"Truffinet": Inférence de réseaux d'interactions microbiennes dans la truffe. Rapport scienti que du PEPS Mirabelles 2014

Aurélie Muller-Gueudin, Marc Buée, Aurélie Deveau, Anne Gégout-Petit, Samuel Martin, Irinel-Constantin Morarescu, Chedy Raïssi

Rapport scientifique du PEPS Mirabelles 2014 "Truffinet": Truffles' microbial interaction inference by network analysis

Porteur : Aurélie Muller-Gueudin, IECL, UMR 7502/ Equipe Inria BIGS.

Participants :
— Marc Buée, INRA Nancy, UMR 1136, Interactions Arbres Micro-organismes
— Aurélie Deveau, INRA Nancy, UMR 1136, Interactions Arbres Micro-organismes
— Anne Gégout-Petit, IECL, UMR 7502 Equipe Inria BIGS
— Samuel Martin, CRAN, UMR 7039
— Constantin Morarescu, CRAN, UMR 7039
— Aurélie Muller-Gueudin, IECL, UMR 7502 / Equipe INRIA BIGS
— Chedy Raïssi, INRIA Nancy Grand Est / LORIA

1 Introduction et contexte

Les truffes abritent des communautés bactériennes extrêmement denses – jusqu'à 100 millions de bactéries par gramme de truffe. Des études récentes suggèrent que ces bactéries pourraient participer à la nutrition du champignon et au développement de son arôme. Des recherches ont été entreprises à l'INRA au sein de l'UMR 1136 Interactions Arbres-Micro organismes pour identifier les bactéries présentes dans les truffes et analyser leur rôle dans le développement des truffes et leurs arômes. A l'aide des nouvelles technologies de séquençage à haut débit, des inventaires des bactéries présentes aux différents stades de développement de la truffe ont été obtenus. Ces données indiquent la présence de communautés complexes et dynamiques composées de centaines d'espèces qui interagissent en réseaux (Antony Babu et al. 2013). L'analyse de ces réseaux d'interactions est nécessaire pour pouvoir comprendre le rôle des communautés bactériennes dans le développement

de la truffe. Ces réseaux ne peuvent pas être visualisés expérimentalement mais ils peuvent être inférés à partir des données de séquençage à haut débit.

2 Présentation des données

2.1 Premier jeu de données

Les modèles statistiques mis au point dans le cadre du projet Truffinet ont été construits à partir d'un premier jeu de données acquis préalablement au début du projet [2], [4]. Ce jeu de données est constitué de 42 échantillons répartis en six classes : bactéries colonisant le sol, les ectomycorrhizes, la zone d'influence de l'ectomycorhize (i.e ectomycorhizosphere), les parties internes et externes de la truffe et le sol adhérant à la truffe. Les échantillons ont été collectés chaque mois entre octobre 2010 et janvier 2011, période correspondant à la saison de maturation de la truffe. Pour chaque type d'échantillon et à chaque date d'échantillonnage, 3 à 4 réplicats biologiques ont été obtenus. La composition des communautés bactériennes associées à chaque échantillon a été obtenue par pyroséquençage 454 de la région 16S de l'ARN ribosomique après extraction et purification des ADNs. Les séquences obtenues se répartissent au sein de 27 phyla et 566 genres bactériens.

2.2 Acquisition des nouvelles données

Afin d'obtenir un plus large jeu de données et de valider les modèles établis à partir d'un premier jeu de données réduit, de nouveaux échantillons ont été prélevés en janvier 2014. Trente truffes noires T. melanosporum et trente échantillons de sols ont été récoltés le 28 janvier 2014 sur la truffière expérimentale de Rollainville (Vosges). Les ADNs ont été extraits selon le protocole décrit par Antony-Babu et al. (2014) puis la région 16S de l'ARN ribosomique a été amplifiée par PCR pour chaque échantillon à l'aide des amorces 787R

$$(xxxxx - ATTAGATACCYTGTAGTCC)$$

et 1073F

$$(xxxxx - ACGAGCTGACGACARCCATG)$$

couplées à un code barre spécifique. Après purification et vérification de la qualité des amplicons, les amplicons ont été rassemblés en un mélange équimolaire puis séquencés par pyroséquençage 454 à l'aide de la technologie GS-FLX 454 Titanium par la société Beckman Counter Genomics. 512 480

séquences ont été obtenues. Après démultiplexage et nettoyage des données à l'aide du logiciel MOTHUR (Schloss et al. 2009), les séquences ont été classifiées en Operational Taxonomic Units (OTUs, 97% de similarité de séquences) et la correspondance taxonomique des OTUs attribuée par comparaison à la base de données SILVA. 6080 OTUs, réparties au sein de 16 phyla et 121 genres bactériens, ont été obtenues.

2.3 Jeux de données utilisés dans la suite

Dans la suite de ce rapport, nous présentons des résultats obtenus pour les jeux de données suivants :

1. Jeu de données 2010 : il s'agit des données dont nous disposions au début de ce projet PEPS. Plus précisément, il s'agit de 613 genres bactériens mesurés sur 30 échantillons (15 échantillons à l'intérieur de la truffe, notés "Inside truffle", et 15 échantillons dans le sol autour de la truffe, notés "Ascocarpic soil").
2. Jeu de données 2014 : il s'agit des données acquises en janvier 2014. Plus précisément, il s'agit de 121 genres bactériens mesurés sur 60 échantillons (30 échantillons à l'intérieur de la truffe, notés "Inside truffle", et 30 échantillons dans le sol autour de la truffe, notés "Ascocarpic soil").

Notons que le premier jeu de données a été récolté sur une période de quatre mois allant de octobre 2010 à janvier 2011. Une étude préliminaire nous a montré qu'il n'était pas nécessaire de distinguer les mois.

Notons enfin que nous disposons, dans l'union des jeux de données de 2010 et 2014, de 223 espèces pour Inside truffle, et 466 espèces pour Ascocarpic soil.

3 Normalisation des données

Les techniques actuelles de séquençage permettent de dénombrer les individus présents dans un échantillon, et de les classer ensuite selon leur OTU, leur phylum, et/ou leur genre. Le nombre total d'individus dénombré varie d'un échantillon à l'autre, et rend impossible une comparaison directe entre échantillons. Ces variations de taille sont dues d'une part à la variabilité biologique, mais aussi à la variabilité intrinsèque à la technologie de séquençage.

Il est donc nécessaire de normaliser les données. Pour diminuer la sensibilité aux espèces rares, la procédure habituellement utilisée en analyse de microbiome est :

— soit de travailler avec les abondances relatives (ratio des abondances de chaque espèce sur le nombre total d'individus mesurés dans l'échantillon) ;
— soit de raréfier (i.e. sous-échantillonner) les échantillons de manière à ce que les échantillons aient tous le même nombre d'individus (i.e. égal à la plus petite taille observée).

Mais aucune de ces méthodes n'est satisfaisante [8]. Dans le cas d'une analyse différentielle entre échantillons de microbiomes provenant de conditions expérimentales différentes, ces deux méthodes entraînent de faux positifs.

D'autres techniques alternatives (assorties de packages R) ont été proposées en analyse de séquences d'ARN. De nombreux auteurs [1, 12, 8, 18] proposent de prendre en compte la variabilité biologique et technologique via un modèle de Poisson surdispersé, i.e. via une loi binomiale négative : en notant Y_{ei} le nombre d'individus correspondant à l'espèce e, dans l'échantillon i, ils supposent que $Y_{ei} \sim NB(\mathbb{E}(Y_{ei}), \Phi_e)$ où (Y_{ei}) est la moyenne théorique de Y_{ei} et Φ_e est un paramètre de dispersion de l'espèce e, à estimer.

D'autres auteurs, [10] proposent de modéliser Y_{ei} par une loi gaussienne à inflation de zéro. Les tableaux de données dont nous disposons comportent en effet un grand nombre de zéros. Ces zéros peuvent être dus à une absence de l'espèce ou bien au hasard d'un sous-échantillonnage. Puisque dans la suite, nous allons inférer un graphe d'interactions entre espèces via un modèle graphique gaussien, nous choisissons la technique de normalisation proposée par [10] et implémentée dans le package R `metagenomeSeq`.

Ainsi, pour chacune des bases des données décrites dans la suite :

1. nous avons tout d'abord appliqué un premier filtre en supprimant les espèces non présentes dans l'ensemble des échantillons. Si pour tout échantillon i, aucun individu de l'espèce e n'est présent, alors nous supprimons l'espèce e de la base des données.
2. Puis, nous avons normalisé la base de données obtenue après ce premier filtre, avec la fonction `MRcounts` du package `metagenomeSeq`.

Les effectifs des bases de données sont présentés dans le tableau 1.

	Inside truffle	Ascocarpic soil
Jeu de données 2010	$n = 15, p = 181$	$n = 15, p = 406$
Jeu de données 2014	$n = 30, p = 76$	$n = 30, p = 121$
Jeu de données commun	$n = 45, p = 59$	$n = 45, p = 73$

Tableau 1 – Effectifs des jeux de données, n représente le nombre d'échantillons disponibles, p le nombre de genres bactériens, après normalisation des données.

4 Modèles graphiques gaussiens

Pour chaque jeu de données, notre objectif est de donner un graphe d'interactions entre les espèces présentes. Un graphe est composé de sommets (les espèces ici) et d'arêtes entre ces sommets. Deux espèces sont reliées par une arête si elles sont en interaction (dans un sens à définir !). Les graphes peuvent être orientés ou pondérés. Il faut donc à la fois définir le sens mathématique de cette interaction et utiliser les données pour faire de l'inférence sur ces interactions. Pour cette étude, une première méthode a été la modélisation par un modèle graphique gaussien que nous présentons ici. Pour une synthèse pédagogique sur les modèles graphiques et leur inférence, on peut consulter [16] dont nous nous sommes inspirés pour la rédaction de cette section.

4.1 Définition

Nous supposons ici que notre vecteur des abondances pour les p espèces étudiées $X = (X_1, \ldots, X_p)$ suit une loi gaussienne de moyenne μ et de matrice de variance-covariance Γ. A ce vecteur gaussien, on associe un graphe à p sommets. Deux sommets i et j de ce graphe sont reliés si la corrélation partielle entre X_i et X_j sachant les $(p-2)$ autres variables est non nulle. En effet il est important de considérer la corrélation partielle entre deux variables et non la corrélation simple qui peut être forte mais due par exemple à l'effet d'une variable tierce sur ces deux variables. Par définition, la corrélation partielle de deux variables X_i et X_j dans un ensemble $X = (X_1, \ldots, X_p)$ est la corrélation des résidus respectifs de la régression de X_i et X_j sur les $(p-2)$ autres variables. Une importante propriété des modèles gaussiens relie la matrice de concentration $S = \Gamma^{-1}$ aux corrélations partielles ρ_{ij} par la relation :

$$\rho_{ij} = -\frac{S_{ij}}{\sqrt{S_{ii}S_{jj}}}$$

pour $i \neq j$. L'enjeu est donc d'estimer la matrice S pour inférer le graphe entre les espèces.

4.2 Inférence

Les références sont nombreuses sur le sujet, on en trouvera une synthèse dans [16] et [6]. La difficulté de l'inférence de S via celle de la matrice de covariance Γ vient du fait que le nombre n d'observations de réplicats du vecteur X est inférieur au nombre d'espèces p. La matrice de covariance empirique $\hat{\Gamma}$ est alors mal conditionnée. Plusieurs méthodes sont alors possibles. On peut faire du "multiple testing" en calculant la corrélation partielle de X_i

et X_j sachant un sous ensemble S de taille $(n-2)$ des $(p-2)$ autres variables et ceci pour beaucoup de tels sous-ensemble. Une autre idée de [13, 14] est d'estimer S par une estimation de la matrice $(\Gamma + \lambda Id)^{-1}$ où $\lambda > 0$ est un paramètre de pénalisation petit.

Le "Graphical Lasso", quant à lui, utilise le lien entre les coefficients $(\beta_k^j)_{k \neq j}$ de la régression de X_j sur les $(p-1)$ autres variables du vecteur X à la matrice de concentration S par la relation :

$$\beta_k^j = -\frac{S_{jk}}{S_{jj}}.$$

Les méthodes Lasso de régression [9, 5] sont alors disponibles, elles consistent en la pénalisation des moindres carrés par l'ajout d'un terme de la forme $\lambda \|\beta^j\|_1$. Le terme de pénalisation "force" la norme de β à être petite et les termes de β s'annulent un à un lorsque λ grandit. Le choix du paramètre λ est donc primordial. Nous avons pour ce travail, utilisé la méthode de [6] qui combine ces approches et propose une méthode pour sélectionner le graphe optimal. Elle est implémentée dans le package `GGMselect`.

4.3 Choix du paramètre λ optimal

Nous avons donc utilisé le package `GGMselect` pour chercher le paramètre λ optimal dans une fenêtre $[\lambda_{\min}, \lambda_{\max}]$:

- $\lambda_{\min}$ correspond au plus petit λ tel que le graphe résultant a un degré inférieur à $n-3$ où n est le nombre d'échantillons disponibles dans le jeu de données.
- $\lambda_{\max}$ correspond au plus petit λ tel que le graphe est vide.

5 Résultats

5.1 Choix du paramètre λ de régularisation

Pour chacun des six jeux de données, nous cherchons le paramètre de régularisation λ optimal au sens défini dans la section précédente. Plus λ est grand, plus la régularisation est forte, et moins il y aura d'arêtes, et ce, jusqu'à ce qu'il n'y ait plus d'arêtes. La figure 1 illustre ce phénomène pour le jeu de données 2010 Ascocarpic soil. En pratique, nous avons montré que la valeur optimale de λ, au sens de [6], est comprise entre 0 et 3, pour des données centrées et réduites.

Les valeurs obtenues pour λ sont reportées dans le tableau 2. Dans certains cas, le paramètre optimal proposé par [6] n'a pas pu être calculé. En

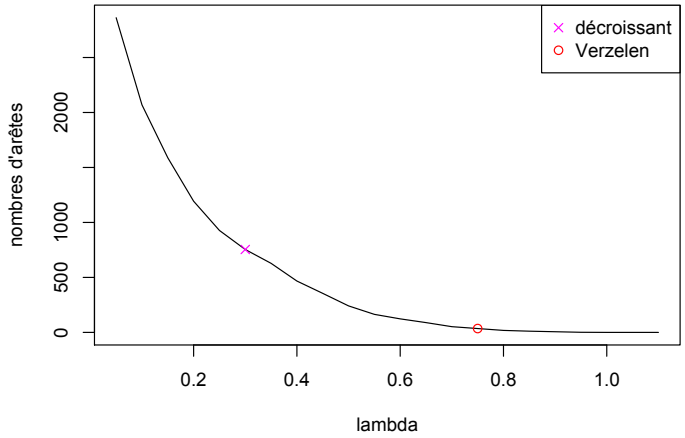

FIGURE 1 – Evolution du nombre d'arêtes dans le modèle graphique gaussien, en fonction du choix de λ. Illustration sur le jeu de données 2010, Ascocarpic soil. La légende indique les valeurs du λ décroissant (i.e. la valeur $\lambda_{\min}$ définie plus haut) et du λ optimal au sens de [6].

effet, pour ces cas, $\lambda_{\min} = \lambda_{\max}$, i.e. nous avons un graphe vide. De plus, pour ces cas pathologiques, nous avons $\lambda_{min} \approx 1+4\cdot 10^{-16}$. Nous avons alors choisi de considérer $\lambda = 1 < \lambda_{\max}$ (ce qui nous assure un graphe parcimonieux, mais non vide).

	Inside truffle	Ascocarpic soil
Jeu de données 2010	$\lambda = 1$	$\lambda = 1$
Jeu de données 2014	$\lambda = 1$	$\lambda = \lambda_V = 0.75$
Jeu de données commun	$\lambda = 1$	$\lambda = \lambda_V = 0.7$

Tableau 2 – Choix du paramètre λ de régularisation. La notation λ_V indique qu'il s'agit du λ optimal au sens de [6].

5.2 Matrices d'adjacence et de corrélations partielles

Après avoir choisi λ, l'inférence de la matrice ρ des corrélations partielles est effectuée via la fonction `glasso` du package `glasso` de R. Nous obtenons

$\widehat{\rho} = (\widehat{\rho}_{ij})_{ij}$. Nous en déduisons la matrice d'adjacence $\widehat{A} = (\widehat{a}_{ij})_{ij}$ définie par :

$$\begin{array}{rcl} \widehat{a}_{ij} = 0 & \Leftrightarrow & \widehat{\rho}_{ij} = 0 \\ \widehat{a}_{ij} = 1 & \Leftrightarrow & \widehat{\rho}_{ij} \neq 0 \end{array}$$

Une fois ces matrices estimées, nous pouvons tracer le graphe d'interactions : deux espèces i et j seront reliées par une arête si et seulement si $\widehat{\rho}_{ij} \neq 0$. Toutefois, pour effectuer un test sur la présence ou non de chaque arête, nous avons calculé un intervalle de confiance à 99.7% des ρ_{ij} via $B = 5000$ tirages Bootstrap avec remise dans les échantillons :

$$\left[\widehat{\rho}_{ij}^{(B)} \pm 3 \frac{\sqrt{\text{var}(\widehat{\rho}_{ij})}}{\sqrt{B}} \right],$$

où

$$\widehat{\rho}_{ij}^{(B)} = \frac{1}{B} \sum_{b=1}^{B} \widehat{\rho}_{ij}^{b}$$

et où $(\widehat{\rho}_{ij}^{b})_{ij}$ désigne la matrice des corrélations partielles obtenue à la $b$ème itération bootstrap. Nous dirons qu'une arête (ij) est significative si l'intervalle de confiance ne contient pas 0. Précisons que $\text{var}(\widehat{\rho}_{ij})$ est calculé comme la variance empirique du vecteur $(\widehat{\rho}_{ij}^{b})_{b}$.

De plus, nous définissons la matrice d'adjacence bootstrap par

$$\widehat{a}_{ij}^{(B)} = \frac{1}{B} \sum_{b=1}^{B} \widehat{a}_{ij}^{b}$$

où $(\widehat{a}_{ij}^{b})_{ij}$ désigne la matrice d'adjacence obtenue à la $b$ème itération bootstrap. La figure 2 illustre la distribution des $(\widehat{a}_{ij}^{(B)})_{(ij) \text{ significatives}}$ pour le lot des arêtes (ij) significatives, et la distribution des $(\widehat{a}_{ij}^{(B)})_{(ij) \text{ non significatives}}$ pour le lot des arêtes (ij) non significatives. La distinction entre les deux distributions est nette.

La figure 3 illustre la matrice $(\widehat{\rho}_{ij}^{(B)})_{ij}$ des corrélations partielles du jeu de données 2014, Inside truffle. Les espèces ont été triées de sorte que les premières lignes de la matrice correspondent aux espèces qui n'intéragissent pas. Pour distinguer les valeurs des corrélations, nous avons artificiellement remplacé les valeurs de la diagonale (majoritairement égales à 1) par des 0. La figure 4 illustre le réseau d'interactions associé. Seules les arêtes (ij) telles que $\widehat{\rho}_{ij}^{(B)} > 2.5 \cdot 10^{-17}$ sont représentées. Notons que les corrélations partielles significatives sont toutes positives. En particulier, les corrélations partielles négatives n'ont pas été considérées comme significatives.

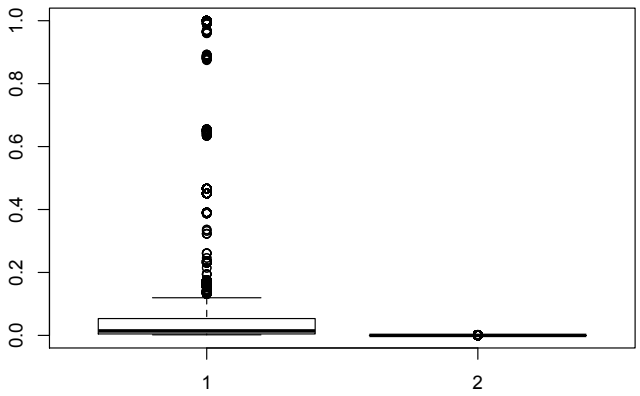

FIGURE 2 – Boxplot des valeurs des $(\widehat{a}_{ij}^{(B)})_{(ij)\text{significative}}$ (à gauche) et boxplot des valeurs des $(\widehat{a}_{ij}^{(B)})_{(ij)\text{non significative}}$ (à droite). Illustration sur le jeu de données 2010, Ascocarpic soil.

Il est important de noter que les valeurs des corrélations partielles sont à interpréter relativement les unes par rapport aux autres au sein d'un même jeu de données. En effet, les corrélations estimées dépendent du paramètre de régularisation λ. Rappelons que plus λ est grand, plus les corrélations sont faibles, jusqu'à toutes s'annuler à partir d'une certaine valeur de λ.

De plus, rappelons que la corrélation partielle et la corrélation brute peuvent être de signes différents. Nos résultats ne sont donc pas en contradiction avec le fait que certains couples d'espèces aient des corrélations brutes négatives.

Tous les calculs effectués pour les six jeux de données du tableau 1 sont récapitulés dans deux tableaux (disponibles sur demande) l'un pour Inside truffle, l'autre pour Ascocarpic soil. Ces tableaux, à 223 lignes pour Inside truffle (correspondant au nombre d'espèces rencontrées dans Inside truffle, dans l'union des jeux de données 2010 et 2014) et 466 lignes pour Ascocarpic soil, comportent 13 colonnes décrites ci-après. Pour chaque espèce e, nous notons :

1. sa présence ou non (1/0) dans le jeu 2010 ; sa présence ou non (1/0) dans le jeu 2014 ; et enfin sa présence ou non (1/0) dans l'intersection des deux jeux.
2. $\min_{i \neq e} \rho_{ie}$ pour i parcourant les espèces du jeu de 2010 ; pour i parcou-

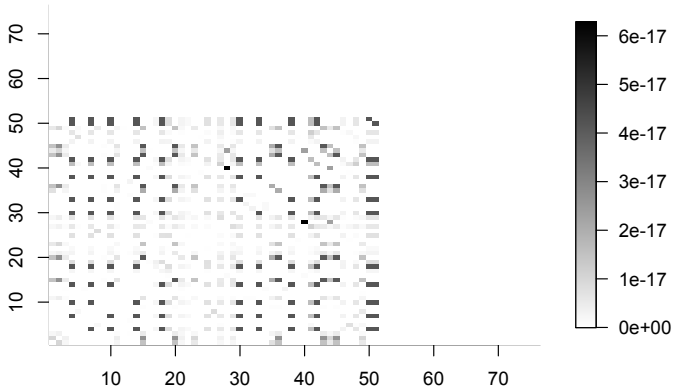

FIGURE 3 – Matrice des corrélations partielles du jeu de données 2014, Inside truffle.

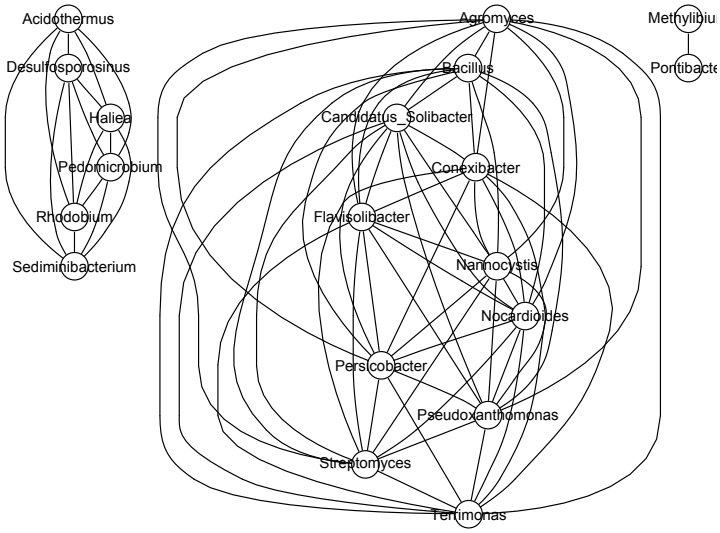

FIGURE 4 – Graphe d'interactions du jeu de données 2014, Inside truffle. Sont représentées seulement les arêtes (ij) telles que $\widehat{\rho}_{ij} > 2.5 \cdot 10^{-17}$.

rant les espèces du jeu de 2014 ; pour i parcourant les espèces communes aux deux jeux.

3. $\max_{i \neq e} \rho_{ie}$ pour i parcourant les espèces du jeu de 2010 ; pour i parcourant les espèces du jeu de 2014 ; pour i parcourant les espèces communes aux deux jeux.
4. $\text{degré}(e) = \frac{1}{2} \sum_{i \neq e} \mathbf{1}_{\rho_{ie} \neq 0}$ le degré de l'espèce e dans le jeu de 2010 ; dans le jeu de 2014 ; et enfin dans l'intersection des deux jeux.

Chaque fois qu'une case de ce tableau est impossible à calculer (si l'espèce e n'est pas dans le jeu considéré), la case est laissée vide.

6 Aspects dynamiques

Ayant constaté que les interactions observées via l'analyse statistique (définies par les corrélations partielles) sont toutes positives, nous avons ensuite mis en place le modèle de dynamique de population basé sur un réseau d'interactions mutualistes décrit dans [15]. Ce modèle permet de décrire l'évolution des populations des différentes bactéries au cours du temps. Il nécessite la connaissance de paramètres : pour chaque noeud, le taux de croissance et la limitation intrinsèque et pour chaque couple de noeuds, la force d'interaction, et le taux de demi saturation. Une étude par simulation a permis de constater la présence de divers comportements dynamiques tels que oscillations, extinctions ou stabilisation, en fonction des paramètres choisis. Cependant, il s'est avéré que les espèces de bactéries pour lesquelles des interactions ont été détectées grâce à l'analyse statistique ne sont jusqu'à présent pas cultivables en laboratoire. Par conséquent, il reste difficile de calibrer les paramètres des noeuds du modèle. D'autre part, la fréquence mensuelle pour l'échantillonnage des truffes implique une potentielle variation de paramètres entre deux échantillonnages, ce qui proscrit la validation du modèle dynamique à paramètre constant.

7 Aspects fouille de données : analyse log-linéaire

L'analyse log-linéaire est une technique statistique bien établie pour trouver des associations entre les variables dans les données [7]. Or, les approches classiques de l'analyse log-linéaires sont exponentielles par rapport au nombre de variables dans le jeu de données initial, puisque ces méthodes calculent la fréquence pour toutes les combinaisons possibles de valeurs. Pour un nombre

M de variables binaires, ces méthodes nécessitent donc 2^M opérations. Dans la pratique, les techniques classiques sont ainsi limitées à une douzaine, voire une vingtaine, de variables tout au plus. Un certain nombre de chercheurs ont étudié des approches qui construisent des modèles log-linéaires sur les sous-ensembles de variables et les combinent dans une étape ultérieure [17, 3]. Toutefois, ces méthodes sont incapables de récupérer les interactions complexes d'ordre élevé.

Dans le cadre de ce PEPS, nous avons utilisé un résultat récent sur une sous-classe puissante de modèles log-linéaires, à savoir, les modèles dits *"décomposables"* qui permettent d'adapter l'analyse log-linéaire à des données de grande dimension sans aucune restriction. En particulier, nous utilisons un résultat récent développé par F. Petitjean et al. [11] qui prouve que pour les modèles décomposables les tests χ^2 peuvent être calculés à partir d'un nombre significativement réduit de tableaux de contingence. Pour cela la méthode, nommée *Chordalysis*, utilise des résultats issus de la théorie des graphes et énonce qu'un modèle log-linéaire graphique est décomposable si le graphe lui correspondant est chordal, c'est à dire, si le graphe ne contient pas de cycles de longueur strictement supérieure à trois.

7.1 Résultats préliminaires

Nous avons appliqué la méthode *Chordalysis* sur les 4 principaux jeux de données proposés dans le cadre du PEPS (données classées respectivement par familles ; par genre ; par ordre ; par phylum de bactéries) et avons pu construire 4 graphes d'interactions permettant un début de compréhension des interactions complexes entre les différentes bactéries présentes dans et autour des truffes. Notons que chaque arête entre les noeuds représente une interaction et que chacun de ces modèles maximise le test de *"likelihood ratio"* (G^2) pour les modèles décomposables.

8 Conclusion et perspectives

Ce projet interdisciplinaire a été l'occasion de réaliser un véritable consortium entre tous les partenaires (biologistes, statisticiens, automaticiens, informaticiens). Nous avons commencé par analyser et modéliser le premier jeu de données. Cela a notamment fait l'objet d'un stage de M2 Statistique durant l'été 2014, financé sur le budget du laboratoire de Mathématiques IECL. Les nouvelles données ont été disponibles à partir de mai 2015. Ces nouvelles données ont permis d'étoffer les données déjà acquises, et de valider les techniques statistiques développées durant l'année 2014 et l'hiver

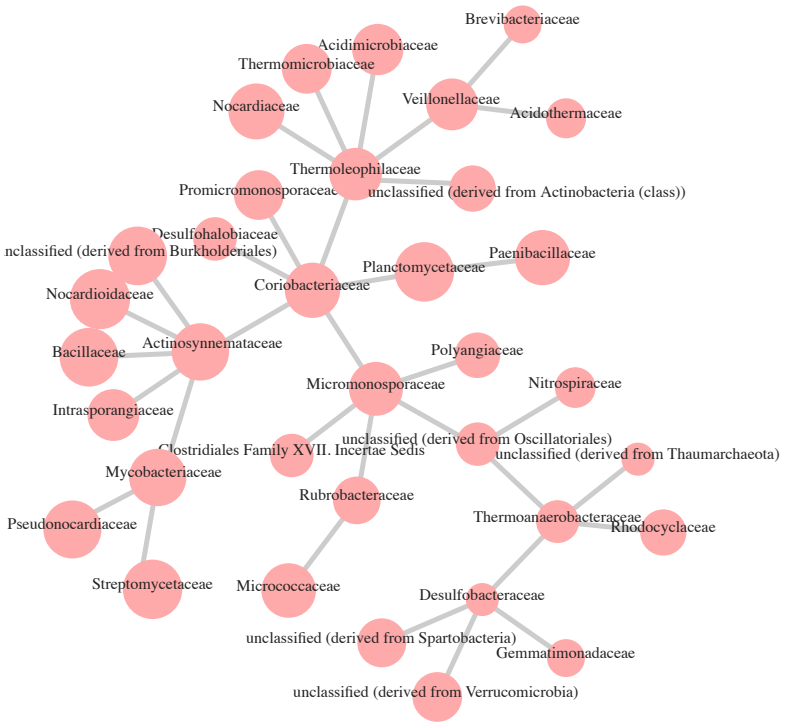

FIGURE 5 – Graphe d'interactions entre familles de bactéries.

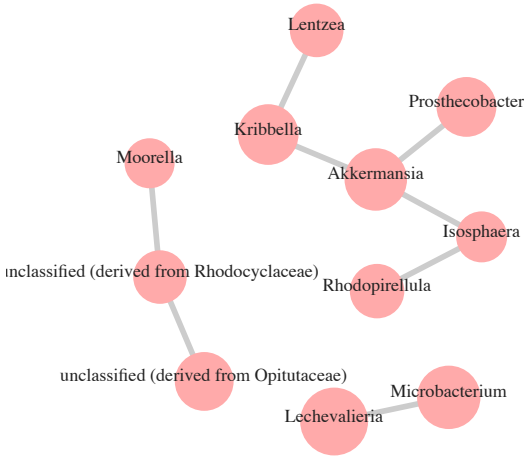

FIGURE 6 – Graphe d'interactions entre genres de bactéries.

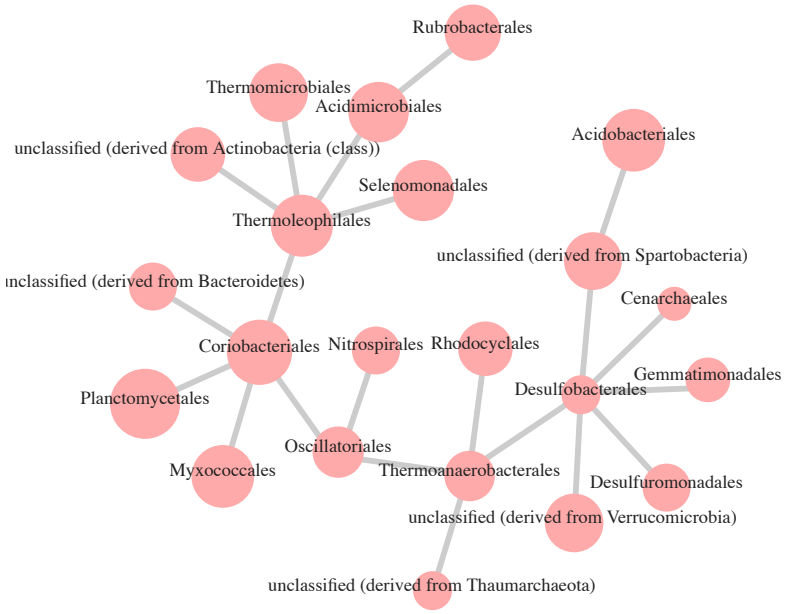

FIGURE 7 – Graphe d'interactions entre ordres de bactéries.

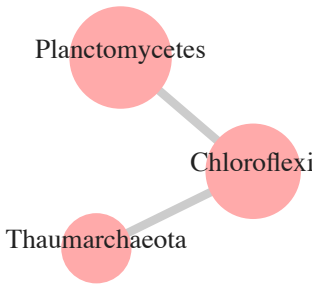

FIGURE 8 – Graphe d'interactions entre phyla de bactéries.

2014-2015. Le financement de l'acquisition de ces données par le PEPS a permis d'atteindre des tailles d'échantillons pour lesquelles les résultats sont beaucoup plus fiables et solides.

Les aspects dynamiques ont été développés, suite à l'analyse statistique. Cependant, l'état actuel des connaissances en biologie ne permet pas de quantifier les paramètres dont nous avons besoin pour exploiter un modèle de dynamique des populations. Nous pouvons à l'heure actuelle seulement proposer des simulations de modèles. Mais la variabilité d'échantillonnage est telle (une mesure mensuelle sur quatre mois) qu'elle ne permet pas pour l'instant de valider un modèle dynamique trivial à paramètre constant.

En parallèle des analyses statistiques, une étude en fouille de données à également été commencée, à partir de modèle log-linéaires. Dans la suite de cette étude, nous souhaitons comparer et mutualiser les informations issues des analyses statistiques et des analyses de fouille de données. Parmi les travaux restant à effectuer, notons que nous devons encore interpréter biologiquement les résultats obtenus. Très peu de résultats sur les interactions sont connus à l'heure actuelle. De plus, quand des résultats sont connus, il s'agit souvent de corrélations simples entre bactéries, mais pas de corrélations partielles i.e. de corrélations en tenant compte de toutes les autres espèces bactériennes présentes dans le milieu. D'autre part, les différentes analyses, menées sur différents types de jeux de données (bactéries classées respectivement par familles ; par genre ; par ordre et par phylum) apporteront des éclairages différents sur le problème étudié. Notons que l'analyse statistique s'est pour l'instant concentrée sur le jeux de données des genres bactériens. Les méthodes développées doivent encore être appliquées aux autres jeux de données (celui plus détaillé des familles, ou ceux moins détaillés des ordres et des phyla).

Enfin, à la suite de ce travail interdisciplinaire prometteur, nous avons décidé de poursuivre une collaboration sur le même type d'analyse de données (inférence de réseaux/fouille de graphe) sur un sujet biologique légèrement différent : évaluer l'existence de lien entre les modifications des communautés microbiennes et la biologie des arbres.

A Bilan financier : pièce jointe

Références

[1] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome biol*, 11(10) :R106, 2010.

[2] Sanjay Antony-Babu, Aurélie Deveau, Joy D Van Nostrand, Jizhong Zhou, François Le Tacon, Christophe Robin, Pascale Frey-Klett, and Stéphane Uroz. Black truffle-associated bacterial communities during the development and maturation of tuber melanosporum ascocarps and putative functional roles. *Environmental microbiology*, 16(9) :2831–2847, 2014.

[3] Corinne Dahinden, Markus Kalisch, and Peter Bühlmann. Decomposition and model selection for large contingency tables. *Biometrical Journal*, 52(2) :233–252, 2010.

[4] Aurélie Deveau, Sanjay Antony-Babu, François Le Tacon, Christophe Robin, Pascale Frey-Klett, and Stéphane Uroz. temporal changes of bacterial communities of the tuber melanosporum ectomycorrhizosphere. *In prep.*

[5] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3) :432–441, 2008.

[6] Christophe Giraud, Sylvie Huet, and Nicolas Verzelen. Graph selection with ggmselect. *Statistical applications in genetics and molecular biology*, 11(3), 2012.

[7] S. J. Haberman. Statistical algorithms : Algorithm AS 51 : Log-linear fit for contingency tables. 21(2) :218–225, June 1972.

[8] Paul J McMurdie and Susan Holmes. Waste not, want not : why rarefying microbiome data is inadmissible. *PLoS Comput Biol*, 10(4) :e1003531, 2014.

[9] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, pages 1436–1462, 2006.

[10] Joseph N Paulson, O Colin Stine, Héctor Corrada Bravo, and Mihai Pop. Differential abundance analysis for microbial marker-gene surveys. *Nature methods*, 10(12) :1200–1202, 2013.

[11] François Petitjean, Geoffrey I. Webb, and Ann E. Nicholson. Scaling log-linear analysis to high-dimensional data. In Hui Xiong, George Karypis, Bhavani M. Thuraisingham, Diane J. Cook, and Xindong Wu, editors, *2013 IEEE 13th International Conference on Data Mining, Dallas, TX,*

USA, December 7-10, 2013, pages 597–606. IEEE Computer Society, 2013.

[12] Mark D Robinson and Gordon K Smyth. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23(21) :2881–2887, 2007.

[13] Juliane Schäfer and Korbinian Strimmer. An empirical bayes approach to inferring large-scale gene association networks. *Bioinformatics*, 21(6) :754–764, 2005.

[14] Juliane Schäfer and Korbinian Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, 4(1), 2005.

[15] Elisa Thébault and Colin Fontaine. Stability of ecological communities and the architecture of mutualistic and trophic networks. *Science*, 329(5993) :853–856, 2010.

[16] Nathalie Villa-Vialaneix. An introduction to network inference and mining. http ://wikistat.fr/. (consulté le 22/07/2015).

[17] Xintao Wu, Daniel Barbará, and Yong Ye. Screening and interpreting multi-item associations based on log-linear modeling. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, pages 276–285, New York, NY, USA, 2003. ACM.

[18] Danni Yu, Wolfgang Huber, and Olga Vitek. Shrinkage estimation of dispersion in negative binomial models for rna-seq experiments with small sample size. *Bioinformatics*, 29(10) :1275–1282, 2013.