Master mathématiques et applications
Ingénierie Statistique

Meta-analyse exploratoire de jeux de données métagénomiques publiques de différents écosystèmes microbiens en lien avec la chaîne alimentaire

Auteur :
Clément Poupelin

Référent universitaire :
Aymeric Stamm

Encadrants :
Christelle Hennequet-Antier
Cédric Midoux
Hélène Chiapello
Eric Dugat-Bony
Julien Tap

Avril 2024 - Septembre 2024

Remerciements

Je souhaite remercier mes encadrants, Christelle Hennequet-Antier, Cédric Midoux, Hélène Chiapello, Eric Dugat-Bony et Julien Tap, pour la confiance qu'ils m'ont accordée tout au long de mon stage, ainsi que pour leurs précieux conseils. Chacun d'entre eux a su m'apporter ses connaissances spécifiques dans son domaine respectif, ce qui a considérablement enrichi mon apprentissage et ma compréhension des différents aspects de la recherche en écologie microbienne.

Je voudrais également exprimer ma gratitude à toutes les personnes de l'unité MaIAGE pour leur accueil chaleureux et leur bienveillance. Leur esprit d'équipe et leur convivialité ont grandement contribué à faire de ce stage une expérience professionnelle unique et enrichissante. Grâce à eux, j'ai pu évoluer dans un environnement stimulant où le partage des idées et l'entraide étaient constamment présents, me permettant ainsi de grandir tant sur le plan personnel que professionnel.

Enfin, je tiens à remercier tous mes camarades de master, avec qui j'ai partagé ces deux dernières années, pour la solidarité et le soutien dont ils ont fait preuve. Leur amitié, leur esprit de collaboration, et leur enthousiasme ont rendu cette période d'études particulièrement mémorable.

Table de sigles, notations et définitions

Sigles

INRAE : Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement

MICA : Microbiologie et Chaîne Alimentaire

MATHNUM : Mathématiques et Informatique pour l'Environnement et l'Agronomie

MaIAGE : Mathématiques et Informatique Appliquées du Génome à l'Environnement

ADN : Acide DésoxyriboNucléique

ARN : Acides RiboNucléiques

POC : *Proof Of Concept*

ENA : *European Nucleotide Archive*

GPS : *Global Positionning System*

PCA : *Principal Component Analysis*

MDS : *MultiDimensional Scaling*

PCoA : *Principal Coordinates Analysis*

ANOVA : *ANalysis Of VAriance*

PERMANOVA : *PERmutational Multivariate ANalysis Of VAriance*

SPIEC-EASI : *SParse InversE Covariance Estimation for Ecological ASsociation Inference*

PLN : *Poisson LogNormal*

MB : Meinshausen et Bühlmann

GLasso : *Graphical Lasso*

CLR : *Centered LogRatio*

StARS : *Stability Approach to Regularization Selection*

TSS : *Total Sum of Squares*

AIC : *Akaike information criterion*

BIC : *bayesian information criterion*

CDD : Contrat à Durée Déterminée

Définitions

ASV (*amplicon sequence variant*) : désigne des séquences d'ADN individuelles récupérées à partir d'une analyse de gène marqueur à haut débit à la suite de l'élimination de séquences artefactes générées pendant les phases de séquençage.

CPU (*Central Processing Unit*) : unité de traitement ou microprocesseur principal d'un ordinateur.

Phylogénie : étude des êtres vivants afin de déterminer leurs liens de parenté.

Notations

$\mathbb{N}_0$: ensemble des entiers naturels avec 0 compris

SDP : ensemble des matrices symétriques définies positives.

X' : transposée de X

$\sum_k$: $\sum_{k=1}^{K}$ pour $k \in \{1, ..., K\}$

$\mathbb{1}$: le vecteur composé de 1, $\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$

Z^i : pour une matrice Z, cela correspond à la colonne i

$Z^{\neg i}$: pour une matrice Z, cela correspond à toute la matrice à laquelle on retire la colonne i

$\bar{x}$: moyenne empirique pour un vecteur $x \in \mathbb{R}^n$: $\frac{1}{n}\sum_{i=1}^{n} x_i$ pour

iid : indépendantes et identiquement distribuées

$F(k, n)$: loi de Fisher de paramètres $k > 0$ et $n > 0$

Table des matières

1 Introduction

Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement (INRAE) est un organisme national de recherche publique renommé au sein duquel j'ai eu l'opportunité de réaliser mon stage. Depuis plusieurs années, INRAE mène des recherches pour mieux décrire et comprendre le fonctionnement des écosystèmes microbiens, notamment au sein des départements MICA (Microbiologie et Chaîne Alimentaire) et MATHNUM (Mathématiques et Informatique pour l'Environnement et l'Agronomie). Ces départements jouent ainsi un rôle important dans l'avancement des connaissances sur les interactions microbiennes et leur impact sur l'environnement, l'agriculture et la santé humaine.
Aujourd'hui, les différents environnements présents dans la chaîne aimentaire ont largement été étudiés à l'aide de méthodes moléculaires et de bioinformatiques. Cependant, le cycle des microorganismes entre ces différents environnements qui constituent la chaîne alimentaire a été peu investi. Aussi, cela amène à s'interroger sur la possibilité d'intégrer des données issues de ces différents environnements afin de les étudier simultanément.

Ces dernières années, avec l'essor de la science ouverte (*open science*), une augmentation de la disponibilité des données de recherche a pu être constatée. En effet, la science ouverte est un mouvement qui cherche à rendre la recherche scientifique et les données qu'elle produit accessibles à tous. L'idée est ainsi de permettre une transparence des processus de recherche, facilitant la collaboration et l'enrichissement des connaissances, ce qui contribue donc à une recherche plus inclusive et participative, répondant aux défis actuels de notre société.
Par ailleurs, cette abondance de données offre des opportunités inédites pour les chercheurs en termes de réutilisation de données. Mais elle engendre également un besoin croissant de spécialistes en gestion et analyse de données, capables de manipuler des jeux de données de grande dimension.

Le sujet de mon stage porte sur l'analyse de jeux de données produits indépendamment, partagés librement dans le cadre de l'open science, qui caractérisent les organismes microbiens présents dans différents écosystèmes. L'objectif est de comparer, intégrer, interpréter et analyser la diversité microbienne et les interactions entre espèces dans divers environnements en lien avec la chaîne alimentaire ; mais également tenter de déterminer s'il existe des groupes d'organismes soit ubiquitaires (présents en différents endroits à la fois), soit spécifiques d'un ou de plusieurs environnements.
Dans ce contexte, plusieurs méthodes et approches statistiques ont été envisagées pour analyser ces données. Mon travail consistait donc à choisir, paramétrer et tester différentes méthodes adaptées aux données manipulées, mais également à identifier les opportunités et les obstacles à l'intégration des jeux de données publiques afin de promouvoir la science ouverte dans le domaine de la recherche en écologie microbienne.

2 Contexte

2.1 Stage et stucture d'accueil

Mon stage s'inscrit dans le cadre du Master mathématiques et applications - Ingénierie Statistique de l'Université de Nantes et s'est déroulé au sein de INRAE à Jouy-en-Josas, d'avril 2024 à septembre 2024.

J'ai intégré l'unité de recherche MaIAGE (Mathématiques et Informatique Appliquées du Génome à l'Environnement) sous la supervision principale des équipes Migale et StatInfOmics. Cette unité regroupe des mathématiciens, des informaticiens, des bioinformaticiens et des biologistes autour de questions de biologie et agro-écologie, allant de l'échelle moléculaire à l'échelle du paysage en passant par l'étude de l'individu, de populations ou d'écosystèmes. Rattachée aux départements MathNum et MICA, l'unité est structurée en cinq équipes :

- Dynenvie : modélisation dynamique et statistique pour les écosystèmes, l'épidémiologie et l'agronomie
- Bibliome : acquisition et formalisation de connaissances à partir de textes
- BioSys : biologie des systèmes
- StatInfOmics : bioinformatique et statistique des données "omiques"
- Migale : plateforme bioinformatique

L'inférence statistique et la modélisation dynamique sont des compétences fortes de l'unité, auxquelles s'ajoutent la bioinformatique, l'automatique et l'algorithmique. Les activités de recherche et d'ingénierie s'appuient également sur une forte implication dans les disciplines destinatrices : écologie, environnement, biologie moléculaire et biologie des systèmes.

Des séminaires sont organisés régulièrement dans l'unité, offrant des opportunités d'enrichir ses connaissances sur divers sujets et de suivre les avancées récentes dans les domaines de recherche de INRAE. J'ai également participé à deux formations organisées par la plateforme migale : Introduction à Linux (Trainings of the "Cycle bioinformatique par la pratique") et Analyse de données métagénomiques 16S (Metabarcoding analyses: from sequences to plots). Ces formations m'ont permis de renforcer mes compétences techniques et scientifiques afin de mieux comprendre le contexte biologique ainsi que les méthodologies associés à mon projet.

Ce rapport de stage présente une grande partie du travail effectué et les résultats obtenus sur le projet Open16S (voir 2.3). En complément, il est accompagné d'un blog [Poupelin 2024] qui est dédié à ce projet et sur lequel peuvent être retrouvés les codes et figures.
Ce travail a été réalisé avec *R* 4.4.1 via l'interface *R*Studio, en utilisant également les capacités de calcul offertes par les noeuds de clusters de la plateforme bioinformatique Migale allant jusqu'à 32 CPUs(*) , assurant ainsi une gestion efficace et performante des calculs en parallèle et des données volumineuses générées par les analyses métagénomiques. De plus, les résultats sont partagés sous la forme de documents Quarto qui sont une nouvelle alternative aux documents Rmarkdown et versionnés avec GitLab dans une démarche de reproductibilité des analyses.

2.2 Etude des écosystèmes microbiens par la métagenomique 16S

L'approche métagénomique permet d'étudier le microbiote, c'est à dire l'ensemble des micro-organismes (bactéries, archées, levures, ...) vivants dans un environnement spécifique appelé microbiome ou biotope. Elle consiste à séquencer simultanément les génomes de plusieurs micro-organismes différents présents dans un milieu donné. Les études métagénomiques connaissent une popularité croissante et représententent une approche moderne pour mieux connaitre les communautés microbiennes dans leurs environnements [Berg et al. 2020].

Après collecte d'échantillons de différents biotopes, vient l'étape de séquençage pour lire les bases nucléotidiques qui forment l'ADN, fournissant des informations sur la composition génétique des micro-organismes présents dans l'échantillon. Dans les études de métagénomique ciblée (ou "metabarcoding") le séquençage du gène codant pour la sous-unité 16S de l'ARN ribosomique (ARNr), est couramment utilisé car il est hautement représentatif des bactéries. De manière générale, le séquençage se concentre principalement sur des gènes marqueurs. Ces gènes sont à la fois présents dans tous les organismes ciblés et constitués de régions variables permettant d'identifier l'espèce correspondant à un gène séquencé.

Ensuite, les organismes séquencés sont identifiés via une comparaison taxonomique à des bases de données de référence pour retrouver les espèces correspondantes. Ainsi cela nous permet d'avoir pour chaque séquence des informations d'affiliation taxonomique allant du règne (bactéries, archées, ...) jusqu'au nom précis de l'espèce (`Figure 1`).

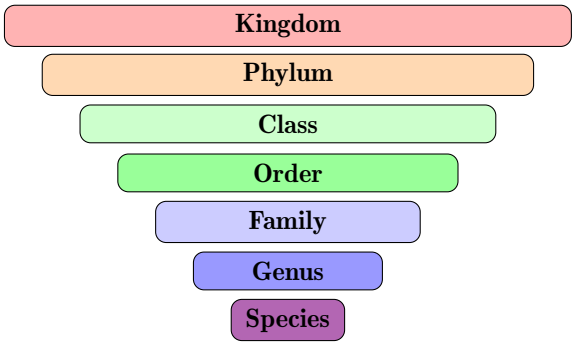


Figure 1 – Classification taxonomique

Enfin, les données sont transformées en table de comptage, où chaque ligne représente un organisme unique et chaque colonne un échantillon, avec les valeurs indiquant le nombre de fois que chaque séquence a été observée dans chaque échantillon.
Cela permet d'effectuer une analyse de diversité pour comprendre la richesse et la répartition des espèces dans différents biotopes, ce qui inclut des analyses de la diversité au sein d'un même échantillon et de la comparaison de la diversité entre différents échantillons.

En conclusion, les études métagénomiques s'enrichissent continuellement grâce aux avancées technologiques et méthodologiques. Cependant, il est important de noter que des biais peuvent apparaître pendant la collecte des échantillons, le séquençage, l'affiliation taxonomique et l'analyse des données, ce qui nécessite une attention particulière pour garantir la qualité et la fiabilité des résultats.

2.3 Projet Open16S

INRAE a été un organisme pionnier en matière de science ouverte et s'est doté d'une direction pour ce domaine dès 2020 (La science ouverte à INRAE).
Le projet Open16S, piloté par le département MICA de INRAE, est un projet original de type POC (preuve de concept) mis en place dans le cadre de la politique du département et qui vise à répondre aux défis liés à la réutilisation de données publiques dans le domaine de la métagénomique afin notamment de dégager de nouvelles hypothèses biologiques originales à tester à partir de l'intégration de jeux de données.

Ainsi, l'objectif principal du projet est d'aborder de manière transversale les écosystèmes microbiens associés à la chaîne alimentaire humaine en utilisant des jeux de données métagénomiques 16S provenant de divers biotopes. Ce projet implique 12 unités de recherche du département MICA, couvrant une gamme complète d'écosystèmes ciblés : des aliments (**food**), des fermenteurs (**digester**), des échantillons d'intestin humain (**human gut**) et des échantillons d'animaux (ici, le trayon de vache - **cow**).

La feuille de route actuelle du projet comprend plusieurs étapes essentielles :

- L'exploration des questions transversales d'écologie microbienne selon deux types d'approches statistiques : une approche *hypothèse-driven* (analyse guidée par les hypothèses) et une approche *data-driven* (analyse guidée par les données).

- L'intégration et l'exploration approfondie des jeux de données pour comprendre et formuler de nouvelles hypothèses sur les déterminants de la structuration des communautés microbiennes.

Ainsi, mon stage avait pour objectifs de réaliser dans un premier temps des analyses statistiques exploratoires et intégratives sur les données métagénomiques 16S puis d'inférer des réseaux d'association entre les espèces microbiennes.

2.4 Description des jeux de données

Au début de mon stage, j'ai disposé de 19 jeux de données contenant les résultats des études de métagénomique 16S effectuées indépendamment par différentes équipes de recherche. Ces jeux de données en lien avec la chaîne alimentaire se répartissent selon 4 biotopes : **food** (10 jeux de données), **digester** (6 jeux de données), **human gut** (3 jeux de données) et **cow** (1 jeux de données). Le nombre d'échantillons traités par projet varie de 24 pour le plus petit projet PRJNA685310 jusqu'à 500 pour PRJNA589612 ; les deux projets portant sur les aliments fermentés (respectivement fromages et levains de boulangerie). Et selon les projets, différentes régions du gène codant pour l'ARNr 16S ont été séquencées (Table 1).

TABLE 1 – Tableau récapitulatif des jeux de données

Métadonnées minimales associées aux projets				
ID projet	**Titre**	**Biotope**	**Echantillons**	**Région 16S**
PRJNA345074	Structural robustness of the gut mucosal microbiota is associated with Crohn's disease remission after surgery (article)	human gut	46	V6-V8
PRJNA459479	Deciphering intra-species bacterial diversity of meat and seafood spoilage microbiota using gyrB amplicon sequencing (article)	food	40	V3-V4
PRJEB39897	Deciphering Microbial Community Dynamics and Biochemical Changes During Nyons Black Olive Natural Fermentations (article)	food	215	V3-V4
PRJNA735449	Microbial community redundance in biomethanation systems lead to faster recovery of methane production rates after starvation (article)	digester	57	V3-V4
PRJEB39821	Indicative Marker Microbiome Structures Deduced from the Taxonomic Inventory of 67 Full-Scale Anaerobic Digesters of 49 Agricultural Biogas Plants (article)	digester	201	V3-V4
PRJNA578621	Robustness and efficacy of an inhibitory consortium against E. coli O26 :H11 in raw milk cheese (article)	food	108	V3-V4
PRJEB15657	Carrot Juice Fermentations as Man-Made Microbial Ecosystems Dominated by Lactic Acid Bacteria (article)	food	310	V4
PRJEB44120	Amplicon sequencing data for publication : Lactic starter dose shapes S. aureus and STEC O26 : H11 growth, and bacterial community patterns in raw milk uncooked pressed cheeses (article)	food	60	V3-V4
PRJEB21187	A Single Community Dominates Structure and Function of a Mixture of Multiple Methanogenic Communities (article)	digester	60	V4
PRJEB21193	A Single Community Dominates Structure and Function of a Mixture of Multiple Methanogenic Communities (article)	digester	42	V4
PRJNA589612	The diversity and function of sourdough starter microbiomes (article)	food	500	V4
PRJNA681555	Description of the temporal dynamics in microbial community composition and beer chemistry in sour beer production via barrel ageing of finished beers (article)	food	60	V4
PRJEB50379	Integration of multiomic data to characterize the influence of milk fat composition on Cantal-type cheese microbiota (article)	food	36	V3-V4
PRJEB51233	Influence of Post-Milking Treatment on Microbial Diversity on the Cow Teat Skin and in Milk (article)	food et cow	245	V3-V4
PRJNA685310	Temporal differences in microbial composition of Epoisses cheese rinds during ripening and storage (article)	food	24	V3-V4
PRJNA450513	Influence of support media supplementation to reduce the inhibition of anaerobic digestion by phenol and ammonia : Effect on degradation performances and microbial dynamics (article)	digester	59	V4-V5
PRJNA450311	Inhibition of anaerobic digestion by phenol and ammonia : Effect on degradation performances and microbial dynamics (article)	digester	96	V4-V5
PRJEB28341	Ceftriaxone and Cefotaxime Have Similar Effects on the Intestinal Microbiota in Human Volunteers Treated by Standard-Dose Regimens (article)	human gut	186	V4
PRJEB6070	Potential of fecal microbiota for early-stage detection of colorectal cancer (article)	human gut	255	V4

Pour garantir la cohérence et la qualité des données, plusieurs critères de sélection des jeux de données avaient été établis :

- **Publication** : les jeux de données devaient déjà avoir été valorisés dans des articles scientifiques. Compte tenu de l'aspect pilote du projet, il a aussi été décidé de prioriser les jeux de données produits par des équipes du département MICA.

- **Disponibilité publique** : les jeux de données devaient être disponibles en libre accès dans des entrepôts publics, ici l'ENA (European Nucleotide Archive), garantissant leur accessibilité et leur transparence pour la communauté scientifique.

- **Variété des biotopes** : les données devaient provenir d'un biotope en lien avec la chaîne alimentaire.

- **Uniformité de technique de séquençage** : les données devaient avoir été obtenues par la même technique de séquencage (ou similaires). Ici, nous sommes sur les techniques de séquençage Illumina et IonTorrent ciblant une région particulière du gène codant pour l'ARNr 16S et qui sont les plus communément utilisées pour identifier des organismes bactériens.

- **Uniformité des régions séquencées** : idéalement, les données devaient provenir du séquençage des mêmes régions variables du gène codant pour l'ARNr 16S. Toutefois, ce critère a été légèrement assoupli pour garantir un nombre suffisant de jeux de données pour représenter les différents biotopes.

Il faut noter que dans la communauté des chercheurs en génomique, il est obligatoire de déposer les données brutes de séquençage dans des bases de données publiques au moment de la publication d'un article. Ces dépôts permettent non seulement de partager les données avec la communauté scientifique mondiale mais aussi d'assurer la transparence et la reproductibilité des recherches. Ainsi, lors du dépôt, les chercheurs renseignent également des métadonnées, incluant des informations sur les conditions expérimentales, les protocoles utilisés, et l'origine des échantillons. Cependant, il s'avère que ces métadonnées sont parfois incomplètes ou imprécises. Néanmoins pour l'analyste des données (bioinformaticien, statisticien), ces métadonnées restent essentielles pour traiter les données et pour réaliser des comparaisons.

Comme mentionné précédemment, tous les jeux de données sont des études de métagénomiques 16S basées sur le séquençage du gène codant pour la sous-unité 16S de l'ARN ribosomique. Celui-ci est composé de régions très conservées et de neuf régions (V1-V9) qui sont dites hypervariables et qui permettent de distinguer efficacement différentes espèces de bactéries (`Figure 2`). Ceci évite donc de séquencer le génome complet permettant ainsi une réduction des coûts.

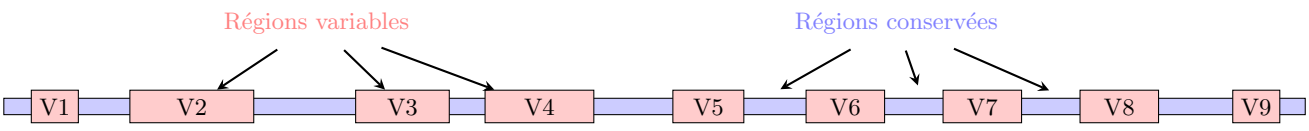


FIGURE 2 – Représentation théorique du 16S et de ses différentes régions

Les données brutes sont donc constituées de séquences (*reads* ou lectures) du gène codant pour l'ARNr 16S pouvant contenir une ou plusieurs régions hypervariables.

2.5 Pré-traitement bioinformatique

Avant mon arrivée, ces données brutes de séquençage et métadonnées associées ont été analysées via un pipeline bioinformatique (`Figure 3`) développé par C.Midoux et O.Rué (Open16s-WP2-results) en accord avec les pratiques définies par INRAE [Falentin et al. 2019] et composé de quatre grandes étapes appliquées pour chacun des projets.

1. Nettoyage des données brutes pour ne conserver que les séquences correspondant aux régions d'intérêts variables du gène codant pour l'ARNr 16S. Les *primers* ou amorces (définit dans les région conservées), utilisés comme points de départ pour la détection des zones d'intérêt lors du séquençage sont éliminés lors de cette étape.
2. Utilisation d'un modèle permettant de corriger les éventuelles erreurs de séquençage afin d'obtenir des séquences plus précises et fiables (utilisation de l'outil DADA2 [Callahan et al. 2016]). Cela permet de former des ASV(*)
3. Détection et suppression des possibles chimères, c'est-à-dire des séquences provenant d'origines différentes fusionnées et pouvant fausser l'analyse (utilisation de l'outil FROGS [Escudié et al. 2017]).
4. Affiliation taxonomique, via la base de référence silva 138.1 (utilisation de l'outil FROGS [Escudié et al. 2017]).
5. Stockage des données et métadonnées dans un objet phyloseq [McMurdie et Holmes 2013] via le package *R* du même nom. Cela permet d'organiser et gérer les données de manière à faciliter les analyses statistiques.

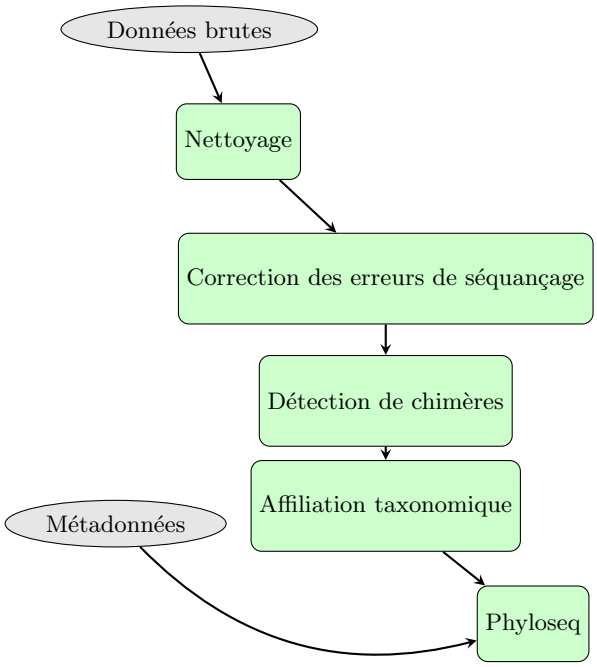


Figure 3 – Pipeline bioinformatique simplifiée

Un objet phyloseq est organisé de la manière suivante (Figure 4) :

- Un tableau d'abondance (**otu_table**)
- Un tableau taxonomique (**tax_table**)
- Une table de métadonnées (**sample_data**)

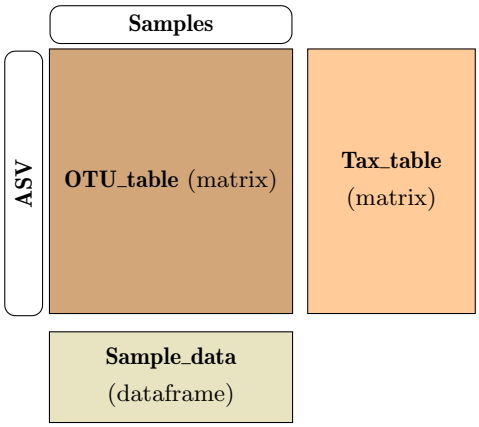


FIGURE 4 – Objet phyloseq

La table d'abondance (**otu_table**) est une matrice de comptage avec en ligne les ASV et en colonne les échantillons (*samples*). Chaque élément de la matrice représente l'abondance de l'ASV dans l'échantillon, c'est à dire le nombre de séquences (*reads*) de l'ASV présentes dans l'échantillon.
Les informations portant sur les échantillons en lien avec les conditions expérimentales (numéro de projet correspondant, organisme associé, condition, etc...) sont contenues dans la table **sample_data**.
Puis, l'assignation taxonomique des ASV (au niveau du règne jusqu'au niveau espèce lorsque cela est possible) est renseignée dans la table **tax_table**.

A l'issue de l'analyse bioinformatique, deux objets phyloseq on été construits. Le premier regroupant toutes les données du projet Open16S avec $n = 74833$ ASV (Figure 5) et le second ayant subi une agglomération par espèce avec $n = 3452$ espèces (Figure 6). C'est à dire que dans ce second objet, tous les ASV possédant la même affiliation taxonomique de règne jusqu'à espèce ont été fusionnées avec addition des comptages.

```
phyloseq-class experiment-level object
otu_table()   OTU Table:         [ 74833 taxa and 2565 samples ]
sample_data() Sample Data:       [ 2565 samples by 140 sample variables ]
tax_table()   Taxonomy Table:    [ 74833 taxa by 7 taxonomic ranks ]
```

FIGURE 5 – Objet "brut"

```
phyloseq-class experiment-level object
otu_table()   OTU Table:         [ 3452 taxa and 2565 samples ]
sample_data() Sample Data:       [ 2565 samples by 140 sample variables ]
tax_table()   Taxonomy Table:    [ 3452 taxa by 7 taxonomic ranks ]
```

FIGURE 6 – Objet avec agglomération par espèce

J'ai donc choisi d'effectuer les analyses statistiques sur l'objet phyloseq généré après agglomération des ASV par espèce. Cela permettait de traiter une matrice avec moins de zéros de plus petite dimension 3452×2565 et de simplifier l'interprétation des résultats en travaillant sur des noms d'espèces.

3 Préparation et premières explorations des données

Un des grands enjeux de la science ouverte est la description normalisée des données de la recherche via leurs métadonnées. Et malgré les efforts de normalisation des métadonnées des entrepôts génomiques (ENA), il a été constaté une grande hétérogénéité des métadonnées associées aux échantillons.
En effet, les différents jeux de données selectionnés étaient des études indépendantes qui n'étaient pas initialement prévues pour être intégrées. Cette intégration a donc conduit à des ensembles de données hétérogènes et souvent incomplets, posant des défis importants autour des données manquantes, de la normalisation des variables et de la grande dimension.

3.1 Création de variable d'intérêt à partir des métadonnées renseignées

Les métadonnées jouent un rôle important dans l'analyse de données, fournissant un contexte essentiel pour analyser et interpréter les résultats. Cependant, du fait de la nature "patchwork" des données qui sont comme un assemblage de plusieurs jeux de données indépendants, celles-ci sont complexes à traiter. En effet, il a été constaté une redondance, notamment avec plusieurs variables décrivant des aspects similaires de la localisation géographique. Par exemple, un jeu de données pouvait avoir les informations géographiques de l'échantillon dans une variable nommée **geo_loc_name** et un autre dans une variable **lat_lon**. Et si dans un jeu de données il avait été décidé de renseigner la ville de prélèvement et un autre les coordonnées GPS, un vrai problème de normalisation de l'information se posait. J'ai donc décidé de sélectionner quelques métadonnées indispensables en regroupant celles qui donnaient les mêmes informations (ou informations équivalentes). De plus, une variable **Biotope**, représentant les écosystèmes étudiés, a été créée manuellement et contient aucune valeur manquante.

J'ai donc sélectionné 6 métadonnées sur les 140 disponibles. La table de métadonnées (sample_data de l'objet phyloseq) est donc composée des variables :

- **Biotope** : lieu de vie associé à l'échantillon
- **Sample** : identifiant de l'échantillon
- **Location** : information de localisation géographique de prélèvement de l'échantillon
- **Source** : information sur la source de prélèvement de l'échantillon
- **organism_name** : organisme ou environnement séquencé
- **PRJN** : numéro de projet auquel l'échantillon appartient

Ici, les variables **Location** et **Source** sont la fusion de plusieurs autres variables (respectivement 8 et 5 variables). Ces fusions ont été réalisées en manipulant les données grâce aux outils du package *R* {*tidyverse*} et se trouve dans la section 0 du blog [Poupelin 2024].

3.2 Co-occurences des espèces dans les différents biotopes

Dans un premier temps, nous souhaitons étudier la présence d'espèces trans-biotopes (communes à plusieurs biotopes) de la chaines alimentaires avec une analyse de co-occurences. Cette analyse exploratoire a permis de détecter, visualiser et quantifier le nombre de co-occurences d'espèces entre deux ou plus biotopes.

A l'aide d'Upset plot (similaires aux diagrammes de Venn) j'ai développé deux fonctions sous *R*. Une première fonction permet de construire la matrice de combinaison entre les biotopes et d'extraire le nom des espèces de cette combinaison (Figure 7). Une espèce est présente dans un biotope si elle est comptée au moins une fois dans un des échantillons de ce biotope. Puis, la seconde fonction permet de visualiser de manière simple l'Upset plot qui découle de cette matrice de combinaison (Figure 8).

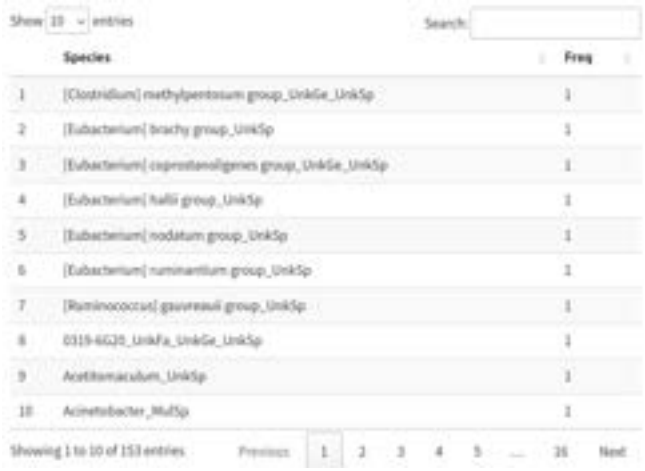


FIGURE 7 – Extraction des espèces communes aux quatre biotopes ($n = 153$)

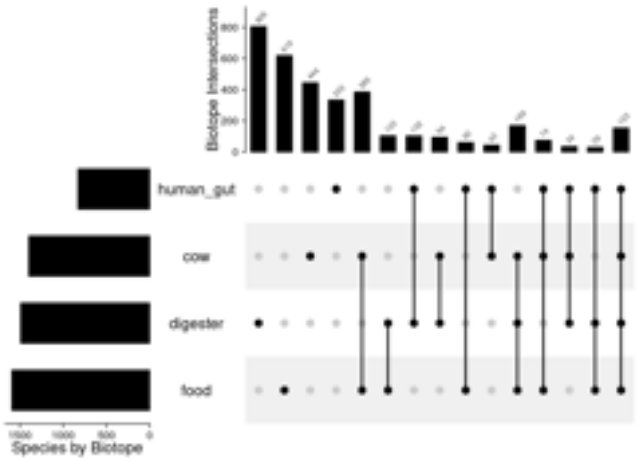
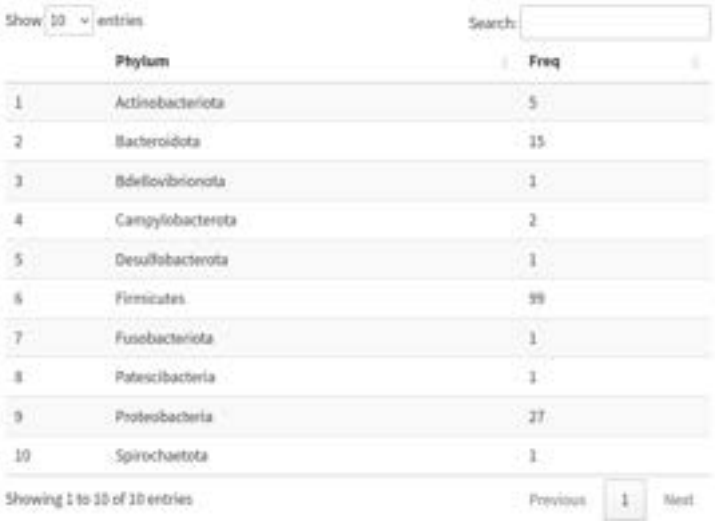


FIGURE 8 – Upset plot des co-occurences d'espèces pour les différents biotopes

Cela nous permet donc, dans un premier temps, de visualiser et quantifier le nombre de co-occurrences d'espèces puis ensuite de les identifier. Nous pouvons alors constater que 2202 espèces restent spécifiques à un biotope mais 1246 sont partagées par nos biotopes. Et nous avons même 153 espèces communes aux quatre biotopes ce qui nous conforte dans la possibilité d' illustrer des relations inter-biotopes.

Maintenant, il faut tout de même préciser que beaucoup de ces espèces communes extraites étaient renseignées comme inconnues (Figure 7) puisque la notation "UnkSp" correspond à *unknown species* et "MulSp" correspond à *multi-affiliation species*. Ce qui veut dire que des séquences correspondent à des espèces notifiées comme inconnues où alors qu'elles correspondent de manière équivalente à plusieurs affiliations possibles dans la base de données de référence.
Cela nous donne un premier aperçu des difficultés rencontrées lorsque l'on se positionne sur une résolution au niveau espèces. Et c'est pourquoi, les fonctions ont été construites de sorte à pouvoir se placer à un autre niveau de taxonomie. Cela nous rendra moins résolutif mais nous pourrons statuer sur des informations plus fiables.

Ainsi, au niveau phylum nous pouvons voir les co-occurrences détectées dans les quatre biotopes (Figure 9). Nous pouvons constater une prédominance d'espèces appartenant au phylum des firmicutes. Mais nous avons aussi la présence élevée d'espèces appartenant aux Proteobacteria et Bacteroidota dont l'un est très présent dans l'homme et l'autre plus répandu dans le sol ou les intestins d'animaux.



Show 10 entries Search:

	Phylum	Freq
1	Actinobacteriota	5
2	Bacteroidota	15
3	Bdellovibrionota	1
4	Campylobacterota	2
5	Desulfobacterota	1
6	Firmicutes	99
7	Fusobacteriota	1
8	Patescibacteria	1
9	Proteobacteria	27
10	Spirochaetota	1

Showing 1 to 10 of 10 entries Previous 1 Next

FIGURE 9 – Extraction des phylum communs aux quatre biotopes

En conclusion, cette analyse nous montre bien que des espèces sont présentes sur plusieurs biotopes de la chaîne alimentaire ce qui amène à envisager la présence de relations entre ces espèces. Mais, cela nous montre aussi que ces études restent très dépendantes des possibilités d'affiliation taxonomique et qu'étudier les espèces implique que certaines soient affiliées à des *unknown* ou qu'elles soient *multi-affiliation*. C'est pour cela qu'une pratique courante est d'effectuer une agglomération à des niveaux taxonomiques supérieurs même si cela implique une perte de résolution (Figure 1).

4 Analyse de diversité

La diversité, ou richesse, fait référence au nombre de taxons (ASV, espèces, ...) différents, comptés ou estimés, dans un système ou échantillon étudié (par simplicité de compréhension, on utilisera tout du long de se rapport le terme espèce plutôt que taxon). Pour estimer la diversité, nous faisons certaines hypothèses. Premièrement, les systèmes d'où viennent les prélèvements sont connus et la taxonomie des différentes espèces peut être identifiée. Ensuite, les systèmes sont considérés comme équidistant dans le sens que si nous rajoutons une espèce dans un système, cela correspondra à une augmentation de diversité d'une unité. Cela ne dépendra donc pas du fait que l'espèce soit proche ou non des autres (sauf si nous utilisons des distances spécifiques reposant sur la phylogénie(*)).

En écologie microbienne, deux grands types d'analyses de diversité sont réalisées en fonction des questions biologiques posées :

- **α-diversité** : elle correspond à la diversité dans un système uniforme de taille fixe. Dans notre contexte, cela signifie que l'on regarde la diversité d'espèces dans chacun des échantillons.

- **β-diversité** : elle mesure à quel point des systèmes locaux sont différents. Dans notre contexte, cela signifie que l'on va regarder si différents échantillons ont une diversité proche ou non.

Ces mesures sont donc très importantes dans les analyses microbiennes, car elles permettent d'identifier les facteurs qui influencent la diversité microbienne et de comprendre les dynamiques des communautés microbiennes. Nous pouvons alors nous poser la question de savoir si les données du projet Open16S issues d'une démarche de science ouverte permettent de retrouver les caractéristiques de diversité propre à chaque biotope même si les échantillons proviennent d'études indépendantes et de matériels biologiques différents.

4.1 Méthodes et théorie

4.1.1 Indices d'alpha-diversité

Différents indices sont utilisés pour quantifier l'α-diversité qui est spécifique à un échantillon donné. Les indices que nous allons présenter sont les plus couramment utilisés et permettent de mesurer non seulement le nombre d'espèces présentes mais aussi leur abondance relative et leur distribution.
L'abondance relative p_s fait référence au rapport entre le nombre de fois qu'une espèce s est observée dans un système par rapport au nombre total d'espèces S dans ce même système, avec $s = 1 \dots S$.

Ainsi, soit p_s avec $s \in \{1, ..., S\}$ la probabilité d'appartenir à l'espèce s et c_i, $i \in \mathbb{N}$, le nombre d'espèces observées i fois.

Richesse observée

$$S_{rich} = \sum_s 1_{p_s>0} = \sum_i c_i$$

Cet indice représente le nombre d'espèces différentes observées dans un échantillon. Il ne prend pas en compte l'abondance relative des espèces, offrant ainsi une mesure brute de la richesse de l'échantillon et est fortement influencé par les espèces rares.

Chao1

$$S_{Chao1} = S_{rich} + \hat{c}_0$$

L'indice de Chao1 est défini par le nombre d'espèces différentes observées dans l'échantillon auquel on associe une estimation du nombre d'espèces non observées $\hat{c}_0$.
Cette estimation se fait à partir de celles observées une et deux fois (annexe).

Shannon entropy

$$S_{Shan} = -\sum_s p_s \ln(p_s)$$

L'indice de Shannon représente l'entropie de la distribution de l'abondance relative des espèces dans un échantillon. Ainsi, il prend en compte à la fois la richesse et l'abondance relative, offrant une mesure plus nuancée de la diversité en tenant compte de l'équité des espèces présentes (annexe).

Inverse Simpson

$$S_{InvSimp} = \frac{1}{p_1^2 + ... + p_s^2}$$

Cet indice évalue quant à lui l'inverse de la probabilité que deux séquences tirées aléatoirement dans un échantillon puissent appartenir à la même espèce. Plus l'indice est élevé, plus la diversité est importante et il est influencé par les espèces très abondantes (annexe).

4.1.2 ANOVA

L'α-diversité mesure la diversité au sein d'un échantillon et produit donc une valeur de diversité par échantillon. Dans ce contexte l'objectif de l'ANOVA, ou analyse de variance, va être de tester l'influence de facteurs expérimentaux sur l'α-diversité en comparant les diversités moyennes entre plusieurs groupes définis par ces facteurs.

Pour expliquer la variabilité de l'α-diversité en fonction d'un facteur contenant I groupes (par exemple le biotope à quatre groupes), l'ANOVA à un facteur se base sur un modèle de la forme :

$$y_{ij} = \mu + \beta_i + \varepsilon_{ij} \qquad \text{pour } i \in \{1, ..., I\} \text{ et } j \in \{1, ..., n_i\}$$

- y_{ij} la variable réponse ou variable à expliquer
- n_i la taille du groupe i
- μ constante (*intercept*)
- β_i l'effet du facteur à tester
- $\varepsilon_{ij} \underset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ l'erreur

Nous posons les moyennes suivantes :

$$\bar{y_i} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} \text{ la moyenne par groupe} \quad , \quad \bar{y} = \frac{1}{\sum_{i=1}^{I} n_i} \sum_{i=1}^{I} \sum_{j=1}^{n_i} y_{ij} \text{ la moyenne totale}$$

Sous l'hypothèse nulle selon laquelle les diversités moyennes des différents groupes sont égales et en supposant que les hypothèses de l'ANOVA sont vérifiées (indépendances des erreurs, homoscédasticité, normalité), la statistique de test s'écrit :

$$F = \frac{\text{SSB}}{\text{SSW}} \frac{n - I}{I - 1} \quad , \quad F \sim F(I - 1, n - I)$$

Avec

$$\text{SSB} = \sum_{i=1}^{I} n_i (\bar{y_i} - \bar{y})^2 \quad \text{et} \quad \text{SSW} = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \bar{y_i})^2$$

SSB représente la variabilité inter-classes et SSW la variabilité intra-classe. Et la variabilité totale se décompose en

$$\text{SST} = \text{SSB} + \text{SSW}$$

.

Cela permet donc de tester l'égalité des moyennes de la variable réponse (α-diversité) entre les groupes (par exemple les biotopes).

4.1.3 Indices de bêta diversité

Contrairement à la diversité α, la diversité β permet de comparer la composition microbienne entre plusieurs communautés microbiennes (d'échantillons différents). Plusieurs distances ou pseudo-distances ont été développées et sont communément utilisées.

Notons $n_{s,1}$ le nombre de séquences de l'espèce s dans l'échantillon 1 et $n_{s,2}$ celui dans l'échantillon 2.

Jaccard

La distance de Jaccard (1 − indice de Jaccard) est définie par

$$d_J = \frac{\sum_s 1_{1_{n_{s,1}>0, n_{s,2}=0}} + 1_{1_{n_{s,2}>0, n_{s,1}=0}}}{\sum_s 1_{n_{s,1}+n_{s,2}>0}}$$

et représente le nombre d'espèces spécifiques de chaque échantillon rapporté à la somme totale des espèces.

Bray-Curtis

L'indice de dissimilarité de Bray-Curtis est un indice de dissimilarité défini par

$$d_{BC} = \frac{\sum_s |n_{s,1} - n_{s,2}|}{\sum_s (n_{s,1} + n_{s,2})}$$

et prend directement en compte l'abondance des espèces. Il prend la valeur 0 si les échantillons sont identiques et 1 s'ils sont complétement dissemblants.

De plus, il existe d'autres indices de β-diversité : *Unifrac* et *Weighted Unifrac*. Ils se basent sur la distance phylogénétique et ne sont pas applicables pour les données du projet Open16S. Ces indices utilisent une longueur de branche de l'arbre phylogénétique construit à partir de l'alignement de tous les ASV (74833, `Figure 5`). Un tel alignement n'est pas possible à produire et par conséquent, ces indices ne sont pas adaptés à la compléxité des données due à l'intégration.

4.1.4 Visualisation par méthode d'ordination

Les méthodes d'ordination sont des méthodes statistiques de représentation graphique utilisées pour visualiser les échantillons dans un espace de dimension inférieure tout en préservant la structure globale des données.
Nous pouvons citer deux méthodes dites d'ordination sans contrainte, c'est à dire sans hypothèse sur la structure des données ou les relations entre variables : Analyse en composantes principales (PCA) et le positionnement multidimensionnel (MDS ou PCoA).
La PCA s'applique directement sur des matrices multidimensionnelles de données individuelles où chaque échantillon (communauté microbienne) est décrit par l'abondance des ASV ou espèces. Et le MDS permet de projeter les échantillons dans un espace euclidien à partir des matrices de distance ou dissimilarité en préservant les proximités. Ces méthodes sont intéressantes pour identifier les facteurs impliqués dans la structuration des communautés (annexe).

4.1.5 PERMANOVA

La β-diversité mesure la diversité entre plusieurs échantillons, souvent représentée par des matrices de distance ou de dissimilarité entre les paires d'échantillons. Ces données de distances ne suivent généralement pas une distribution normale et peuvent avoir des structures complexes de dépendance, ce qui rend l'ANOVA inappropriée.
Nous avons alors réalisé l'analyse PERMANOVA (analyse de la variance par permutation) qui :

- permet de partitionner la variance à partir de distances ou dissimilarités en fonction de facteurs expérimentaux (modélisation similaire à l'ANOVA)
- se base sur des permutations et ne fait pas d'hypothèses strictes sur la distribution des données (normalité).
- teste si les centres des groupes (les centroîdes) diffèrent, en prenant en compte les distances multidimensionnelles entre les échantillons.

Cette partie s'inspire de la publication de [J.Anderson 2005]. L'analyse PERMANOVA à un facteur (par exemple le biotope) se base sur une matrice de distances ou dissimilarités $D = \{d_{ij}\}_{i,j\in\mathbb{R}}$.
Soit N le nombre d'échantillons.

La somme des carrés des écarts totale se définit comme suit :

$$\text{SST} = \frac{1}{N}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{ij}^2$$

Il faut ensuite définir la somme des carrés des écarts entre les observations et les centroïdes au sein des groupes :

$$\text{SSW} = \frac{1}{n}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{ij}^2\delta_{ij}$$

avec n le nombre de répétitions (réplicats) dans le groupe et $\delta_{ij} = 1$ si i et j sont dans le même groupe et 0 sinon

De plus, nous avons la relation $\text{SST} = \text{SSB} + \text{SSW}$ où SSB représente la somme des carrés des écarts entre les centroïdes des groupes et le centroîde global.

Cela permet de construire la statistique de test

$$F = \frac{\text{SSB}}{\text{SSW}}\frac{N-I}{I-1} \quad \text{avec } I \text{ le nombre de groupes}$$

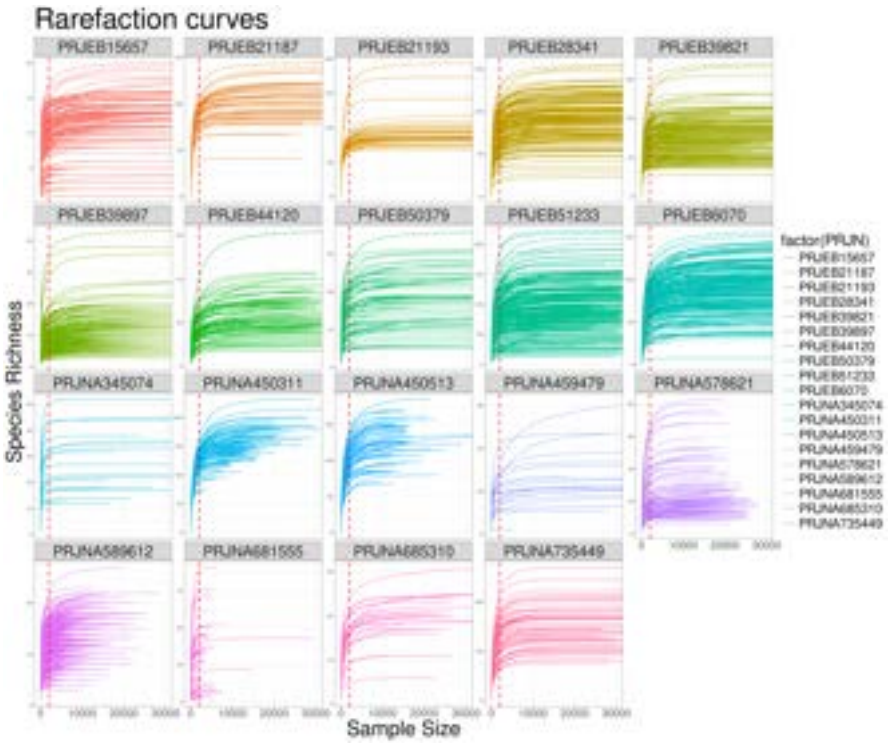
Ensuite, les données sont permutées K fois. À chaque permutation, de nouveaux labels sont réattribués aux échantillons et une statistique F_k est calculée, pour k allant de 1 à K.
Puis la p-value est déterminée par

$$p = \frac{\text{Nombre de } |F_k| \geq F}{\text{Nombre total de permutations}}$$

4.1.6 Rarefaction et Breakaway

Pour l'α-diversité, de nombreux indices dépendent de la taille de l'échantillon, caractérisée par le nombre de séquences à l'intérieur de celui-ci. Dans les données disponibles du projet Open16s, le nombre total de séquences varie de 122 pour un des échantillons du projet PRJEB15657 à 473768 pour un échantillon du projet PRJEB6070. Cette différence vient de ce qu'on appelle la profondeur (ou couverture) de séquençage qui est un effet expérimental. Et puisque les projets ont été fait indépendamment, les profondeurs de séquençage sont également différentes. L'objectif étant de comparer l'α-diversité entre les échantillons, les données d'abondances sont parfois raréfiées sur la base du minimum de séquences détectées dans un échantillon.

Supposons que nous avons deux échantillons de tailles n_1 et n_2 telles que $n_1 < n_2$. De manière simple, la raréfaction consiste à tirer aléatoirement (tirage avec remise) n_1 séquences que l'on gardera pour le deuxième échantillon de taille $n_2^{rar} = n_1$. Les espèces détectées dans ce nouvel échantillon raréfié varient en fonction du tirage effectué et c'est pourquoi la procédure est itérée plusieurs fois afin d'obtenir différentes valeurs d'espèces détectées. Par la suite, cela nous permet d'obtenir le nombre moyen de fois qu'une espèce est observé dans un échantillon. Les courbes de raréfaction peuvent ensuite mettre en évidence que la richesse observée augmente en fonction de la taille de l'échantillon (Figure 10).



La majorité des courbes approche une asymptote, ce qui signifie qu'augmenter le nombre de séquences ne fera probablement pas découvrir de nouvelles espèces. En m'appuyant sur la littérature, j'ai choisi un seuil de raréfaction à 2000 séquences pour effectuer les analyses de diversité. Ainsi, environ 1.29% des échantillons seront retirés de l'analyse car ils possèdent moins de 2000 séquences.

FIGURE 10 – Courbe de raréfaction d'échantillons selon le projet

La raréfaction permet de rendre comparable les mesures de diversité malgré l'exclusion de séquences au cours du processus et potentiellement la perte d'espèces rares, ce qui suscite la controverse depuis quelques années [McMurdie et Holmes 2014].

En réponse à cette controverse, une autre méthode a été développée [Willis et Bunge 2015] et implémentée dans le package R *Breakaway*. Pour estimer et modéliser la richesse observée, l'objectif est d'estimer le nombre d'espèces non observées même si les échantillons sont de tailles différentes en utilisant une régression non linéaire sur les ratios de fréquence $\frac{c_{i+1}}{c_i}$ (c_i défini en 4.1.1). J'ai donc testé cette méthode afin de voir si elle apportait des résultats différents.

4.2 Résultats

4.2.1 Alpha Diversité

Nous avons exploré l'α-diversité des échantillons selon le biotope en utilisant plusieurs méthodes d'estimation.

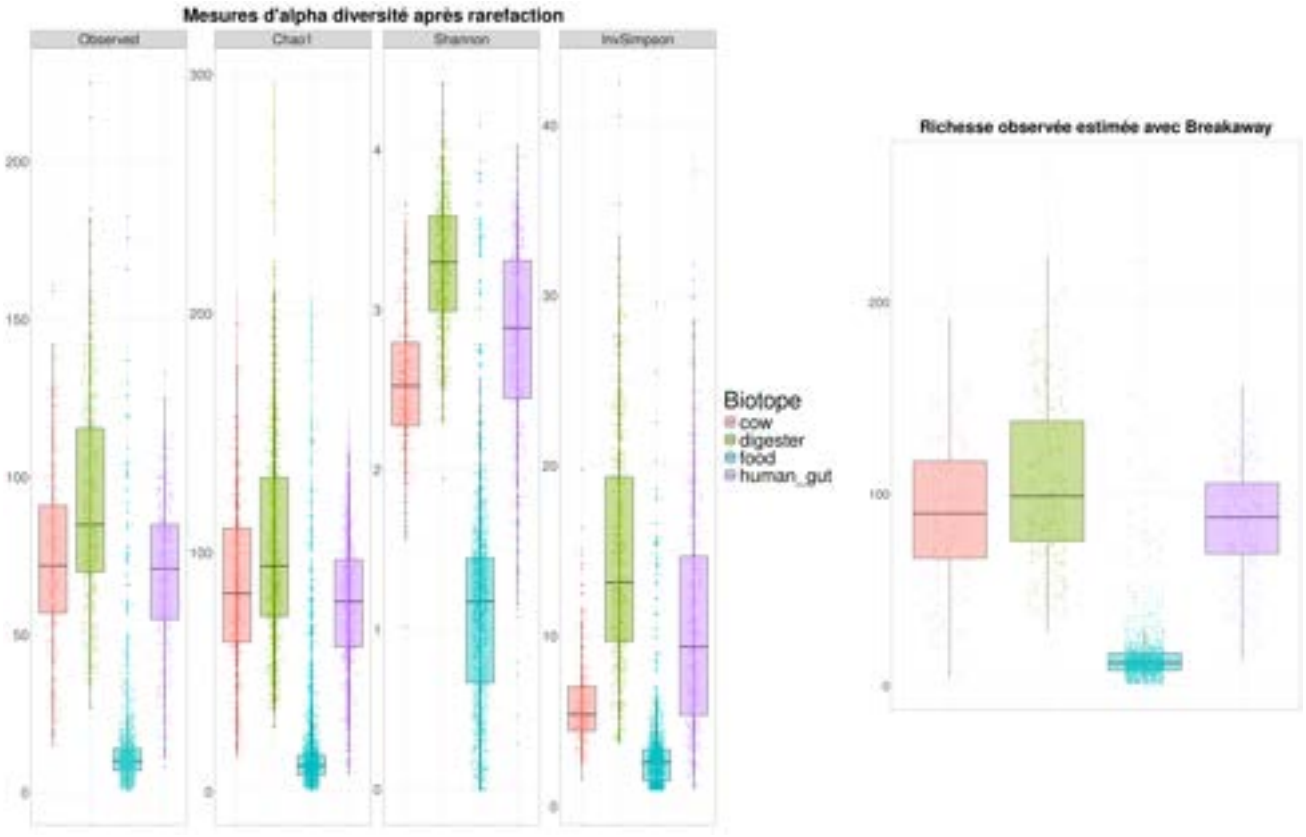


FIGURE 11 – Boxplot de l'α-diversité des échantillons selon leur biotope

Ces résultats permettent de visualiser une diversité élevée pour quelques échantillons des biotopes tels que **digester**, **cow** et **human gut**, et cela, quelle que soit l'indice utilisé (Figure 11).
Toutefois, il faut également noter que le grand nombre d'échantillons du biotope **food** contribue à ce que l'α-diversité soit étendue et très variable : des échantillons ont une richesse faible et d'autres une richesse aussi élevée que dans les autres biotopes. En effet, le biotope **food** est représenté par des échantillons provenant de sources très variées (Figure 12).

Puis, nous pouvons noter également une différence forte dans les valeurs obtenues avec l'indice de Shannon s'expliquant par une meilleur prise en compte l'équité entre les espèces.

Aussi, l'estimation de la richesse observée avec le package *breakaway* est similaire à celle estimée par Chao1 après la raréfaction. Par conséquent, nous ne développerons pas d'avantage sur cette méthode.

Ainsi, il semble clair que le biotope a un effet sur la diversité présente dans les échantillons. Pour confirmer cela, nous avons effectué plusieurs ANOVA en utilisant différentes mesures de l'α-diversité (Observed, Chao1, Shannon, Inverse Simpson) comme variables de réponse, avec le biotope comme facteur explicatif.

Pour chaque mesure de l'α-diversité, nous avons testé l'hypothèse nulle selon laquelle les diversités moyennes des différents biotopes sont égales. Un rejet de l'hypothèse nulle indiquerait que le biotope a un effet significatif sur la diversité.

TABLE 2 – Analyse de variance

Response : Observed					
	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
Biotope	3	2833802	944601	1864.5	<2.2e-16
Residuals	2523	1278199	507		

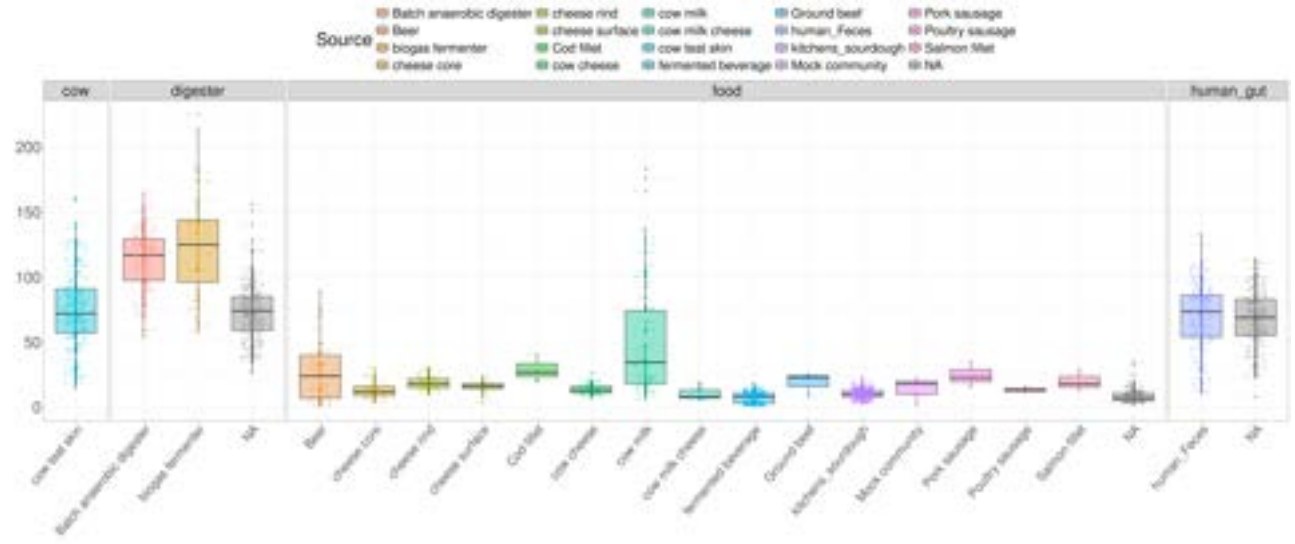
Response : Chao1					
	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
Biotope	3	3631410	1210470	1669	<2.2e-16
Residuals	2523	1829810	725		

Response : Shannon					
	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
Biotope	3	2192.66	730.89	2200.8	< 2.2e-16
Residuals	2523	837.88	0.33		

Response : Inverse Simpson					
	Df	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
Biotope	3	59429	19809.7	978.51	< 2.2e-16
Residuals	2523	51077	20.2		

Pour toutes les mesures de l'α-diversité, les p-values obtenues sont inférieures à 0.05 (Table 2), ce qui conduit au rejet de l'hypothèse nulle. Cela signifie que les diversités moyennes diffèrent significativement entre les biotopes, confirmant ainsi l'impact du biotope sur la variabilité de l'α-diversité.

De plus, grâce aux métadonnées, nous pouvons analyser plus en détail quelles sources et quels organismes contribuent également à la diversité intra-biotope.



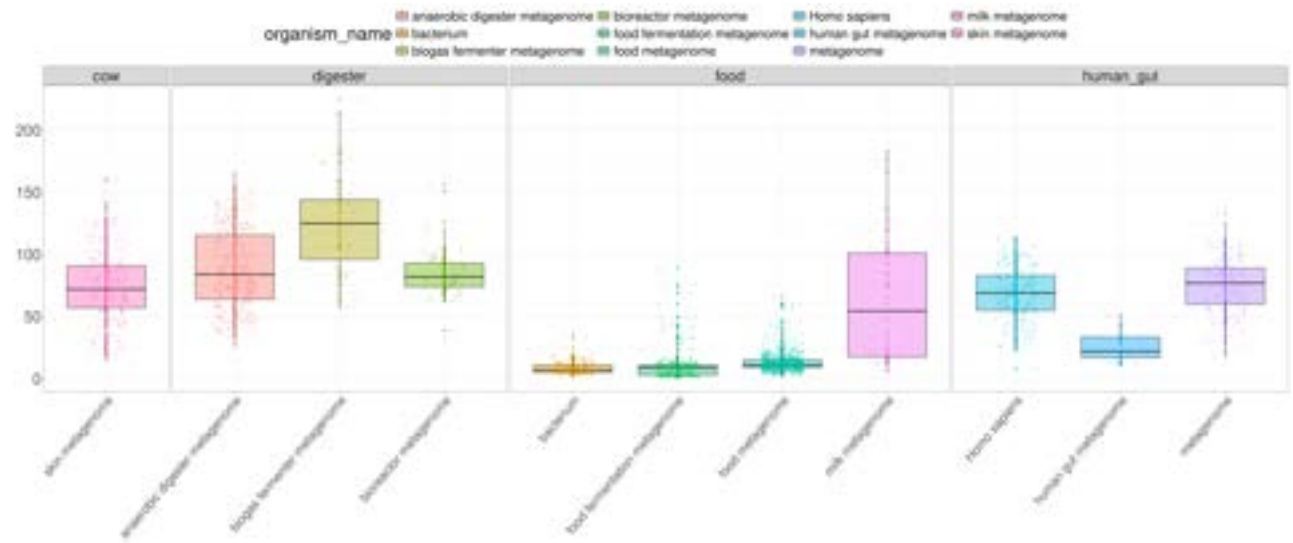


FIGURE 12 – Richesse observée pour les échantillons selon les variables **Source** et **organism_name**

Nous constatons dans un premier temps que *cow teat skin, batch anaerobic digester, biogas fermenter, beer, cow milk et human feces* ont une richesse plus élevée au sein des biotopes et une plus grande variabilité (`Figure 12`).

Aussi, dans le biotope **food** le lait (avant quelconque transformation) présente la diversité la plus élevée. En revanche, les autres sources alimentaires montrent une diversité plus faible (`Figure 12`), ce qui peut être dû à des procédés de transformation ou de conservation qui réduisent la diversité microbienne.

Bien entendu, ces analyses dépendent fortement de la qualité des métadonnées disponibles. Certaines informations peuvent être manquantes ou imprécises comme par exemple *métagenome* de la variable **organism_name** et dans le biotope **human gut** qui est sans précision particulière (`Figure 12`).

Cette analyse révèle donc que le biotope a un impact sur la diversité microbienne. Cependant, le manque de précision des métadonnées ne nous permet pas d'aller plus loin dans les analyses de l'α-diversité pour la compréhension des environnements microbiens de la chaîne alimentaire.

4.2.2 Bêta Diversité

Pour évaluer la β-diversité, nous avons donc utilisé les indices de Jaccard et de Bray-Curtis avec la méthode de positionnement multidimensionnel (MDS).

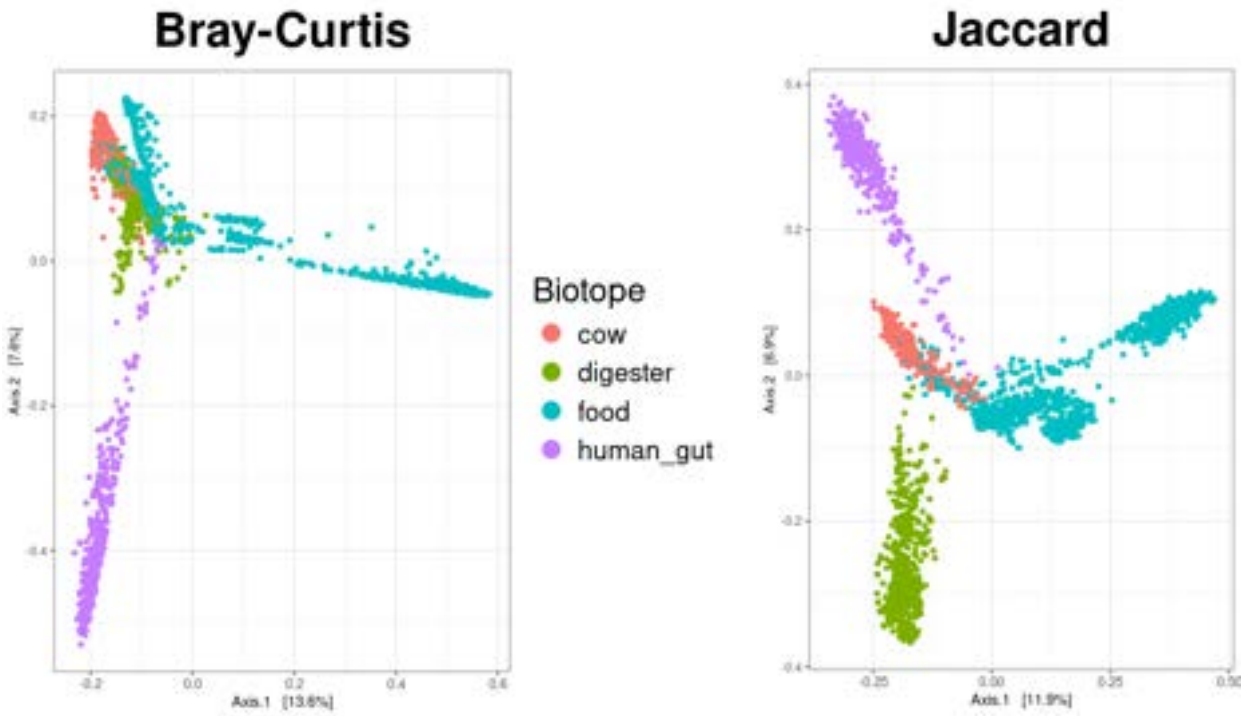


FIGURE 13 – Représentation en deux dimensions de la $\beta-$diversité en fonction du biotope

Ici encore, le biotope semble impacter la β-diversité. En effet, nous observons une structuration des échantillons en fonction des biotopes dans le plan euclidien (Figure 13). Notons aussi que, bien que le pourcentage d'explication des axes semble faible, il est conforme à la littérature des analyses de données métagénomiques en écologie microbiennes. La réduction de dimension nous fait passer de plus de 2500 axes (l'équivalent d'un axe par échantillons) à deux axes. Et donc, l'analyse de β-diversité bénéficie des avancées méthodologiques dans le domaine de la réduction de dimension.

De plus, il est intéressant de noter que le pourcentage d'explication est plus élevé en utilisant l'indice de Bray-Curtis qui prend en compte l'abondance des espèces contrairement à Jaccard.

Aussi, il est intéressant de regarder la β-diversité en fonction d'autres variables et principalement voir l'impact des sources de prélèvement. Nous pouvons retrouver ici présence de structuration selon la source de prélèvement (Figure 14). Mais nous constatons à nouveau une limitation due aux données manquantes et à la normalisation des termes (voir 3.1).

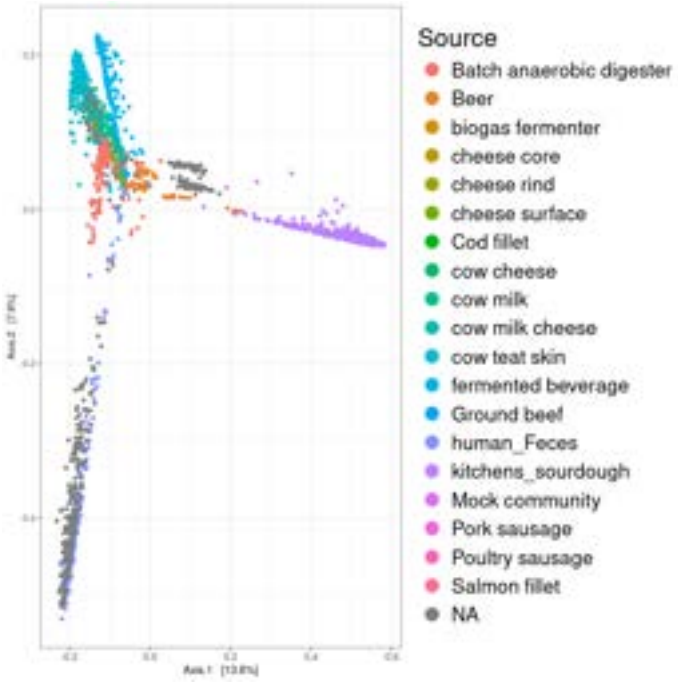


FIGURE 14 – Représentation en deux dimensions de la $\beta-$diversité en fonction de la variable **Source** pour la dissimilarité de Bray Curtis

L'analyse PERMANOVA, basée sur une matrice de distance ou dissimilarité, a été utilisée pour tester l'impact du biotope. L'hypothèse nulle est définie par les centroïdes et la dispersion des groupes équivalents entre les différents biotopes (cela signifie que les échantillons des différents biotopes ne diffèrent pas significativement en terme de composition microbienne).

TABLE 3 – Analyse PERMANOVA

Jaccard					
	Df	**SumOfSqs**	**R2**	**F value**	**Pr(>F)**
Biotope	3	236.91	0.20212	215.66	0.01
Residuals	2554	935.22	0.79788		
Total	2557	1172.13	1.00000		

Bray Curtis					
	Df	**SumOfSqs**	**R2**	**F value**	**Pr(>F)**
Biotope	3	225.97	0.19723	209.17	0.01
Residuals	2554	919.74	0.80277		
Total	2557	1145.71	1.00000		

Nous constatons un rejet de l'hypothèse nulle pour les deux indices, avec des p-values inférieures à 0.05, indiquant que les échantillons des différents biotopes diffèrent significativement (Table 3). Nous pouvons aussi voir que la variable biotope semble expliquer environ un cinquième de la variablilité totale.

Pour conclure cette partie, l'étude de la diversité, tant α que β, a confirmé que les communautés microbiennes sont différentes selon les biotopes et aussi les sources de la chaîne alimentaire. Nous avons retrouvé des résultats connus de la littérature, notamment une richesse d'espèce plus importante dans les écosystèmes humains (**human gut**) et animaux (**cow**), ainsi que pour les fermenteurs (**digester**). Tout cela offre une base solide pour des études plus approfondies sur les interactions entre les différentes espèces.

5 Inférence de réseau

L'inférence de réseau dans l'étude des données microbiennes permet de modéliser et de visualiser les interactions, au sens mathématique, entre différentes espèces microbiennes (en biologie, une interaction entre des espèces doit être validée expérimentalement). Ainsi, il est possible d'identifier les espèces liées à un environnement spécifique ou plus transversales.

De manière concrète, un réseau est composé de nœuds et d'arêtes. Chaque nœud correspond à une espèce tandis que les arêtes correspondent aux interactions entre les espèces.

L'inférence de réseau est une thématique de recherche en plein essor en biologie et en particulier en écologie microbienne. Plusieurs modèles de réseaux ont été développés récemment reposant sur des méthodologies différentes. Certaines méthodes peuvent donc être plus adaptés à des types spécifiques de données ou des questions de recherche.

Parmi ces méthodes, il a été choisi dans le cadre du stage de se concentrer sur celles reposant sur l'estimation des dépendances conditionnelles et de tester plus particulièrement :

- SPIEC-EASI (*Sparse InversE Covariance estimation for Ecological Association and Statistical Inference*) [D.Kurtz et al. 2015]
- PLN (*Poisson LogNormal*) [Chiquet, Mariadassou et Robin 2021].

L'application des réseaux dans l'étude des données microbiennes apporte une richesse d'informations et d'interprétations qui sont essentielles pour comprendre la complexité des interactions microbiennes. Mon idée ici était avant tout de découvrir et tester plusieurs approches d'inférence de réseau. De plus, cela a permis l'utilisation du modèle PLN co-développé par un chercheur de l'unité MaIAGE.

5.1 Méthodes et théorie

5.1.1 Filtration et agglomération

Dans l'objectif d'inférer un réseau calculable et interprétable, j'ai élaboré plusieurs stratégies pour construire des jeux de données adaptés. En effet, les techniques de réseau nécessite que les données ne soient pas de trop grande dimension sinon le calcul serait trop coûteux (voir impossible) et il y a aussi un risque qu'il ne soit pas interprétable car beaucoup trop dense.

Tout d'abord, un filtrage global basé sur la profondeur de séquençage (quantité totale de séquences obtenues par échantillon) est effectué. Nous rappelons que le nombre total de séquences produites est différent entre les échantillons et est lié à la technique de séquençage (illustration lors de la raréfaction 4.1.6)).
Le seuil, fixé à 10000 séquences, a été choisi en s'inspirant de la littérature. Cela implique la suppression des échantillons possèdant moins de 10000 séquences lues et quelque soit le projet d'où viennent les échantillons.
Ainsi, 2210 échantillons sont conservés sur les 2558 (plus de 86.4% des échantillons).

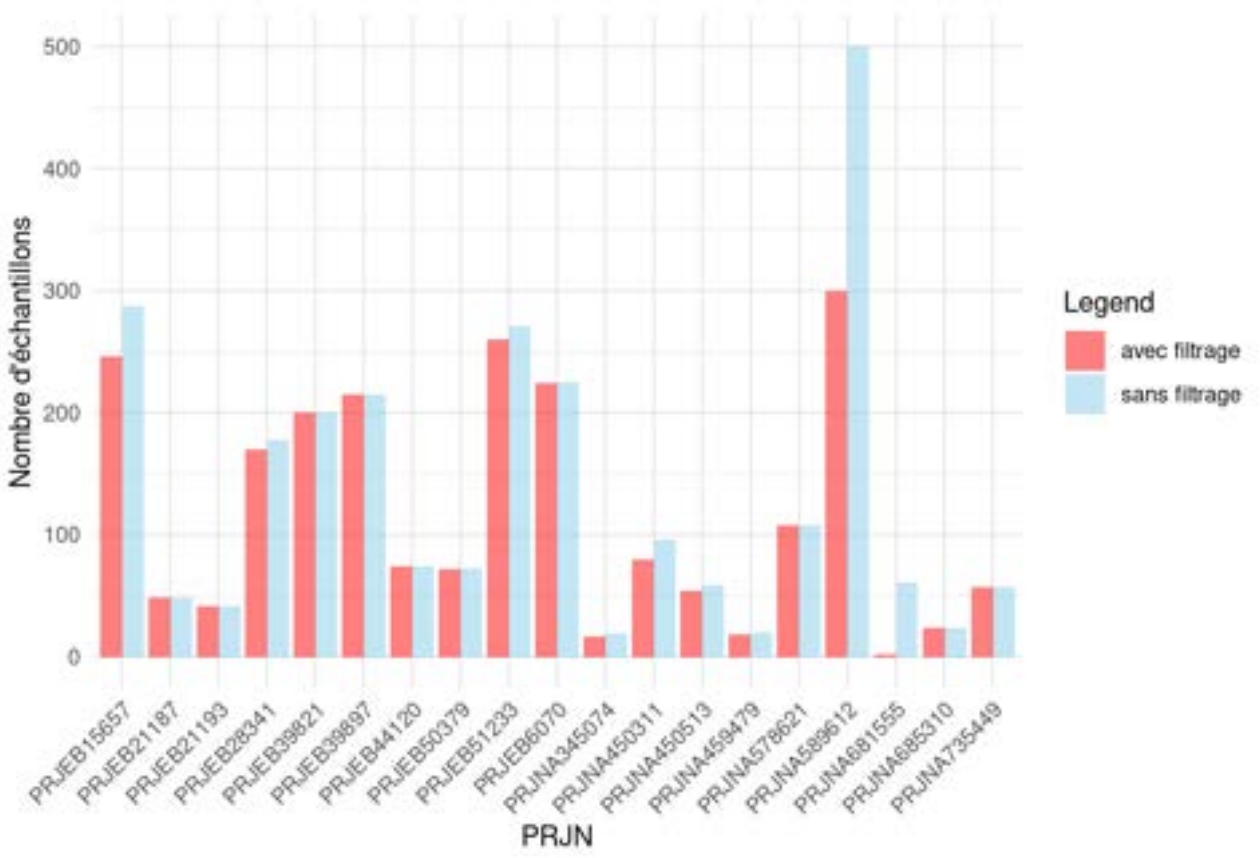


FIGURE 15 – Nombre d'échantillons par projet avant et après filtrage sur la profondeur de séquençage

Par contre, nous pouvons remarquer que le projet PRJNA681555 (Table 1) est fortement impacté par ce filtrage (Figure 15) avec 96,72% des échantillons du projet supprimés. En effet, la profondeur de séquençage du projet sur la bière (biotope **food**) est largement inférieure aux autres projets, nous avons alors décidé de retirer tous les échantillons de ce projet pour la suite des analyses afin de ne pas introduire un effet spécifique à celui-ci.

Ensuite, j'ai développé une approche de sélection des espèces selon des critères de prévalence et d'abondance. Ces deux notions se définissent comme suit :

$$\text{Abondance} = \frac{\text{Nombre de fois où l'espèce est présente dans l'échantillon}}{\text{Nombre total d'espèces}}$$

$$\text{Prévalence} = \frac{\text{Nombre d'échantillons où l'espèce est présente}}{\text{Nombre d'échantillons}}$$

Le but de notre étude étant de regarder les interactions d'espèces de différents biotopes, j'ai défini une grille de valeurs (Figure 17) pour les seuils de prévalence et d'abondance afin de choisir une combinaison adaptée pour l'inférence de réseau. Ces seuils varient entre $1e^{-05}$ et 0.4 pour la prévalence (un seuil à 0.4 signifie qu'une espèce est présente au minimum dans 40% des échantillons du biotope) puis entre 0 et 0.3 pour l'abondance (un seuil à 0.3 indique qu'une espèce est présente au minimum à 30% par rapport aux autres espèces du biotope). Cela donne donc, pour chaque biotope, un total de 56 combinaisons de couples de seuils.
En d'autres termes, pour un biotope, nous avons sélectionné les espèces dont la prévalence est supérieure à un seuil minimal par échantillon et dont l'abondance est supérieure à un seuil minimal dans au moins un échantillon. Ce qui permet la présence d'espèces rares dans certains échantillons.

Enfin, une dernière stratégie communément utilisée consiste à agglomérer les comptages de la table d'abondance à un rang taxonomique supérieur à celui de l'espèce (Figure 1).
Cette table d'abondance est alors de dimension plus réduite et l'affiliation taxonomique est mieux renseignée.

5.1.2 SPIEC EASI

Cette partie se base sur la publication dédiée à l'inférence de réseau via la méthode SPIEC-EASI (*SParse InversE Covariance Estimation for Ecological ASsociation Inference*)[D.Kurtz et al. 2015] et par simplicité de compréhension, le terme espèce sera utilisé plutôt que ASV ou taxon. Cette méthode a été développée spécifiquement pour l'inférence de réseau sur des données d'écologie microbienne et elle est implémentée dans le package *R* {*SpiecEasi*}. L'inférence via SPIEC-EASI se déroule en deux étapes principales :

1. Transformation des données à partir de la matrice de comptage des espèces (*otu_table*).

2. Inférence d'un réseau d'interactions à partir des données transformées.

Contrairement aux réseaux d'association basés sur des corrélations empiriques comme la corrélation de Pearson, SPIEC-EASI vise à inférer un modèle basé sur les dépendances conditionnelles. Deux nœuds sont conditionnellement indépendants s'ils n'apportent aucune information supplémentaire l'un à l'autre, une fois que l'effet de tous les autres nœuds du réseau est pris en compte.

Le réseau est considéré comme un graphe non orienté $\mathcal{G} = \{V, E\}$, avec $V = \{v_1, ..., v_p\}$ représentant l'ensemble des sommets ou nœuds du graphe et $E \subset V \times V$ contenant les couples de noeuds (ASV, espèces, ...) étant en interaction. L'interaction mesure l'association potentielle entre deux espèces.

Pour inférer ce graphe, SPIEC-EASI propose deux approches :

- **MB** (Meinshausen et Bühlmann) : sélection par proche voisin en effectuant une régression multiple pénalisée pour chaque nœud.

- **GLasso** (Graphical Lasso) : estimation de la matrice de précision par maximum de vraisemblance pénalisé.

5.1.2.1 Transformation des données

La transformation utilisée pour l'inférence via SPIEC-EASI est la transformation *centered log-ratio* (CLR). Cette transformation est essentielle car elle permet de gérer la nature compositionnelle des données métagénomiques, où les abondances relatives des espèces sont exprimées en proportion. Les abondances relatives permettent de prendre en compte la différence de profondeur de séquençage entre les échantillons (illustration lors de la raréfaction 4.1.6)).

Soit W la matrice de comptage (**otu_table**, `Figure 4`) transposée avec les n échantillons en ligne et les p espèces en colonne. Cela donne $W \in \mathbb{N}_0^{n \times p}$ avec $w^{(j)} = [w_1^{(j)}, ..., w_p^{(j)}]$ représentant les comptages pour l'échantillon j.

Une première étape consiste à normaliser les comptages par la somme totale des comptages. Ainsi, la matrice X des données compositionnelles est définies par les vecteurs

$$x^{(j)} = [\frac{w_1^{(j)}}{m^{(j)}}, ..., \frac{w_p^{(j)}}{m^{(j)}}]$$

où $x^{(j)}$ représente les abondances relatives des espèces dans l'échantillon j avec $m^{(j)} = \sum_{i=1}^{p} w_i^{(j)}$.
Nous pouvons constater ici que les comptages relatifs des espèces ne peuvent plus être considérés comme indépendants puisque l'espace de chaque échantillon est le simplex unité de dimension p, $\mathbb{S}^p = \{x | x_i > 0, \sum_{i=1}^{p} x_i = 1\}$.

La transformation CLR est ensuite appliquée pour obtenir la matrice Z, où chaque élément $z^{(j)}$ est le logarithme du ratio entre une abondance relative et la moyenne géométrique des abondances relatives de l'échantillon.

$$z^{(j)} = \text{clr}(x^{(j)}) = \left[\log\left(\frac{x_1^{(j)}}{g(x^{(j)})}\right), ..., \log\left(\frac{x_p^{(j)}}{g(x^{(j)})}\right)\right] \quad , \quad g(x^{(j)}) = \left(\prod_{i=1}^{p} x_i^{(j)}\right)^{\frac{1}{p}}$$

5.1.2.2 Approche MB

Soient $Z^i \in \mathbb{R}^n$ la colonne i de Z et $Z^{\neg i} \in \mathbb{R}^{n \times (p-1)}$ les autres colonnes de Z.
Pour chaque noeud $v_i \in V$, on résout alors le problème suivant :

$$\hat{\beta}^{i,\lambda} = \underset{\beta \in \mathbb{R}^{p-1}}{\text{argmin}} \left(\frac{1}{n}||Z^i - Z^{\neg i}\beta||^2 + \lambda||\beta||_1\right)$$

avec $||.||_1$ correspond à la norme 1 (pour $a \in \mathbb{R}^n$, $||a||_1 = \sum_{i=1}^{n} |a_i|$) et $\lambda \geq 0$ est un scalaire de pénalisation. Les estimations de β sont obtenues en réalisant p régressions linéaires régularisées par une approche Lasso.
Par la suite, un voisinage local de v_i peut être construit $N_i^\lambda = \left\{j \in \{1, ..., p\} \backslash i \quad | \quad \hat{\beta}^{i,\lambda} \neq 0\right\}$. C'est en fonction de l'appartenance à l'intersection ou l'union des voisinages N_i^λ et N_j^λ qu'une arête est construite entre les noeuds v_i et v_j avec le poids de l'arête défini par la moyenne des β correspondants.

Le choix de λ est déterminé ensuite de telle sorte à contrôler la sparsité en utilisant la méthode StARS (voir 5.1.4).

5.1.2.3 Approche GLasso

L'estimation de la matrice de précision (inverse de la matrice de covariance) à partir des données transformées CLR se fait via le problème d'optimisation suivant :

$$\hat{\Omega} = \underset{\Omega \in SPD}{\text{argmin}} \left(-\log(\det(\Omega) + \text{tr}(\Omega\hat{\Gamma}) + \lambda||\Omega||_1\right)$$

Avec $\hat{\Gamma}$ la covariance empirique de la matrice Z, $||.||_1$ correspond à la norme 1 et $\lambda \geq 0$ est un scalaire de pénalisation. De plus, l'ensemble SDP correspond aux matrices symétriques définies positives.

Ainsi, les valeurs non nulles et hors diagonale de $\hat{\Omega}$ définissent les arêtes (interactions) et leurs poids associés.

Comme pour l'approche MB, le choix de λ est déterminé de telle sorte à contrôler la sparsité en utilisant la méthode StARS (voir 5.1.4).

5.1.3 PLN models

Cette partie concerne l'inférence de réseaux par le modèle Poisson Lognormal (PLN) [Chiquet, Mariadassou et Robin 2021] et toujours par simplicité de compréhension, le terme espèce sera utilisé plutôt que ASV ou taxon. Le modèle PLN est spécialement conçu pour l'analyse de données de comptage, ce qui en fait un outil particulièrement adapté aux données d'écologie microbienne où les tables d'abondances représentent le nombre de fois qu'un ASV ou une espèce est comptée dans un échantillon. L'inférence du modèle PLN est implémentée dans le package *R* {PLNmodels} et suit une démarche en deux étapes principales :

1. Modélisation des données par un modèle PLN à partir de la matrice de comptage des espèces

2. Inférence d'un réseau d'interactions à partir du modèle ajusté

Le réseau obtenu est un graphe non orienté $\mathcal{G} = \{V, E\}$, où $V = \{v_1, ..., v_p\}$ représentant l'ensemble des sommets ou nœuds du graphe et $E \subset V \times V$ contenant les couples de noeuds (ASV, espèces, ...) en interaction.

L'une des forces du modèle PLN réside dans sa capacité à prendre en compte l'hétérogénéité des données de comptage, tout en inférant un réseau d'interactions basé sur des dépendances conditionnelles, similaire à ce que propose SPIEC-EASI. Contrairement à SPIEC-EASI, le modèle PLN à l'avantage de permettre d'intégrer des covariables provenant des métadonnées qui peuvent avoir un effet sur les comptages observés.

5.1.3.1 Modèle

Soit W la matrice de comptage des espèces (**otu_table**, `Figure 4`) transposée avec les n échantillons en ligne et les p espèces en colonne. Cela donne $W \in \mathbb{N}_0^{n \times p}$ avec $w^{(i)} = [w_1^{(i)}, ..., w_p^{(i)}]$ représentant les comptages pour l'échantillon i.

Le modèle PLN-Network se définit pour chaque échantillon i avec un vecteur latent Z_i gaussien :

$$\text{Variable latente : } Z_i \sim \mathcal{N}(\mu, \Omega^{-1}) \quad \text{avec} \quad ||\Omega||_{1,0} < c$$

$$\text{Observation : } W_{ij}|Z_{ij} \sim \mathcal{P}(\exp(Z_{ij}))$$

Avec μ qui correspond à un effet principal, Ω la matrice de précision (inverse de la matrice de covariance) qui décrit la structure de dépendance entre les p espèces, c une constante réelle positive qui représente la contrainte de sparsité mise sur la matrice de précision et $||\Omega||_{1,0}$ correspond à la somme des valeurs absolue des termes hors de la diagonale.
Les variables Z_i sont supposées indépendantes et donc les W_{ij} sont conditionnellement indépendants par rapport aux variables Z_i.

La quantité fixe o_i, appelée *Offset*, est utilisée pour ajuster les différences de profondeur de séquençage entre les échantillons (illustration lors de la raréfaction 4.1.6).

Dans ce modèle, l'*Offset* est défini par la méthode TSS (*Total Sum Scaling*) qui consiste à transformer la matrice d'abondance en une matrice d'abondance relative où pour un échantillon i on a :

$$[\frac{w_1^{(i)}}{m^{(i)}}, ..., \frac{w_p^{(i)}}{m^{(i)}}] \quad \text{avec} \quad m^{(i)} = \sum_{j=1}^{p} w_j^{(i)}$$

De plus, lorsque des variables introduisent une structure dépendante sur les comptages observés (par exemple la profondeur de sequançage, le biotope, ...), celles-ci peuvent être intégrées dans le modèle via des covariables. On définit alors la variable latente :

$$\text{Variable latente : } Z_i \sim \mathcal{N}(\mu_i, \Omega^{-1}) \quad \text{avec} \quad ||\Omega||_{1,0} < c$$

L'effet fixe se décompose alors en $\mu_i = o_i + x_i'\theta_i$ où o_i est l'*Offset*, $x_i \in \mathbb{R}^d$ est le vecteur des covariables pour l'échantillon i et $\theta_i \in \mathbb{R}^d$ le vecteur des coefficients de régression associés aux d covariables. Ces vecteurs de régression forment la matrice Θ de dimension $d \times p$.

Comme pour l'approche GLasso, les valeurs de la matrice de précision hors de la diagonale définissent les arêtes (interactions) et leurs poids associés.
Et le paramètre de pénalisation est déterminé en utilisant la méthode StARS (voir 5.1.4) de telle sorte à contrôler la sparsité de la matrice de précision .

5.1.4 StARS selection

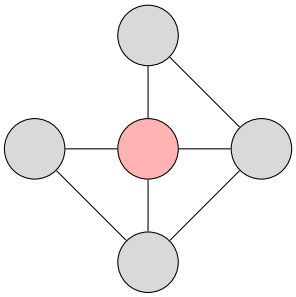
La sélection StARS (*Stability Approach to Regularization Selection*) [Liu, Roeder et Wasserman 2010] est une méthode conçue pour déterminer le paramètre de pénalisation/régularisation λ dans les modèles d'inférence de réseau afin de contrôler la sparsité, particulièrement dans les contextes de haute dimension. Les méthodes classiques comme la validation croisée K-fold, le critère d'information d'Akaike (AIC) et le critère d'information bayésien (BIC) sont efficaces pour des problèmes de petite dimension, mais elles se révèlent inadéquates pour les problèmes de grande dimension. Par exemple, en grande dimension, la validation croisée peut entraîner un surajustement des données, tandis que les critères AIC et BIC tendent vers des valeurs infinies, rendant difficile la sélection de modèles pertinents.

C'est dans ce contexte que la méthode StARS a été développée. L'idée principale de StARS est de trouver le paramètre de régularisation optimal qui permet de créer un graphe à la fois sparse et stable. Une grande valeur de λ tend à construire un réseau sparse alors qu'une faible valeur de λ correspond à des réseaux plus denses.

Le fonctionnement de StARS repose sur le sous-échantillonnage de 80% des échantillons (valeur par défaut). Contrairement à des techniques comme la validation croisée K-fold, où les échantillons sont disjoints, les sous-échantillons de StARS peuvent se chevaucher. Pour chaque sous-échantillon, un graphe dépendant de λ est construit. Le but est de sélectionner un paramètre λ qui permet d'obtenir une stabilité supérieure à un seuil pré-déterminé (par exemple 90%). La stabilité est mesurée par la variabilité globale du réseau, en lien avec la fréquence de sélection des arêtes des réseaux obtenus à chaque sous-échantillonnage.

5.1.5 Mesures de robustesse du réseau

Pour évaluer la robustesse et la stabilité d'un réseau, je me suis particulièrement intéressé à la distribution des degrés et à la centralité d'intermédiarité (*betweenness centrality*). La méthode d'évaluation que j'ai utilisée consiste à supprimer progressivement les nœuds les plus "centraux" du réseau et à chaque étape de suppression regarder l'impact que cela a eu sur la structure du réseau. Ces nœuds sont déterminés selon leur centralité d'intermédiarité qui correspond au nombre de fois qu'un nœud est sur le chemin le plus court entre deux autres nœuds quelconque du graphe.



Le noeud en rouge est celui qui se trouve le plus souvent sur le chemin le plus court entre deux autres noeuds (Figure 16).

FIGURE 16 – Illustration de la centralité d'intermédiarité

Ainsi, en supposant que les chemins les plus courts correspondent à des relations d'interaction rapprochées entre deux espèces, un nœud possédant une grande centralité d'intermédiarité a une grande influence sur les interactions du réseau.
Donc, à chaque suppression de nœuds centraux, l'indice de connectivité naturelle (*natural connectivity*) [WU et al. 2010] a été calculé pour représenter la robustesse du réseau. Il est basé sur les valeurs propres de la matrice d'adjacence (matrice binaire contenant des 1 lorsque deux nœuds sont en interaction et 0 sinon). Ces valeurs propres caractérisent des aspects importants de la robustesse globale du graphe et l'indice peut donc permettre d'évaluer la structure du réseau et sa résistance aux perturbations.
Soit N le nombre de noeuds du réseau et λ_i, i allant de 1 à N, les valeurs propres de la matrice d'adjacence. L'indice de connectivité naturelle est de la forme :

$$\mathrm{NC} = \ln\left(\frac{1}{N}\sum_{i=1}^{N} e^{\lambda_i}\right)$$

J'ai donc développé deux fonctions R. La première calcule l'indice de connectivité naturelle du réseau (NC). La seconde construit à chaque étape de suppression d'un noeud central, une nouvelle matrice d'adjacence et calcule un nouvel indice de connectivité naturelle.
Par la suite, nous pouvons visualiser l'impact de la suppression successive des noeuds centraux sous la forme d'une courbe représentant la connectivité naturelle en fonction de la proportion de nœuds centraux enlevés. Et en complément, nous proposons un histogramme qui représente les fréquences des différences $NC_t - NC_{t-1}$, t allant de 1 à T nombre de suppressions.

5.2 Résultats

Dans cette partie, tous les réseaux ont été construits de manière interactive, permettant ainsi d'afficher pour chaque nœud (espèce ou classe) toutes les informations taxonomiques affiliées, ainsi que les valeurs de prévalence et d'abondance dans le ou les biotopes concernés. Cette interactivité est intéressante pour étudier les réseaux et rendre accessible les informations à tous les acteurs dans un contexte multi-disciplinaire.
Vous pouvez retrouver les réseaux interactifs présentés dans les sections 7 et 9 du blog [Poupelin 2024], offrant une exploration détaillée et dynamique sur chaque noeuds.

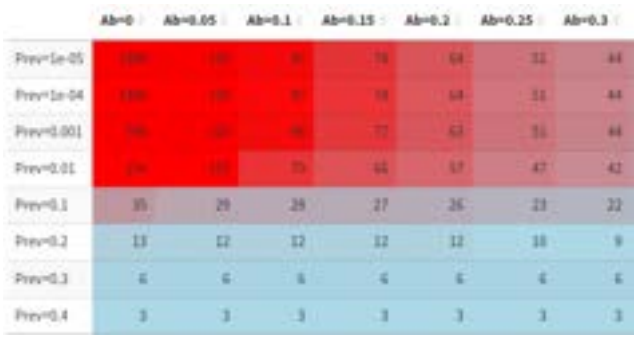
5.2.1 Création de jeux de données pour l'inférence de réseau

L'inférence de réseau de grande dimension représente un défi et une complexité à prendre en considération. En effet, au niveau espèce les données sont de grande dimension (3448 espèces $\times$ 2210 échantillons) et il n'est pas raisonnable d'inférer un réseau directement sur cette matrice d'abondance (faisabilité et temps de calcul). J'ai donc réfléchi à plusieurs approches permettant de créer des jeux de données pertinents pour l'inférence de réseau.

Tous les réseaux ont été construits à partir du jeu de données dont la profondeur de séquençage des échantillons est supérieure à 10000 (voir 4.1.1).

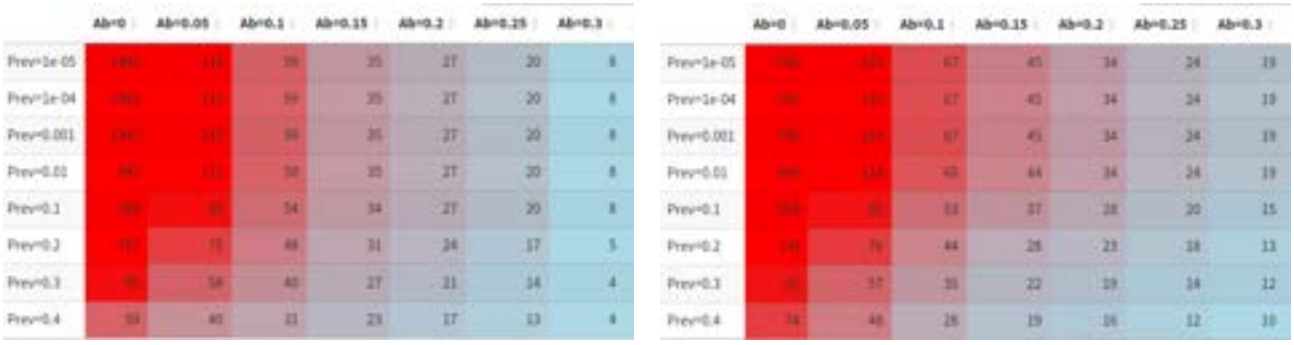
5.2.1.1 Jeu de données pour l'inférence de réseau avec SPIEC EASI

Dans l'optique de rester au niveau de la résolution espèce, j'ai décidé de filtrer les espèces selon leur abondance et prévalence au sein de leur biotope à partir des grilles (`Figure 17`) construites à cet effet. Nous pouvons y retrouver le nombre d'espèces selon les différents couples de seuils. Par exemple, pour un seuil de prévalence à $1e^{-05}$ ($Prev$=$1e^{-05}$) et aucun seuil minimal d'abondance (Ab=0) nous obtenons 1300 espèces dans le biotope **food**.
Aussi, concernant les 1246 espèces partagées par au moins deux biotopes (`Figure 7`), j'ai choisi de considérer qu'une espèce appartenait au biotope dans lequel elle était la plus prévalente et elle sera donc filtrée en fonction de ce biotope principal.



food

cow



digester human gut

FIGURE 17 – Nombre d'espèces différentes selon les seuils de prévalence et d'abondance

Nous constatons que le nombre d'espèces présentes dans les biotopes **cow**, **digester** et **human gut** diminue rapidement lorsque le critère d'abondance augmente (Figure 17). Ceci suggère que peu d'espèces sont très abondantes dans ces biotopes (par exemple, 9 espèces sont abondantes à plus de 25% avec un prévalence minimale de $1e^{-05}$ pour le biotope **cow**, Figure 17). Par contre, beaucoup d'espèces sont prévalentes dans ces biotopes (par exemple, 69 espèces sont présentes dans plus de 40% des échantillons du biotope **cow**, Figure 17). J'ai donc choisi pour ces trois biotopes de favoriser des couples de seuils où la prévalence était plus élevée afin de ne garder qu'une cinquantaine d'espèces.
Puis la même démarche a été suivie pour le biotope **food** mais cette fois-ci les espèces semblent majoritairement être peu prévalentes mais très abondantes (par exemple, 44 espèces sont abondantes à plus de 30% avec un prévalence minimale de $1e^{-05}$ pour le biotope **food**, Figure 17). À noter que, comme pour les analyses d'α-diversité, la faible prévalence peut s'expliquer par le nombre important de sources différentes des échantillons liés à ce biotope (voir 4.2.1 Figure 12).

Cela m'a donc amené à choisir des couples de seuils adaptés à chaque biotope. Plusieurs jeux de données filtrés ont alors été construits afin de tester la démarche. Mais ici, un seul de ces jeux de données sera utilisé contenant un nombre d'espèces "raisonnable" pour inférer un réseau d'interactions tout en préservant au maximum les caractéristiques des échantillons de chaque biotope en terme d'abondance et prévalence (Table 4).

TABLE 4 – Seuils de filtrage appliqués sur le jeu de données

Biotope	Seuil de prévalence	Seuil d'abondance	Nombre d'espèces
Food	0.001	0.25	51
Cow	0.1	0.05	41
Digester	0.3	0.05	58
Human gut	0.4	0.05	48

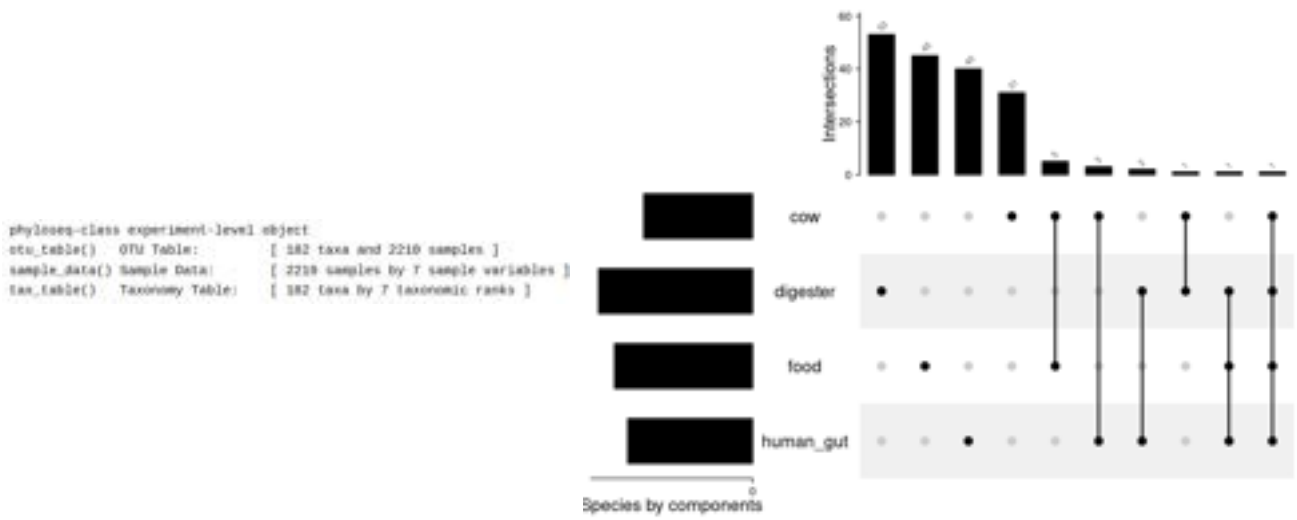


FIGURE 18 – Objet phyloseq filtré pour les réseaux SPIEC-EASI et l'Upset plot associé

Cette stratégie de filtrage a permis d'obtenir :

- un jeu de données contenant 182 espèces, les plus représentatives d'un biotope en termes de prévalence et d'abondance.

- un jeu de données où des espèces partagées entre plusieurs biotopes ($n = 13$) sont présentes avec possiblement des espèces très représentées (abondante et prévalente) dans un biotope mais plus rare dans un autre (Figure 18).

Cependant, cette sélection ne permet pas de se concentrer sur des espèces transversales au biotope (les espèces communes à plusieurs biotopes). En réponse à cela, un autre jeu de données a été construit.

5.2.1.2 Jeu de données pour l'inférence de réseau avec PLN

Pour cette stratégie, j'ai décidé d'agglomérer les comptages à un rang taxonomique supérieur. En effet, on constate que 1139 espèces sont inconnues parmi les 3448 (Figure 19).


Show 10 entries Search:

	Species	Freq
1247	unknown species	1139
792	Multi-affiliation	782
754	metagenome	128
572	gut metagenome	17
371	Clostridium sp.	14
92	anaerobic digester metagenome	13
147	Bacillus sp.	9
516	Eubacterium sp.	7
535	Flavobacterium sp.	5
1068	Ruminococcus sp.	5

Showing 1 to 10 of 1,279 entries Previous 1 2 3 4 5 ... 128 Next

FIGURE 19 – Liste des espèces et leur fréquence dans dans le jeu de données aggloméré par espèces

Lorsque les comptages sont agglomérés au niveau ordre ($n = 317$) ou classe ($n = 138$), l'affiliation taxonomique est plus complète mais nous perdons en précision sur la nature des organismes présents. Au niveau ordre on constate la présence de 54 *unknown* et 10 *multi affiliation* et au niveau classe nous avons 12 *unknown* et 3 *multi affiliation* (Figure 20).


Show 10 entries Search:

	Order	Freq
245	unknown order	54
162	Multi-affiliation	10
1	0319-6G20	1
2	1-20	1
3	32-520	1
4	Abditibacteriales	1
5	Absconditabacteriales (SR1)	1
6	Acetobacterales	1
7	Acholeplasmatales	1
8	Acidaminococcales	1

Showing 1 to 10 of 255 entries Previous 1 2 3 4 5 ... 26 Next

Show 10 entries Search:

	Class	Freq
120	unknown class	12
77	Multi-affiliation	3
1	028H05-P-BN-P5	1
2	Abditibacteria	1
3	ABY1	1
4	Acidimicrobiia	1
5	Acidobacteriae	1
6	Actinobacteria	1
7	Alphaproteobacteria	1
8	Aminicenantia	1

Showing 1 to 10 of 125 entries Previous 1 2 3 4 5 ... 13 Next

FIGURE 20 – Liste des ordres et classes ainsi que leur fréquence

Puis, afin de proposer un jeu de données principalement centré sur la transversatilité entre les biotopes, j'ai sélectionné les taxa au rang taxonomique ordre et classe qui sont communs à au moins deux biotopes.

Par conséquent, au rang ordre ($n = 154$) on constate la présence de 10 *unknown* et 4 *multi affiliation* et au rang classe ($n = 62$) nous avons 3 *unknown* et 2 *multi affiliation* (Figure 21).

	Order	Freq
137	unknown order	10
94	Multi-affiliation	4
1	0319-6G20	1
2	Absconditabacteriales (SR1)	1
3	Acetobacterales	1
4	Acholeplasmatales	1
5	Acidaminococcales	1
6	Acidobacteriales	1
7	Actinomarinales	1
8	Actinomycetales	1

Showing 1 to 10 of 142 entries

	Class	Freq
57	unknown class	3
40	Multi-affiliation	2
1	ABY1	1
2	Acidimicrobiia	1
3	Acidobacteriae	1
4	Actinobacteria	1
5	Alphaproteobacteria	1
6	Anaerolineae	1
7	Babeliae	1
8	Bacilli	1

Showing 1 to 10 of 59 entries

FIGURE 21 – Liste des ordres et classes communs à au moins deux biotopes ainsi que leur fréquence

Finalement, le jeu de données contenant les 62 classes communes à au moins deux biotopes a été retenu (Figure 22). Cela permet d'avoir un jeu de données dont la taille est adaptée à l'inférence de réseau avec la méthode PLN.

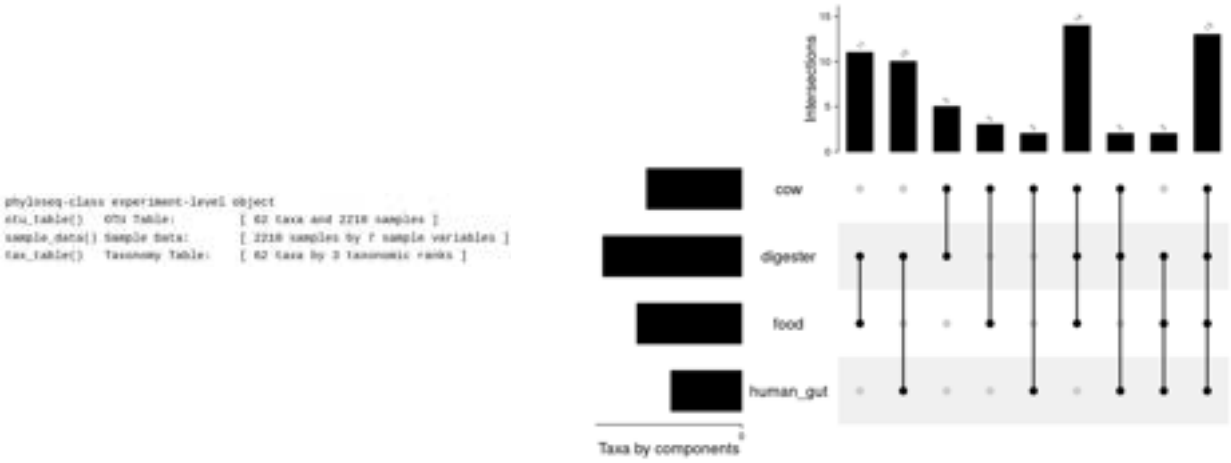


FIGURE 22 – Objet phyloseq filtré pour les réseaux PLN et l'Upset plot associé

En conclusion de cette section, nous avons créé deux jeux de données associés aux deux objets phyloseq distincts (Figure 18 et Figure 22) pour l'inféence de réseaux avec les méthodes *SPIEC-EASI* (au rang espèce) et *PLN* (au rang classe). Cette méthodologie a pour objectif d'obtenir des résultats pertinents et différenciés, apportants ainsi une meilleure compréhension des interactions microbiennes au sein des biotopes étudiés.

5.2.2 SPIEC EASI

Les premières approches de réseau que j'ai testé étaient Glasso et MB via le package {*SpiecEasi*} dans *R*. L'idée était ici, en utilisant la capacité d'estimation sparse de *SPIEC-EASI*, de créer des réseaux avec un nombre de nœuds compris entre 100 et 200 afin de voir si des structures se dégageaient. J'ai donc inféré les réseaux sur l'objet phyloseq, contenant les 182 espèces, présenté précédemment (voir 5.2.1.1 et Figure 18).

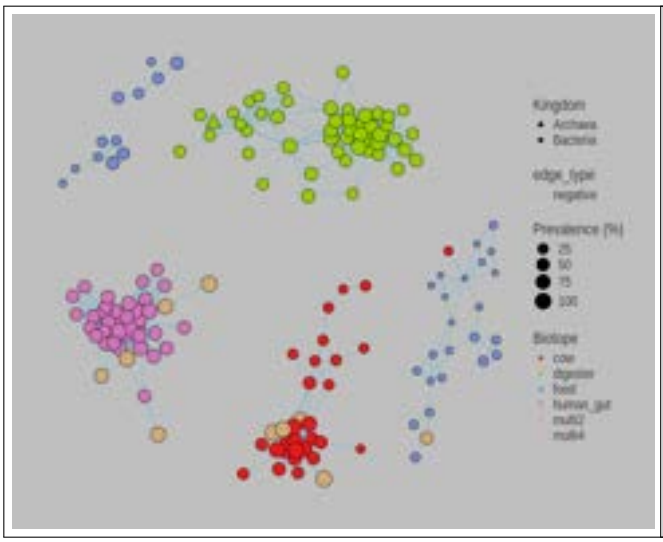


FIGURE 23 – Réseau d'interaction via GLasso avec 182 espèces

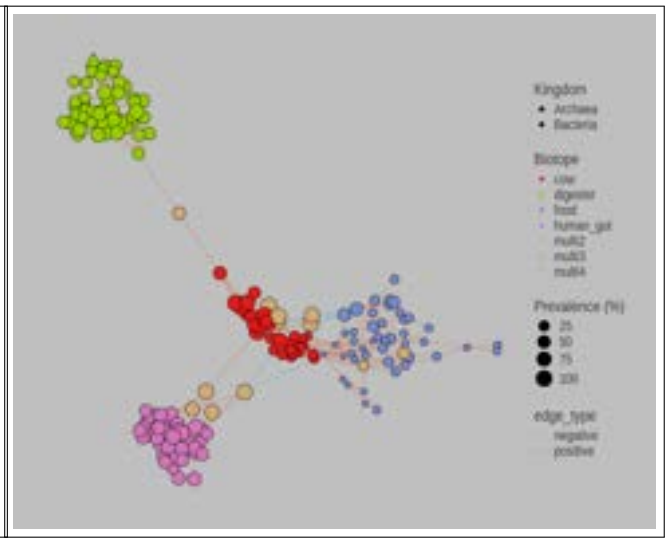


FIGURE 24 – Réseau d'interaction via MB avec 182 espèces

Nous constatons une structuration par biotope (Figure 23 et Figure 24) ce qui signifie que les espèces appartenant à un même biotope sont en interaction entre elles. Les espèces qui appartiennent à plusieurs biotopes (multi2, multi3 et multi4 selon qu'elles soient communes à 2, 3 ou 4 biotopes, en orange sur les figures) sont principalement en interaction avec d'autres espèces du même biotope avec l'approche GLasso (Figure 23) et des espèces d'un biotope différent avec l'approche MB (Figure 24).
Par exemple, pour une des espèces située entre **human gut** et **cow** (Figure 24), en regardant les informations sur cette espèce (blog [POUPELIN 2024] section 7), nous constatons qu'elle est majoritairement prévalente et abondante pour le biotope **cow** (abondance à 1.88% et prévalence à 78.51%) mais qu'elle est également présente au sein du biotope **human gut** (abondance à 0.04% et prévalence à 0.47%). Cette espèce est renseignée en *unknown* mais nous pouvons tout de même savoir qu'elle fait partie de la famille des Sphingomonadaceae.

Il est aussi intéressant de remarquer que les espèces du biotope **cow** occupent une place centrale dans le réseau inféré via la méthode MB (Figure 24) ce qui peut amener à une envie d'analyser plus en profondeur les espèces constituant le biotope **cow**.
De plus, nous pouvons remarquer la formation de deux groupes distincts au sein du biotope **food** (Figure 23). Cela pourrait s'expliquer par la diversité des sources alimentaires présentes dans ce biotope, qui engendre une variation importante au sein même du groupe. Puis, les interactions entre les biotopes **cow** et **food** (Figure 23 et surtout Figure 24) semblent illustrer une proximité entre ces deux biotopes qui n'est pas incohérente étant donné que beaucoup d'échantillons du biotope **food** proviennent de lait de vache (Table 1).

Par la suite, différents critères sont utilisés pour décrire les réseaux. Dans un premier temps, on s'intéresse au degré associé à chacun des nœuds. Le degré d'un nœud représente le nombre d'arêtes qui le relient à d'autres nœuds.

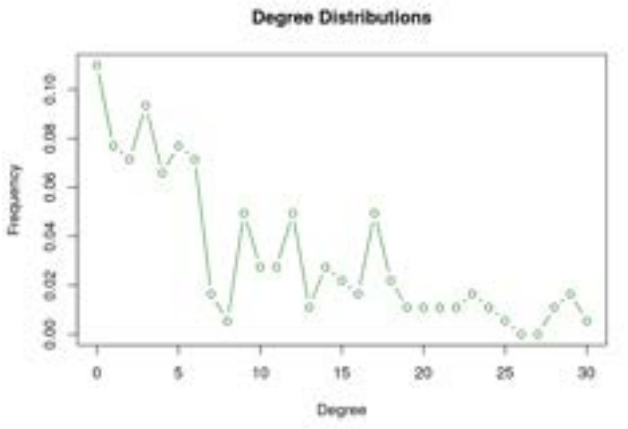


FIGURE 25 – Distribution des degrés des noeuds du réseau via GLasso

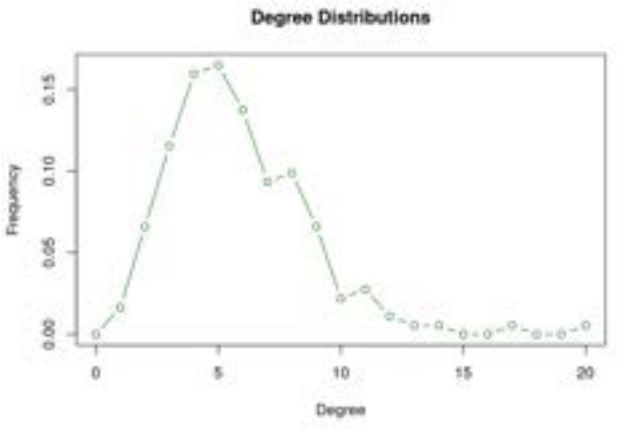
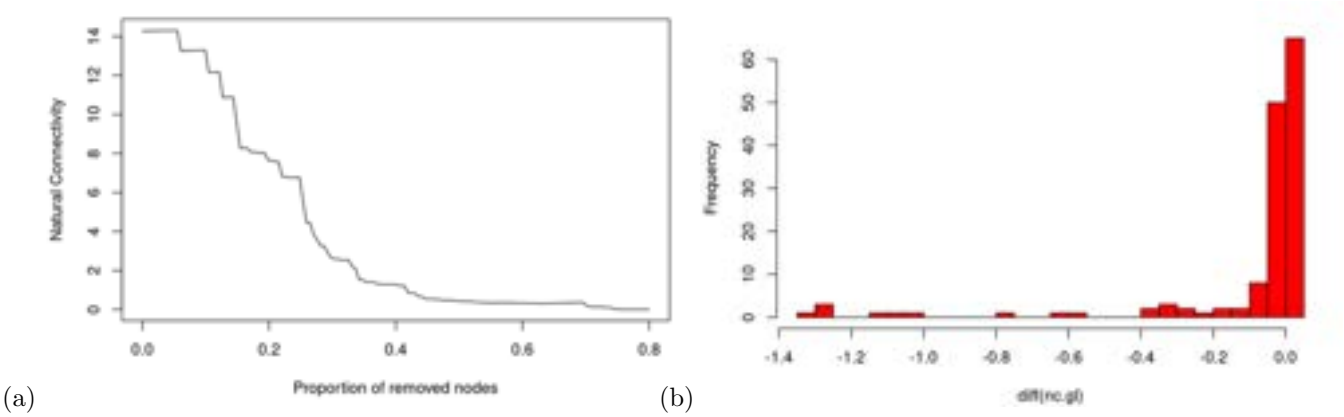


FIGURE 26 – Distribution des degrés des noeuds du réseau via MB

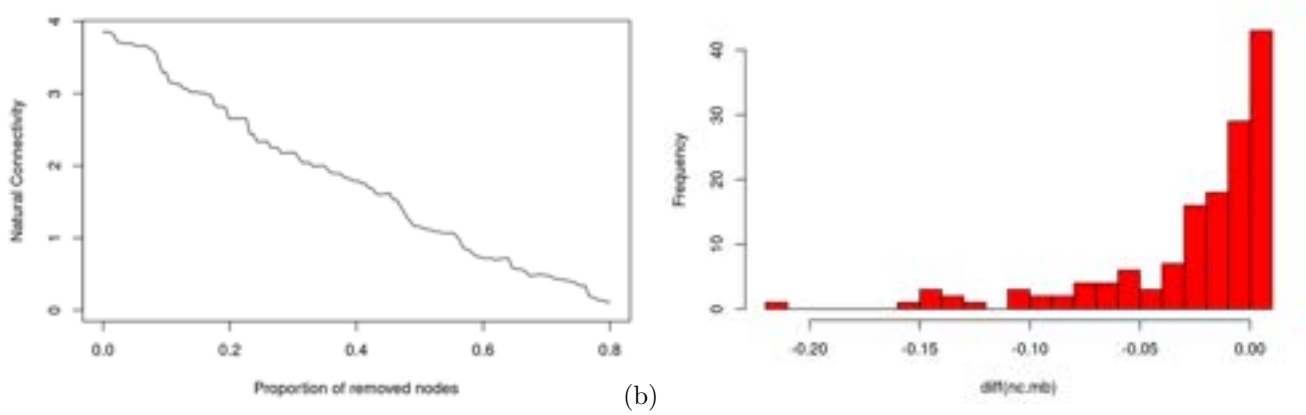
De cette manière, on voit bien une différence entre le réseau d'interaction construit avec l'approche GLasso et celui construit avec l'approche MB (`Figure 25` et `Figure 26`). Pour l'approche GLasso, les espèces sont souvent liées majoritairement à un petit nombre d'espèces (1 à 5) alors que pour l'approche MB elles sont majoritairement liées à 6 ou 7 autres espèces. Cela témoigne donc d'une plus forte densité du réseau construit par l'approche MB. Par contre, dans les deux approches, peu d'espèces sont fortements liées aux autres.

De plus, nous pouvons également regarder la robustesse des réseaux construits lorsque l'on retire successivement des noeuds centraux (voir 5.1.5) correspondants aux espèces qui intéragissent le plus avec d'autres.



(a) (b)

FIGURE 27 – Connectivité du réseau via GLasso



(a) (b)

FIGURE 28 – Connectivité du réseau via MB

Avec une proportion de 40% de nœuds supprimés, le réseau basé sur la méthode GLasso, malgré de fortes valeurs de connectivité naturelle (NC), est beaucoup plus fragile à la perte des noeuds centraux que le réseau basé sur la méthode MB (Figure 27 (a) et Figure 28 (a)).
En complément, on regarde aussi la distribution des différences de connectivité à chaque étape de suppression d'un noeud central (Figure 27 (b) et Figure 28 (b)). Les histogrammes montrent les fréquences des différences $NC_t - NC_{t-1}$ avec t allant de 1 à T suppressions. Alors, une valeur de différence proche de zéro indique que la suppression du noeud a peu d'influence sur la connectivité du réseau.
Ces figures, complémentaires, semblent indiquer que les deux réseaux sont relativelent robustes. Cependant, le réseau basé sur la méthode MB est le plus stable à la suppression de nœuds centraux avec des valeurs de différences moins étendues et une fréquence plus élevée de différence proche de zéro.
Ces mesures liées à la connectivité sont une aide à la décision pour le choix du modèle (GLasso, MB) et pour l'interprétation des interactions entre les espèces.

En conclusion, la méthode SPIEC EASI nous a permis dans un premier temps de visulaiser la cohérence des données issues du projet Open16S où les espèces appartenant au même biotope ont tendance à plus interagir entre elles. Puis nous avons montré que l'approche MB permet de mettre en évidence des interactions entre les espèces appartenant à des biotopes différents (par exemple des espèces des familles Rikenellaceae, Yersiniaceae, Streptococcaceae ou encore Lachnospiraceae).

5.2.3 Les réseaux PLN

La seconde approche d'inférence de réseau que j'ai utilisée est celle construite à partir du modèle PLN (Poisson LogNormal) via le package *R* {*PLNmodels*}. Les réseaux ont été construits à partir du jeu de données transversal aux biotopes contenant 62 classes (agglomération des comptages au niveau taxonomique classe, voir 5.2.1.2 et Figure 22). C'est à dire que les 62 classes du réseau sont toutes communes à au moins deux biotopes. Il est important de noter que nous parlerons donc maintenant de classe et non plus d'espèce. Le niveau classe étant un autre niveau plus large d'affiliation taxonomique (Figure 1).

Et donc, bien que toutes les classes soient partagées par les différents biotopes, pour la simplicité de lecture j'ai décidé de les colorier en fonction du biotope où elles sont le plus prévalente.
L'idée ici est d'utiliser la possibilité qu'offre PLN d'inclure des covariables dans le modèle afin de visualiser les interactions conditionnellement au biotope.
En effet, nous avons constaté lors de l'inférence de réseau avec la méthode SPIEC EASI que les espèces appartenant à un même biotope interagissent davantage entre elles. Cependant, nous souhaitons aussi identifier des interactions entre des espèces de biotopes différents.
J'ai donc estimé deux modèles statistiques différents :

$$Abundance \sim 1 + Offset \tag{1}$$

$$Abundance \sim 0 + Biotope + Offset \tag{2}$$

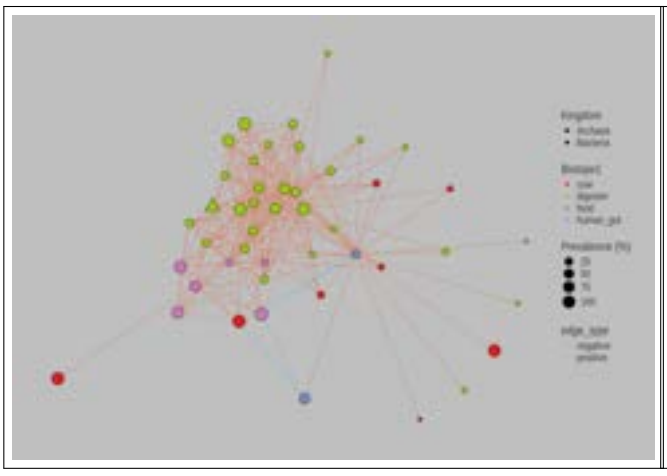


FIGURE 29 – Réseau d'interaction via PLN avec 46 classes pour le modèle (1)

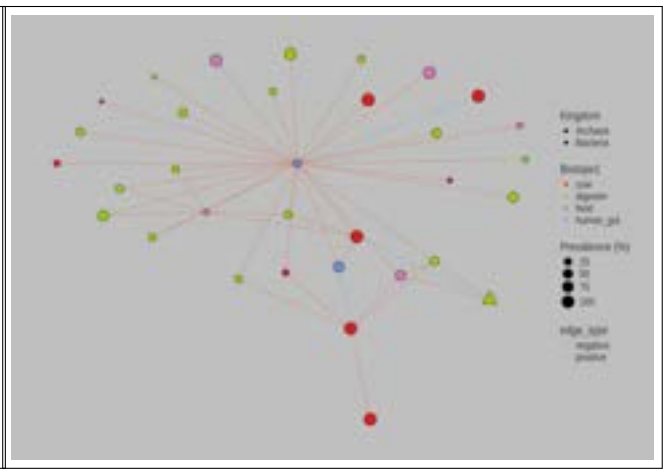


FIGURE 30 – Réseau d'interaction via PLN avec 33 classes pour le modèle (2)

Les résultats obtenus avec les réseaux PLN montrent des particularités intéressantes sur les interactions microbiennes en fonction du biotope et sont très distincts selon que l'on prenne en compte l'effet du biotope ou non dans le modèle.

Aussi, parmi les 62 classes présentes dans le jeu de données, les réseaux suggèrent que de nombreuses classes ne semblent pas interagir (Figure 29 et Figure 30) puisqu'ils sont constitués respectivement de 46 et 33 noeuds. Il est également intéressant de noter que le nombre d'interactions augmente lorsque le biotope n'est pas inclus dans le modèle (surtout dans le biotope **digester**, Figure 29).
Et, selon les résultats du réseau avec le modèle (2) (Figure 30), les interactions entre les classes sont peu nombreuses, ce qui renforce l'idée que le biotope joue un rôle crucial dans la structuration des réseaux microbiens.

Un autre point d'intérêt est le nœud "central" visible dans le réseau avec le mdoèle (2) (Figure 30), qui malheureusement reste non renseigné au niveau de la classe et du phylum (nous savons donc juste qu'il s'agit d'une bactérie). Cette observation souligne une fois de plus les limitations liées aux informations manquantes. Mais, le résultat reste tout de même prometteur concernant l'existence d'une bactérie jouant un rôle central dans les interactions entre biotopes.

Pour décrire les réseaux construits avec le modèle PLN, nous pouvons nous intéresser au degré associé à chacun des noeuds.

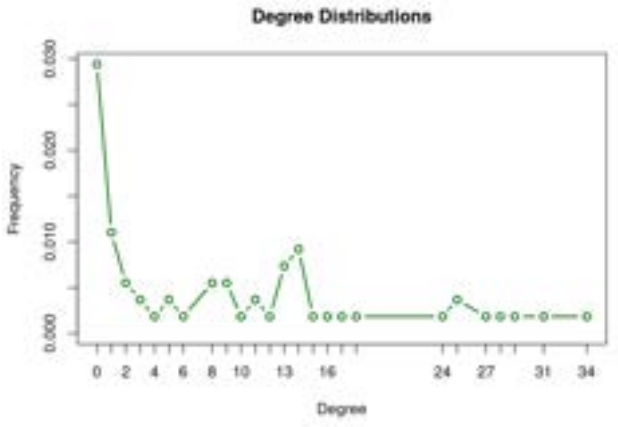


FIGURE 31 – Distribution des degrés des noeuds du réseau pour le modèle (1)

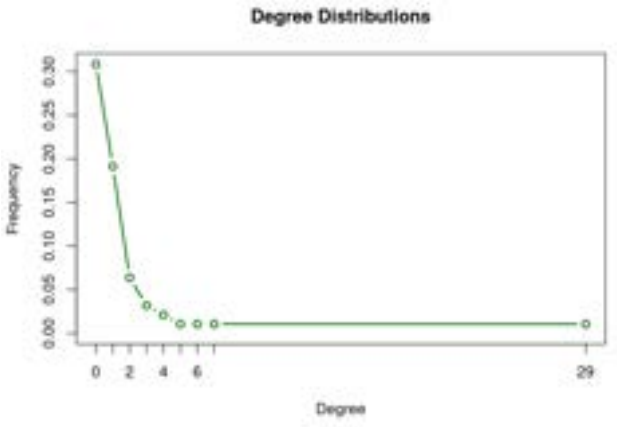


FIGURE 32 – Distribution des degrés des noeuds du réseau pour le modèle (2)

Nous pouvons constater qu'il y a majoritairement des interactions de degrés faibles (entre 1 et 3) pour les deux réseaux (Figure 31 et Figure 32). Lorsque le biotope n'est pas inclus dans le modèle, le réseau possède quelques noeuds de degré supérieur (entre 13 et 16, Figure 31) correspondant probablement aux classes du biotope **digester**.

Nous pouvons ensuite illustrer la stabilité avec les graphiques de stabilité disponible via le package *R* {*PLNmodels*} et calculés lors de la procédure de sélection de modèle StARS.

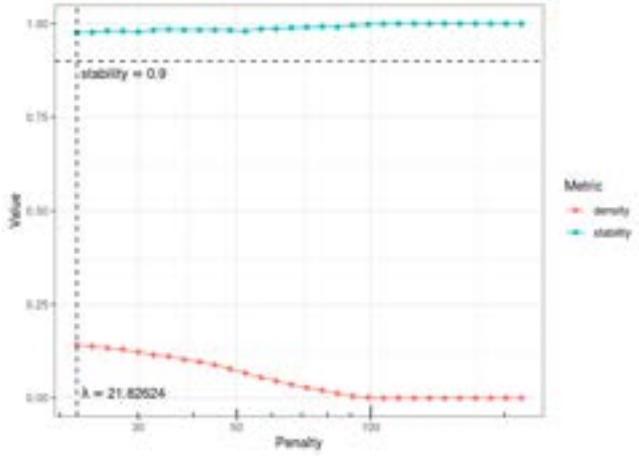


FIGURE 33 – Stabilité et densité du réseau pour le modèle (1)

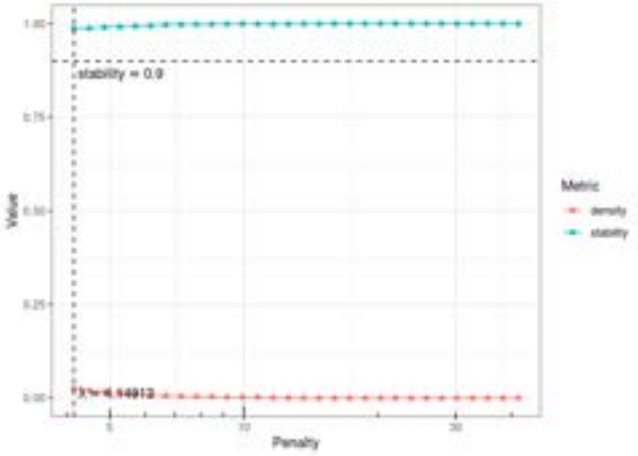


FIGURE 34 – Stabilité et densité du réseau pour le modèle (2)

Nous pouvons constater une très bonne stabilité des réseaux, ce qui indique que les interactions identifiées sont robustes et fiables pour identifier les relations inter-classes (`Figure 33` et `Figure 34`).

De plus, les réseaux obtenus ne sont pas trop denses, ce qui suggère un bon équilibre entre la détection d'interactions pertinentes et la sparsité du réseau. Et nous pouvons également retrouver que le réseau pour le modèle (1) est plus dense que celui pour le modèle (2) (`Figure 33` et `Figure 34`).

Cependant, il est important de noter que si l'on réduit la pénalisation, la densité des réseaux augmente fortement, ce qui complique l'interprétation des résultats en rendant difficile la distinction entre interactions significatives et bruit.

En conclusion, la méthode PLN nous a permis d'obtenir des premiers résultats intéressants et exploitables pouvant constituer une base solide pour l'analyse des interactions microbiennes. Ces résultats soulignent la forte influence du biotope dans les interactions microbiennes et suggèrent que si on inclut cet effet dans le modèle, les interactions deviennent considérablement moins nombreuses.

6 Discussion et perspectives

Ce travail a permis de montrer que la réutilisation et l'intégration de jeux de données publiques de métagénomique bactérienne est une approche pertinente et prometteuse en écologie microbienne.

En ce qui concerne les réseaux d'interaction, bien qu'une partie du travail ait porté sur certaines méthodes spécifiques, il est important de rappeler qu'il n'existe pas encore de consensus sur la méthode optimale à utiliser. Par conséquent, une perspective intéressante serait de tester d'autres approches pour mieux comprendre leurs avantages et leurs limites dans le contexte de l'analyse des communautés microbiennes. Plusieurs méthodes différentes sont développées dans le package *R* {*PLNmodels*} et elles mériteraient d'être approfondies.

Il faut rappeler également que des questions importantes se posent aussi lors de la construction du jeu de données étudié que ce soit sur le choix de la résolution taxonomique (espèce, genre, ...) ou sur l'impact des filtres pouvant être nécessaires avant l'inférence de réseau.

Aussi, les métadonnées publiques associées aux différents jeux de données restent un facteur limitant et la normalisation des termes est un point crucial pour améliorer l'intégration des données issues de différentes études. Une piste intéressante pour la continuité de ce travail serait de contribuer à la conception de nouveaux guides pour les études métagénomiques 16S, dans le but de faciliter les étapes d'intégration dans un contexte de science ouverte. L'utilisation d'ontologies telles que OntoBiotope [Nédellec et al. 2018] pourrait jouer un rôle clé pour uniformiser les termes sur les sources de prélèvement des échantillons (par exemple, les aliments, les intestins, ...). De plus, le recours à des techniques de fouille de texte (*text mining*) pour extraire des informations directement depuis les publications associées aux jeux de données est une perspective à envisager.

Une prolongation en CDD sur le projet m'a donc été proposée afin de continuer à approfondir ces différentes pistes et renforcer les bases méthodologiques posées pendant ce stage. Cette prolongation permettra non seulement de consolider les résultats obtenus, mais aussi de contribuer de manière plus significative au développement de la science ouverte à INRAE.

7 Conclusion

Ce stage a été une expérience très enrichissante, marquée par une grande liberté d'exploration et d'initiative. J'ai eu l'opportunité de travailler sur des problématiques nouvelles, ce qui m'a permis de développer ma créativité et mon autonomie. J'ai aussi pu améliorer mes compétences sur la mise en œuvre d'outils interactifs. Ces visuels peuvent ainsi être explorés par tous (des biologistes, des bioinformaticiens, ...) et permettent une meilleure communication entre les différents acteurs de la recherche.

Cependant, cette liberté s'est aussi révélée être un véritable challenge. Travailler sur un projet où peu de travaux préalables existent signifie qu'il n'y a pas de modèle ou de référence claire à suivre. Cela a rendu le travail plus complexe dans la mesure où il était plus difficile de se raccrocher à des méthodes ou des solutions déjà éprouvées. Néanmoins, cette situation est inhérente au monde de la recherche, où l'innovation et la prise de risque sont indispensables pour avancer.

En somme, le stage a été une excellente introduction à la réalité du travail d'ingénieur dans le domaine de la recherche, m'offrant à la fois un défi stimulant et une expérience très formatrice.

Références

[1] Gabriele Berg et al. "Microbiome definition re-visited : old concepts and new challenges". In : *Microbiome* 8.1 (2020), p. 103. ISSN : 2049-2618. DOI : 10.1186/s40168-020-00875-0. URL : https://doi.org/10.1186/s40168-020-00875-0.

[2] Benjamin J Callahan et al. "DADA2 : High-resolution sample inference from Illumina amplicon data". en. In : *Nat Methods* 13.7 (mai 2016), p. 581-583. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4927377/pdf/nihms782534.pdf.

[3] Julien Chiquet, Mahendra Mariadassou et Stéphane Robin. "The Poisson-Lognormal Model as a Versatile Framework for the Joint Analysis of Species Abundances". In : *Frontiers in Ecology and Evolution* 9 (mars 2021). DOI : 10.3389/fevo.2021.588292. URL : https://hal.sorbonne-universite.fr/hal-03215628.

[4] Zachary D.Kurtz et al. "Sparse and Compositionally Robust Inference of Microbial Ecological Networks". In : *PLOS Computational Biology* (2015). URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4423992/pdf/pcbi.1004226.pdf.

[5] Frédéric Escudié et al. "FROGS : Find, Rapidly, OTUs with Galaxy Solution". In : *Bioinformatics* 34.8 (déc. 2017), p. 1287-1294. ISSN : 1367-4803. DOI : 10.1093/bioinformatics/btx791. eprint : https://academic.oup.com/bioinformatics/article-pdf/34/8/1287/48915593/bioinformatics_34_8_1287.pdf. URL : https://doi.org/10.1093/bioinformatics/btx791.

[6] Hélène Falentin et al. "Guide pratique à destination des biologistes, bioinformaticiens et statisticiens qui souhaitent s'initier aux analyses métabarcoding". In : *Cahier des Techniques de l'INRA* 97 (2019), p. 46-69. URL : https://hal.science/hal-02311421.

[7] Marti J.Anderson. *PERMANOVA Permutational multivariate analysis of variance*. 2005. URL : https://www.yumpu.com/en/document/read/8286549/permanova-department-of-statistics.

[8] Wu Jun et al. "Natural Connectivity of Complex Networks". In : *Chinese Physics Letters* 27.7 (2010), p. 078902. DOI : 10.1088/0256-307X/27/7/078902. URL : https://dx.doi.org/10.1088/0256-307X/27/7/078902.

[9] Sungkyu Jung. *Lecture 8 : Multidimensional scaling*. 2013. URL : https://www.stat.pitt.edu/sungkyu/course/2221Fall13/lec8_mds_combined.pdf.

[10] Han Liu, Kathryn Roeder et Larry Wasserman. "Stability Approach to Regularization Selection (StARS) for High Dimensional Graphical Models". en. In : *Adv Neural Inf Process Syst* 24.2 (déc. 2010), p. 1432-1440. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4138724/.

[11] Eric Marcon. "Mesures de la Biodiversité". Master. Lecture. Kourou, France, sept. 2015. URL : https://agroparistech.hal.science/cel-01205813.

[12] Paul J McMurdie et Susan Holmes. "phyloseq : an R package for reproducible interactive analysis and graphics of microbiome census data". en. In : *PLoS One* 8.4 (avr. 2013), e61217. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3632530/pdf/pone.0061217.pdf.

[13] Paul J McMurdie et Susan Holmes. "Waste not, want not : why rarefying microbiome data is inadmissible". en. In : *PLoS Comput Biol* 10.4 (avr. 2014), e1003531. URL : https://pubmed.ncbi.nlm.nih.gov/24699258/.

[14] Claire NÉDELLEC et al. "L'ontologie OntoBiotope pour l'étude de la biodiversité microbienne". In : *EGC'2018 : 18èmes Conférence Internationale sur l'Extraction et la Gestion des Connaissances*. Paris, France, jan. 2018. URL : https://hal.inrae.fr/hal-02737399.

[15] Clément POUPELIN. *Blog projet Open16s*. 2024. URL : http://clement.poupelin.pages.mia.inra.fr/open16s_blog/.

[16] Drew WILIMITIS. *Multidimensional Scaling*. 2019. URL : https://github.com/drewwilimitis/Manifold-Learning/blob/master/Multidimensional_Scaling.ipynb.

[17] Amy WILLIS et John BUNGE. "Estimating Diversity via Frequency Ratios". In : *Biometrics* 71.4 (juin 2015), p. 1042-1049. ISSN : 0006-341X. DOI : 10.1111/biom.12332. eprint : https://academic.oup.com/biometrics/article-pdf/71/4/1042/55135365/biometrics_71_4_1042.pdf. URL : https://doi.org/10.1111/biom.12332.

[18] Jun WU et al. *Robustness of Random Graphs Based on Natural Connectivity*. 2010. arXiv : 1009.3430 [cond-mat.stat-mech]. URL : https://arxiv.org/pdf/1009.3430.

Annexe

Chao1 [Marcon 2015]

L'indice de Chao1 est donc de la forme suivante :

$$S_{Chao1} = S_{rich} + \hat{c}_0$$

Cela représente le nombre d'espèces observées (S_{rich}) dans un échantillon auquel on associe une estimation du nombre d'espèces non observées.
Nous allons montrer ici comment est construit $\hat{c}_0$.

Soit p_s avec $s \in \{1, ..., S\}$ la probabilité qu'une séquence appartienne à l'espèce s et c_i, $i \in \mathbb{N}$, le nombre d'espèce observées i fois. Supposons que dans un échantillon de taille n résultant d'un tirage indépendant de séquences, la probabilité que l'espèce s soit observée k fois suit une loi binomiale. L'espérance du nombre d'espèces observées k fois est alors de la forme :

$$\mathbb{E}(c_k) = \sum_s \mathbb{P}(s \text{ observée } k \text{ fois}) = \sum_s \binom{n}{k} p_s^k (1-p_s)^{n-k}$$

L'idée est d'estimer le nombre d'espèces non observées à partir de celles observées 1 et 2 fois.
Ainsi on obtient que

$$\mathbb{E}(c_0) = \sum_s (1-p_s)^n \qquad \mathbb{E}(c_1) = n\sum_s p_s(1-p_s)^{n-1} \qquad \mathbb{E}(c_2) = \frac{n(n-1)}{2}\sum_s p_s^2(1-p_s)^{n-2}$$

Puis grâce à l'inégalité de Cauchy-Schwarz, on a

$$\left(\sum_s p_s(1-p_s)^{n-1}\right)^2 \leq \left(\sum_s (1-p_s)^n\right)\left(\sum_s p_s^2(1-p_s)^{n-2}\right)$$

d'où l'inégalité suivante

$$\mathbb{E}(c_0) \geq \frac{n-1}{2n}\frac{\mathbb{E}(c_1)^2}{\mathbb{E}(c_2)}$$

Ainsi, on peut utiliser les moyennes observées pour remplacer les espérances de c_1 et c_2. Cela nous permet de construire un estimateur minimum où l'espérance du nombre d'espèces observées zéro fois est supérieure ou égale au nombre estimé.

$$\hat{c}_0 = \frac{n-1}{2n}\frac{(\bar{c}_1)^2}{(\bar{c}_2)}$$

Shannon entropy [Marcon 2015]

Soit un échantillon à n séquences avec n grand et p_s la probabilité qu'une séquence appartienne à l'espèce s.
On enregistre alors la liste (ordonnée) des espèces de n séquences. Le nombre de séquence correspondant à l'espèce s pourra être quantifié par np_s.
Ainsi, en calculant le nombre de positions possibles dans la liste des séquences appartenants à une première espèce on a $\binom{n}{np_1}$. Pour une deuxième on aurait $\binom{n-np_1}{np_2}$ et pour la S-ième $\binom{n-np_1-\ldots-np_{S-1}}{np_S}$. Le produit de ces combinaisons est :

$$L = \frac{n!}{\Pi_s(np_s)!}$$

En passant au logarithme, on obtient

$$\ln(L) = \ln(n!) - \sum_s \ln((np_s)!)$$

Ensuite, on utilise l'approximation de stirling qui, pour un x suffisament grand, nous donne que $\ln(x!) \approx x\ln(x) - x$.
On obtient alors que

$$\begin{aligned}
\ln(L) &\approx n\ln(n) - n - \sum_s np_s\ln(np_s) - np_s \\
&\approx n\left(\ln(n) - 1\right) - n\left(\sum_s p_s\ln(np_s) - \sum_s p_s\right) \\
&\text{Or, } \sum_s p_s = 1 \\
\ln(L) &\approx n\left(\ln(n) - 1\right) - n\left(\sum_s p_s\ln(np_s) - 1\right) \\
&\approx n\left(\ln(n) - 1\right) - n\left(\sum_s p_s\ln(n) + \sum_s p_s\ln(p_s) - 1\right) \\
&\approx n\left(\ln(n) - 1\right) - n\left(\ln(n) + \sum_s p_s\ln(p_s) - 1\right) \\
&\approx n\left(\ln(n) - 1 - \ln(n) - \sum_s p_s\ln(p_s) + 1\right) \\
&\approx -n\sum_s p_s\ln(p_s)
\end{aligned}$$

Ainsi, on obtient l'indice de Shannon

$$S_{Shan} = -\sum_s p_s\ln(p_s) \approx \frac{\ln(L)}{n}$$

Inverse Simpson [Marcon 2015]

Soit p_s la probabilité qu'une séquence tirée aléatoirement dans un échantillon appartienne à l'espèce s.
Soit A et B deux séquences, en supposant les tirages indépendants, on a

$$\mathbb{P}(\text{"}A \text{ appartient à l'espèce } s\text{"}, \text{"}B \text{ appartient à l'espèce } s\text{"}) = p_s^2$$

Alors, on peut poser λ comme étant probabilité que 2 séquences tirées aléatoirement puissent appartenir à la même espèce :

$$\begin{aligned}\lambda &= \mathbb{P}\left(\bigcup_s (\text{"}A \text{ appartient à l'espèce } s\text{"}, \text{"}B \text{ appartient à l'espèce } s\text{"})\right) \\ &= \sum_s p_s^2\end{aligned}$$

Ainsi, l'indice d'inverse de Simpson est $\frac{1}{\lambda}$ de telle sorte que si λ grand, alors beaucoup de séquences sont de la même espèce ce qui implique moins de diversité. Et en prenant l'inverse on aura que si l'indice est faible, alors la diversité sera faible. D'où

$$S_{InvSimp} = \frac{1}{p_1^2 + ... + p_s^2}$$

MDS (*MultiDimensional scaling*) [Wilimitis 2019] et [Jung 2013]

Il existe plusieurs techniques référencées comme du positionnement multidimensionnel. Ici, c'est la technique du positionnement multidimensionnel classique (*classical multidimensional scaling*) dont il est question.
Le positionnement multidimensionnel classique aussi connu sous le nom d'analyse en coordonnées principales (PCoA) permet de visualiser les données de grande dimension dans un espace euclidien de dimension réduite.

Le principe de la MDS est proche de la PCA (Analyse en Composantes Principales) à l'exception que la PCA est basée uniquement sur des matrices de distance euclidienne alors que la MDS peut être appliquée à toute matrices $D = \{d_{ij}\}_{i,j \in \mathbb{R}}$ de distance ou de dissimilarité (par exemple Bray-Curtis pour les données d'écologie microbienne). Ainsi, MDS diffère des autres méthodes de réduction de dimensionnalité dans la mesure où l'entrée dans MDS est uniquement la matrice de distance/dissimilarité, au lieu des vecteurs de position réels des données. Alors, étant donné une matrice de distance/dissimilarité $D \in \mathbb{R}^{n \times n}$ avec d_{ij} représentant la distance/dissimilarité entre i et j, nous avons $x_1, ..., x_n \in \mathbb{R}^k$, avec k qui va définir la dimension de sortie, tel que :

$$\underbrace{d_{ij}^2}_{\text{distance/dissimilarité original}} \approx \underbrace{||x_i - x_j||^2}_{\text{configuration de sortie}}$$

On retrouve une configuration qui maintient les distances euclidiennes dans $\mathbb{R}^k$ (généralement dans $\mathbb{R}^2$) aussi proche que possible de nos distances/dissimilarité d'origine.

Soit $X = (x_1, ..., x_n)'$ et $G = XX'$.
En transformant la matrice de distance/dissimilarité par un double centrage, nous obtenons la relation :

$$G = -\frac{1}{2} C D^2 C \quad \text{avec} \quad C = I - \frac{1}{N} \mathbb{1}\mathbb{1}'$$

I désigne la matrice identité de dimension N et $\mathbb{1}$ le vecteur de 1 de dimension N.

Une décomposition via les valeurs propres est ensuite effectuée sur G pour définir X dont les lignes contiennent les coordonnées principales.

$$G = U \Lambda U'$$

$$X = U \Lambda^{\frac{1}{2}}$$

avec U la matrice des vecteurs propres symétrique définie positive.

Le MDS est particulièrement adapté pour l'analyse de données métagénomiques 16S afin de représenter graphiquement les échantillons tout en préservant les distances ou dissimilarités entre eux. La dissimilarité de Bray-Curtis étant la plus souvent choisie.