nouvelle voie s'ouvre actuellement avec le développement de techniques de *capture de gènes*, qui permettent entre autres l'enrichissement en ARNr SSU des jeux de séquençage métagénomique complet. La combinaison de ce type de données avec des méthodes de reconstruction de séquences de marqueurs conservés pourrait permettre dans un avenir proche de réaliser des analyses taxonomiques routinières non biaisées, précises et sensibles aux espèces de faibles abondances.

Une autre perspective est liée aux développements récents des technologies de séquençage de troisième génération, permettant de séquencer de longs fragments d'ADN à faible coût. Cependant, les taux d'erreurs encore élevés (environ 10%) ne permettent pas actuellement l'utilisation de telles lectures pour l'analyse taxonomique fiable d'un échantillon, car les lectures obtenues diffèrent plus des séquences originales que ne diffèrent entre eux de nombreux gènes partagés par des espèces relativement différentes. Dans ces conditions, on peut toutefois envisager des approches hybrides qui combineraient lectures courtes de type Illumina et lectures longues. Les contigs générés par MATAM, dont nous avons vu au chapitre III qu'ils présentaient un taux d'erreur extrêment faible, seraient une bonne entrée en matière.

Enfin, dans ce manuscrit, nous avons travaillé avec un seul marqueur phylogénétique, l'ARNr SSU. L'analyse de ce marqueur est un passage obligé, car il s'agit d'un des gènes conservés les plus étudiés. Il est utilisé comme clé pour le référencement des nouvelles espèces, et il a servi à l'évaluation de la majorité des autres outils d'analyse taxonomique aujourd'hui disponibles. Nous observons toutefois que les limitations de ce marqueur commencent à apparaître clairement. Entre autres, la quantité de signal phylogénétique totale de l'ARNr SSU est limitée (environ 1500 sites disponibles pour l'ARNr 16S), et son analyse taxonomique ne permettra potentiellement pas d'aller au-delà de l'assignation à l'espèce dans de nombreux cas. En effet, chez de nombreuses souches qu'il serait pourtant utile de discriminer, l'ARNr 16S pourrait ne pas présenter de polymorphisme, aucune différence de séquence des ARNr 16S ne serait alors disponible pour identifier ces souches. L'approche de MATAM pourrait être appliquée avec profit à d'autres marqueurs, plus long, afin d'augmenter le signal et donc la résolution de l'analyse taxonomique. Elle pourrait en effet permettre de considérer des marqueurs tels que l'ARNr 23S / 28S de la grande sous-unité du ribosome (environ 3000 sites), voire éventuellement des opérons ribosomes complets. Il serait également possible, comme le proposent MetaPhyler [50] ou MetaPhlan [85] de considérer l'analyse de plusieurs marqueurs conservés de manière simultanée, dont certains ne sont présents que dans certains clades. À ce sujet, nous avons réalisé quelques tests concluants avec MATAM sur la reconstruction du marqueur mitochondrial *cytochrome oxidase sous-unité 1* dans des données simulées de séquençage métagénomique complet de communautés d'eucaryotes.

Bibliographie

[1] Stephen F. Altschul et al. « Basic Local Alignment Search Tool ». In : *Journal of Molecular Biology* 215.3 (oct. 1990), p. 403–410. ISSN : 0022-2836. DOI : 10.1016/S0022-2836(05)80360-2.

[2] Florent E. Angly et al. « Grinder: A Versatile Amplicon and Shotgun Sequence Simulator ». In : *Nucleic Acids Research* 40.12 (juil. 2012), e94. ISSN : 0305-1048. DOI : 10.1093/nar/gks251.

[3] Manimozhiyan Arumugam et al. « Enterotypes of the Human Gut Microbiome ». In : *Nature* 473.7346 (mai 2011), p. 174–180. ISSN : 0028-0836. DOI : 10.1038/nature09944.

[4] Anton Bankevich et al. « SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing ». In : *Journal of Computational Biology* 19.5 (mai 2012), p. 455–477. ISSN : 1066-5277. DOI : 10.1089/cmb.2012.0021.

[5] F. Baquero et C. Nombela. « The microbiome as a human organ ». English. In : *Clinical Microbiology and Infection* 18 (juil. 2012), p. 2–4. ISSN : 1198-743X. DOI : 10.1111/j.1469-0691.2012.03916.x.

[6] Mathieu Bastian, Sebastien Heymann, Mathieu Jacomy et al. « Gephi: An Open Source Software for Exploring and Manipulating Networks. » In : *Icwsm* 8 (2009), p. 361–362.

[7] Hayedeh Behzad, Takashi Gojobori et Katsuhiko Mineta. « Challenges and Opportunities of Airborne Metagenomics ». In : *Genome Biology and Evolution* 7.5 (mai 2015), p. 1216–1226. DOI : 10.1093/gbe/evv064.

[8] Yasmine Belkaid et Timothy W. Hand. « Role of the Microbiota in Immunity and Inflammation ». English. In : *Cell* 157.1 (mar. 2014), p. 121–141. ISSN : 0092-8674, 1097-4172. DOI : 10.1016/j.cell.2014.03.011.

[9] E. Blöchl et al. « Pyrolobus Fumarii, Gen. and Sp. Nov., Represents a Novel Group of Archaea, Extending the Upper Temperature Limit for Life to 113 Degrees C ». eng. In : *Extremophiles: Life Under Extreme Conditions* 1.1 (fév. 1997), p. 14–21. ISSN : 1431-0651.

[10] Celine Brochier-Armanet, Patrick Forterre et Simonetta Gribaldo. « Phylogeny and Evolution of the Archaea: One Hundred Genomes Later ». eng. In : *Current Opinion in Microbiology* 14.3 (juin 2011), p. 274–281. ISSN : 1879-0364. DOI : 10.1016/j.mib.2011.04.015.

[11] J. Gregory Caporaso et al. « QIIME Allows Analysis of High-Throughput Community Sequencing Data ». en. In : *Nature Methods* 7.5 (2010), p. 335–336. ISSN : 1548-7091. DOI : 10.1038/nmeth.f.303.

[12] Rayan Chikhi et Guillaume Rizk. « Space-Efficient and Exact de Bruijn Graph Representation Based on a Bloom Filter. » In : *WABI*. T. 7534. Lecture Notes in Computer Science. Springer, 2012, p. 236–248.

[13] Dylan Chivian et al. « Environmental Genomics Reveals a Single-Species Ecosystem Deep Within Earth ». en. In : *Science* 322.5899 (oct. 2008), p. 275–278. ISSN : 0036-8075, 1095-9203. DOI : 10.1126/science.1155495.

[14] The Human Microbiome Project Consortium. « Structure, function and diversity of the healthy human microbiome ». en. In : *Nature* 486.7402 (juin 2012), p. 207–214. ISSN : 0028-0836. DOI : 10.1038/nature11234.

[15] Cymon J. Cox et al. « The Archaebacterial Origin of Eukaryotes ». en. In : *Proceedings of the National Academy of Sciences* 105.51 (déc. 2008), p. 20356–20361. ISSN : 0027-8424, 1091-6490. DOI : 10.1073/pnas.0810647105.

[16] Joel B. Dacks et al. « The Changing View of Eukaryogenesis – Fossils, Cells, Lineages and How They All Come Together ». en. In : *J Cell Sci* 129.20 (oct. 2016), p. 3695–3703. ISSN : 0021-9533, 1477-9137. DOI : 10.1242/jcs.178566.

[17] A. P. Dempster, N. M. Laird et D. B. Rubin. « Maximum Likelihood from Incomplete Data via the EM Algorithm ». In : *Journal of the Royal Statistical Society, Series B* 39.1 (1977), p. 1–38.

[18] A. Dereeper et al. « Phylogeny.fr: robust phylogenetic analysis for the non-specialist ». In : *Nucleic Acids Research* 36.suppl_2 (juil. 2008), W465–W469. ISSN : 0305-1048. DOI : 10.1093/nar/gkn180.

[19] T. Z. DeSantis et al. « Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB ». en. In : *Applied and Environmental Microbiology* 72.7 (jan. 2006), p. 5069–5072. ISSN : 0099-2240, 1098-5336. DOI : 10.1128/AEM.03006-05.

[20] Andreas Döring et al. « SeqAn An efficient, generic C++ library for sequence analysis ». In : *BMC Bioinformatics* 9.1 (2008), p. 1–9. DOI : 10.1186/1471-2105-9-11.

[21] Robert C. Edgar. « MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput ». In : *Nucleic Acids Research* 32.5 (mar. 2004), p. 1792–1797. ISSN : 0305-1048. DOI : 10.1093/nar/gkh340.

[22] Robert C. Edgar. « Search and Clustering Orders of Magnitude Faster than BLAST ». In : *Bioinformatics* 26.19 (oct. 2010), p. 2460–2461. ISSN : 1367-4803. DOI : 10.1093/bioinformatics/btq461.

[23] Robert C Edgar. « UPARSE: Highly Accurate OTU Sequences from Microbial Amplicon Reads ». In : *Nature Methods* 10.10 (août 2013), p. 996–998. ISSN : 1548-7091, 1548-7105. DOI : 10.1038/nmeth.2604.

[24] Robert C. Edgar et al. « UCHIME improves sensitivity and speed of chimera detection ». In : *Bioinformatics* 27.16 (août 2011), p. 2194–2200. ISSN : 1367-4803. DOI : 10.1093/bioinformatics/btr381.

[25] Cyrielle Gasc et Pierre Peyret. « Revealing Large Metagenomic Regions through Long DNA Fragment Hybridization Capture ». In : *Microbiome* 5 (mar. 2017). ISSN : 2049-2618. DOI : 10.1186/s40168-017-0251-0.

[26] O. Gotoh. « An Improved Algorithm for Matching Biological Sequences ». eng. In : *Journal of Molecular Biology* 162.3 (déc. 1982), p. 705–708. ISSN : 0022-2836.

[27] R. Grossi et J. Vitter. « Compressed Suffix Arrays and Suffix Trees with Applications to Text Indexing and String Matching ». In : *SIAM Journal on Computing* 35.2 (jan. 2005), p. 378–407. ISSN : 0097-5397. DOI : 10.1137/S0097539702402354.

[28] Stéphane Guindon et al. « New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0 ». In : *Systematic Biology* 59.3 (mai 2010), p. 307–321. ISSN : 1063-5157. DOI : 10.1093/sysbio/syq010.

[29] Katharina J. Hoff. « The Effect of Sequencing Errors on Metagenomic Gene Prediction ». In : *BMC Genomics* 10 (2009), p. 520. ISSN : 1471-2164. DOI : 10.1186/1471-2164-10-520.

[30] M. Holtgrewe. « Mason – A Read Simulator for Second Generation Sequencing Data ». en. In : *Technical Report FU Berlin* (oct. 2010).

[31] Weichun Huang et al. « ART: A next-Generation Sequencing Read Simulator ». In : *Bioinformatics* 28.4 (fév. 2012), p. 593–594. ISSN : 1367-4803. DOI : 10.1093/bioinformatics/btr708.

[32] Weichun Huang et al. « ART: a next-generation sequencing read simulator ». In : *Bioinformatics* 28.4 (fév. 2012), p. 593–594. ISSN : 1367-4803. DOI : 10.1093/bioinformatics/btr708.

[33] Xiaoqiu Huang et Anup Madan. « CAP3: A DNA Sequence Assembly Program ». en. In : *Genome Research* 9.9 (jan. 1999), p. 868–877. ISSN : 1088-9051, 1549-5469. DOI : 10.1101/gr.9.9.868.

[34] Laura A. Hug et al. « A New View of the Tree of Life ». en. In : *Nature Microbiology* 1 (avr. 2016), p. 16048. ISSN : 2058-5276. DOI : 10.1038/nmicrobiol.2016.48.

[35] William R. Jeck et al. « Extending Assembly of Short DNA Sequences to Handle Error ». In : *Bioinformatics* 23.21 (nov. 2007), p. 2942–2944. ISSN : 1367-4803. DOI : 10.1093/bioinformatics/btm451.

[36] George Kunnackal John et Gerard E. Mullin. « The Gut Microbiome and Obesity ». en. In : *Current Oncology Reports* 18.7 (juil. 2016), p. 45. ISSN : 1523-3790, 1534-6269. DOI : 10.1007/s11912-016-0528-7.

[37] Konstantinos T. Konstantinidis et James M. Tiedje. « Genomic insights that advance the species definition for prokaryotes ». en. In : *Proceedings of the National Academy of Sciences of the United States of America* 102.7 (fév. 2005), p. 2567–2572. ISSN : 0027-8424, 1091-6490. DOI : 10.1073/pnas.0409727102.

[38] Evguenia KOPYLOVA, Laurent NOÉ et Hélène TOUZET. « SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data ». In : *Bioinformatics* 28.24 (2012), p. 3211–3217. DOI : 10.1093/bioinformatics/bts611.

[39] Evguenia KOPYLOVA et al. « SortMeRNA 2: ribosomal RNA classification for taxonomic assignation ». In : *Workshop on Recent Computational Advances in Metagenomics, ECCB 2014*. 2014.

[40] Eric S. LANDER et al. « Initial sequencing and analysis of the human genome ». en. In : *Nature* 409.6822 (fév. 2001), p. 860–921. ISSN : 0028-0836. DOI : 10.1038/35057062.

[41] Ben LANGMEAD et al. « Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome ». In : *Genome Biology* 10 (mar. 2009), R25. ISSN : 1474-760X. DOI : 10.1186/gb-2009-10-3-r25.

[42] Rasko LEINONEN, Hideaki SUGAWARA et Martin SHUMWAY. « The Sequence Read Archive ». In : *Nucleic Acids Research* 39.Database issue (jan. 2011), p. D19–D21. ISSN : 0305-1048. DOI : 10.1093/nar/gkq1019.

[43] Dinghua LI et al. « MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph ». In : *Bioinformatics* 31.10 (mai 2015), p. 1674–1676. ISSN : 1367-4803. DOI : 10.1093/bioinformatics/btv033.

[44] Heng LI. *Notes on Pairwise Alignment with Dynamic Programming*. Juil. 2017. DOI : 10.6084/m9.figshare.5223973.v2.

[45] Heng LI et Richard DURBIN. « Fast and Accurate Short Read Alignment with Burrows–Wheeler Transform ». In : *Bioinformatics* 25.14 (juil. 2009), p. 1754–1760. ISSN : 1367-4803. DOI : 10.1093/bioinformatics/btp324.

[46] Heng LI et Nils HOMER. « A Survey of Sequence Alignment Algorithms for Next-Generation Sequencing ». In : *Briefings in Bioinformatics* 11.5 (sept. 2010), p. 473–483. ISSN : 1467-5463. DOI : 10.1093/bib/bbq015.

[47] Ruiqiang LI et al. « De Novo Assembly of Human Genomes with Massively Parallel Short Read Sequencing ». en. In : *Genome Research* (déc. 2009). ISSN : 1088-9051, 1549-5469. DOI : 10.1101/gr.097261.109.

[48] Weizhong LI et Adam GODZIK. « Cd-Hit: A Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences ». In : *Bioinformatics* 22.13 (juil. 2006), p. 1658–1659. ISSN : 1367-4803. DOI : 10.1093/bioinformatics/btl158.

[49] Chaim LINHART et Ron SHAMIR. « The Degenerate Primer Design Problem: Theory and Applications ». In : *Journal of Computational Biology* 12.4 (mai 2005), p. 431–456. DOI : 10.1089/cmb.2005.12.431.

[50] Bo LIU et al. « Accurate and Fast Estimation of Taxonomic Profiles from Metagenomic Shotgun Sequences ». In : *BMC Genomics* 12.2 (2011), S4. ISSN : 1471-2164. DOI : 10.1186/1471-2164-12-S2-S4.

[51] Kenneth J. LOCEY et Jay T. LENNON. « Scaling laws predict global microbial diversity ». en. In : *Proceedings of the National Academy of Sciences* 113.21 (mai 2016), p. 5970–5975. ISSN : 0027-8424, 1091-6490. DOI : 10.1073/pnas.1521291113.

[52] Bin MA, John TROMP et Ming LI. « PatternHunter: Faster and More Sensitive Homology Search ». In : *Bioinformatics* 18.3 (mar. 2002), p. 440–445. ISSN : 1367-4803. DOI : 10.1093/bioinformatics/18.3.440.

[53] Udi MANBER et Gene MYERS. « Suffix Arrays: A New Method for On-Line String Searches ». In : *Proceedings of the First Annual ACM-SIAM Symposium on Discrete Algorithms*. SODA '90. Philadelphia, PA, USA : Society for Industrial and Applied Mathematics, 1990, p. 319–327. ISBN : 978-0-89871-251-3.

[54] Marcel MARTIN. « Cutadapt removes adapter sequences from high-throughput sequencing reads ». en. In : *EMBnet.journal* 17.1 (mai 2011), pp. 10–12. ISSN : 2226-6089.

[55] Marcel MARTINEZ-PORCHAS et al. « How Conserved Are the Conserved 16S-rRNA Regions? » en. In : *PeerJ* 5 (fév. 2017), e3036. ISSN : 2167-8359. DOI : 10.7717/peerj.3036.

[56] Konstantinos MAVROMATIS et al. « Use of simulated data sets to evaluate the fidelity of metagenomic processing methods ». en. In : *Nature Methods* 4.6 (juin 2007), p. 495–500. ISSN : 1548-7091. DOI : 10.1038/nmeth1043.

[57] Ernst MAYR. *The Growth of Biological Thought: Diversity, Evolution, and Inheritance*. en. Google-Books-ID: pHThtE2R0UQC. Harvard University Press, 1982. ISBN : 978-0-674-36446-2.

[58] Celine MERCIER et al. *SUMATRA and SUMACLUST: fast and exact comparison and clustering of sequences*. Available: http://metabarcoding.org/sumaclust. 2013.

[59] Barbara A. METHÉ et al. « A Framework for Human Microbiome Research ». In : *Nature* 486.7402 (juin 2012), p. 215–221. ISSN : 0028-0836. DOI : 10.1038/nature11209.

[60] Alla MIKHEENKO, Vladislav SAVELIEV et Alexey GUREVICH. « MetaQUAST: evaluation of metagenome assemblies ». In : *Bioinformatics* 32.7 (avr. 2016), p. 1088–1090. ISSN : 1367-4803. DOI : 10.1093/bioinformatics/btv697.

[61] Christopher S. MILLER et al. « EMIRGE: reconstruction of full-length ribosomal genes from microbial community short read sequencing data ». In : *Genome Biology* 12.5 (2011), R44. ISSN : 1474-760X. DOI : 10.1186/gb-2011-12-5-r44.

[62] Eugene W. MYERS et al. « A Whole-Genome Assembly of *Drosophila* ». en. In : *Science* 287.5461 (mar. 2000), p. 2196–2204. ISSN : 0036-8075, 1095-9203. DOI : 10.1126/science.287.5461.2196.

[63] Niranjan NAGARAJAN et Mihai POP. « Sequence Assembly Demystified ». en. In : *Nature Reviews Genetics* 14.3 (jan. 2013), p. 157–167. ISSN : 1471-0056. DOI : 10.1038/nrg3367.

[64] Eric P. NAWROCKI et Sean R. EDDY. « Infernal 1.1: 100-Fold Faster RNA Homology Searches ». In : *Bioinformatics* 29.22 (nov. 2013), p. 2933–2935. ISSN : 1367-4803. DOI : 10.1093/bioinformatics/btt509.

[65] S. B. Needleman et C. D. Wunsch. « A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins ». eng. In : *Journal of Molecular Biology* 48.3 (mar. 1970), p. 443–453. ISSN : 0022-2836.

[66] Nam-Phuong Nguyen et al. « A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity ». en. In : *npj Biofilms and Microbiomes* 2 (avr. 2016), npjbiofilms20164. ISSN : 2055-5008. DOI : 10.1038/npjbiofilms.2016.4.

[67] Sergey Nurk et al. « metaSPAdes: a new versatile de novo metagenomics assembler ». In : *arXiv:1604.03071 [q-bio]* (avr. 2016). arXiv: 1604.03071.

[68] Brian D Ondov, Nicholas H Bergman et Adam M Phillippy. « Interactive Metagenomic Visualization in a Web Browser ». In : *BMC Bioinformatics* 12 (sept. 2011), p. 385. ISSN : 1471-2105. DOI : 10.1186/1471-2105-12-385.

[69] Yu Peng et al. « IDBA – A Practical Iterative de Bruijn Graph De Novo Assembler ». en. In : *Research in Computational Molecular Biology*. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, avr. 2010, p. 426–440. DOI : 10.1007/978-3-642-12683-3_28.

[70] Yu Peng et al. « IDBA-UD: A de Novo Assembler for Single-Cell and Metagenomic Sequencing Data with Highly Uneven Depth ». In : *Bioinformatics* 28.11 (juin 2012), p. 1420–1428. ISSN : 1367-4803. DOI : 10.1093/bioinformatics/bts174.

[71] Pierre Pericard et al. « MATAM: Reconstruction of Phylogenetic Marker Genes from Short Sequencing Reads in Metagenomes ». In : *Bioinformatics* (). DOI : 10.1093/bioinformatics/btx644.

[72] Pavel A. Pevzner, Haixu Tang et Michael S. Waterman. « An Eulerian Path Approach to DNA Fragment Assembly ». In : *Proceedings of the National Academy of Sciences of the United States of America* 98.17 (août 2001), p. 9748–9753. ISSN : 0027-8424. DOI : 10.1073/pnas.171285098.

[73] Miguel Pignatelli et Andrés Moya. « Evaluating the Fidelity of De Novo Short Read Metagenomic Assembly Using Simulated Data ». In : *PLOS ONE* 6.5 (mai 2011), e19984. ISSN : 1932-6203. DOI : 10.1371/journal.pone.0019984.

[74] Junjie Qin et al. « A Human Gut Microbial Gene Catalog Established by Metagenomic Sequencing ». In : *Nature* 464.7285 (mar. 2010), p. 59–65. ISSN : 0028-0836. DOI : 10.1038/nature08821.

[75] Christian Quast et al. « The SILVA ribosomal RNA gene database project: improved data processing and web-based tools ». In : *Nucleic Acids Research* 41.D1 (jan. 2013), p. D590–D596. ISSN : 0305-1048. DOI : 10.1093/nar/gks1219.

[76] A L Reysenbach et al. « Differential Amplification of rRNA Genes by Polymerase Chain Reaction. » In : *Applied and Environmental Microbiology* 58.10 (oct. 1992), p. 3417–3418. ISSN : 0099-2240.

[77] Daniel C. Richter et al. « MetaSim—A Sequencing Simulator for Genomics and Metagenomics ». In : *PLOS ONE* 3.10 (oct. 2008), e3373. ISSN : 1932-6203. DOI : 10.1371/journal.pone.0003373.

[78] Torbjørn Rognes et al. « VSEARCH: A Versatile Open Source Tool for Metagenomics ». In : *PeerJ* 4 (oct. 2016). issn : 2167-8359. doi : 10.7717/peerj.2584.

[79] F. Sanger, S. Nicklen et A. R. Coulson. « DNA Sequencing with Chain-Terminating Inhibitors ». en. In : *Proceedings of the National Academy of Sciences* 74.12 (jan. 1977), p. 5463–5467. issn : 0027-8424, 1091-6490.

[80] Michael C. Schatz, Arthur L. Delcher et Steven L. Salzberg. « Assembly of Large Genomes Using Second-Generation Sequencing ». en. In : *Genome Research* 20.9 (jan. 2010), p. 1165–1173. issn : 1088-9051, 1549-5469. doi : 10.1101/gr.101360.109.

[81] Patrick D. Schloss et al. « Introducing Mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities ». en. In : *Applied and Environmental Microbiology* 75.23 (jan. 2009), p. 7537–7541. issn : 0099-2240, 1098-5336. doi : 10.1128/AEM.01541-09.

[82] Robert Schmieder et Robert Edwards. « Quality control and preprocessing of metagenomic datasets ». In : *Bioinformatics* 27.6 (mar. 2011), p. 863–864. issn : 1367-4803. doi : 10.1093/bioinformatics/btr026.

[83] *Scientists Find Traces of Sea Plankton on ISS Surface*. http://tass.com/non-political/745635. (Visité le 03/05/2017).

[84] Alexander Sczyrba et al. « Critical Assessment of Metagenome Interpretation - a benchmark of computational metagenomics software ». en. In : *bioRxiv* (jan. 2017), p. 099127. doi : 10.1101/099127.

[85] Nicola Segata et al. « Metagenomic Microbial Community Profiling Using Unique Clade-Specific Marker Genes ». en. In : *Nature Methods* 9.8 (août 2012), p. 811–814. issn : 1548-7091. doi : 10.1038/nmeth.2066.

[86] Migun Shakya et al. « Comparative metagenomic and rRNA microbial diversity characterization using Archaeal and Bacterial synthetic communities ». In : *Environmental microbiology* 15.6 (juin 2013), p. 1882–1899. issn : 1462-2912. doi : 10.1111/1462-2920.12086.

[87] Léa Siegwald et al. « Assessment of Common and Emerging Bioinformatics Pipelines for Targeted Metagenomics ». In : *PLOS ONE* 12.1 (jan. 2017), e0169563. issn : 1932-6203. doi : 10.1371/journal.pone.0169563.

[88] Jared T. Simpson et Richard Durbin. « Efficient Construction of an Assembly String Graph Using the FM-Index ». In : *Bioinformatics* 26.12 (juin 2010), p. i367–i373. issn : 1367-4803. doi : 10.1093/bioinformatics/btq217.

[89] Jared T. Simpson et Richard Durbin. « Efficient de Novo Assembly of Large Genomes Using Compressed Data Structures ». en. In : *Genome Research* 22.3 (jan. 2012), p. 549–556. issn : 1088-9051, 1549-5469. doi : 10.1101/gr.126953.111.

[90] Jared T. Simpson et Richard Durbin. « Efficient de novo assembly of large genomes using compressed data structures ». In : *Genome Research* 22.3 (2012), p. 549–556. doi : 10.1101/gr.126953.111.

[91] T. F. Smith et M. S. Waterman. « Identification of Common Molecular Subsequences ». In : *Journal of Molecular Biology* 147.1 (mar. 1981), p. 195–197. ISSN : 0022-2836. DOI : 10.1016/0022-2836(81)90087-5.

[92] Anja Spang et al. « Complex archaea that bridge the gap between prokaryotes and eukaryotes ». en. In : *Nature* 521.7551 (2015), p. 173–179. ISSN : 0028-0836. DOI : 10.1038/nature14447.

[93] Eric J. Stewart. « Growing Unculturable Bacteria ». en. In : *Journal of Bacteriology* 194.16 (août 2012), p. 4151–4160. ISSN : 0021-9193, 1098-5530. DOI : 10.1128/JB.00345-12.

[94] Shinichi Sunagawa et al. « Structure and function of the global ocean microbiome ». en. In : *Science* 348.6237 (mai 2015), p. 1261359. ISSN : 0036-8075, 1095-9203. DOI : 10.1126/science.1261359.

[95] « The Metagenomics and Metadesign of the Subways and Urban Biomes (MetaSUB) International Consortium Inaugural Meeting Report ». In : *Microbiome* 4 (juin 2016). ISSN : 2049-2618. DOI : 10.1186/s40168-016-0168-z.

[96] *The Tree of Life Web Project*. 2007. URL : http://tolweb.org.

[97] Colomban de Vargas et al. « Eukaryotic plankton diversity in the sunlit ocean ». en. In : *Science* 348.6237 (mai 2015), p. 1261605. ISSN : 0036-8075, 1095-9203. DOI : 10.1126/science.1261605.

[98] Qiong Wang et al. « Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy ». en. In : *Applied and Environmental Microbiology* 73.16 (août 2007), p. 5261–5267. ISSN : 0099-2240, 1098-5336. DOI : 10.1128/AEM.00062-07.

[99] C R Woese et G E Fox. « Phylogenetic Structure of the Prokaryotic Domain: The Primary Kingdoms. » In : *Proceedings of the National Academy of Sciences of the United States of America* 74.11 (nov. 1977), p. 5088–5090. ISSN : 0027-8424.

[100] C R Woese, O Kandler et M L Wheelis. « Towards a Natural System of Organisms: Proposal for the Domains Archaea, Bacteria, and Eucarya. » In : *Proceedings of the National Academy of Sciences of the United States of America* 87.12 (juin 1990), p. 4576–4579. ISSN : 0027-8424.

[101] Derrick E. Wood et Steven L. Salzberg. « Kraken: Ultrafast Metagenomic Sequence Classification Using Exact Alignments ». In : *Genome Biology* 15 (2014), R46. ISSN : 1474-760X. DOI : 10.1186/gb-2014-15-3-r46.

[102] Pablo Yarza et al. « The All-Species Living Tree Project: A 16S rRNA-Based Phylogenetic Tree of All Sequenced Type Strains ». In : *Systematic and Applied Microbiology* 31.4 (sept. 2008), p. 241–250. ISSN : 0723-2020. DOI : 10.1016/j.syapm.2008.07.001.

[103] Pablo Yarza et al. « Uniting the Classification of Cultured and Uncultured Bacteria and Archaea Using 16S rRNA Gene Sequences ». en. In : *Nature Reviews Microbiology* 12.9 (sept. 2014), p. 635–645. ISSN : 1740-1526. DOI : 10.1038/nrmicro3330.

[104] Cheng YUAN et al. « Reconstructing 16S rRNA genes in metagenomic data ». In : *Bioinformatics* 31.12 (juin 2015), p. i35–i43. ISSN : 1367-4803. DOI : 10.1093/bioinformatics/btv231.

[105] Daniel R. ZERBINO et Ewan BIRNEY. « Velvet: Algorithms for de Novo Short Read Assembly Using de Bruijn Graphs ». en. In : *Genome Research* 18.5 (jan. 2008), p. 821–829. ISSN : 1088-9051, 1549-5469. DOI : 10.1101/gr.074492.107.