Université Paris Diderot - Paris 7 2016/2017

Master 2 Biologie-Informatique-Bioinformatique

Comparaison de méthodes statistiques d'inférence de réseaux de co-occurrences au sein d'écosystèmes microbiens à partir de données métagénomiques

Julie Lao

Sous la direction de :
Sophie Schbath - Mahendra Mariadassou

Mathématiques et Informatique Appliquées du Génome à l'Environnement
INRA Jouy-en-Josas

Remerciements

Je tiens tout d'abord à remercier Mahendra Mariadassou et Sophie Schbath pour leur encadrement, leurs conseils, leurs relectures et leur aide précieuse.

Je souhaite également remercier tous les membres du laboratoire pour leur accueil et les discussions échangées. Je tiens tout particulièrement à remercier Slim, Ibrahim, Ta et Sam qui ont rythmé ma vie pendant ces six derniers mois. Je me souviendrais longtemps des parties de go avec Arnaud.

Je souhaite également remercier la promotion M2BI 2016-2017 et plus particulièrement Florence, Lucie et Jaysen. Ils ont été d'un grand soutien tout le long de cette année.

Je voudrais finalement remercier Catherine Etchebest et Jean-Christophe Gelly, ainsi que l'ensemble de l'équipe pédagogique, pour cette année.

Table des matières

Remerciements **1**

Table des figures **3**

Liste des tableaux **3**

1 Introduction **1**

1.1 Une définition de la métagénomique 1
1.2 Problématique 1
1.3 Réseaux de co-occurrences en métagénomique 1
1.4 Méthodes d'inférences et leurs problèmes 2
1.4.1 Par similarité et par régressions 2
1.4.2 Problèmes des méthodes d'inférences par similarité et régression 2
1.4.3 Outils s'appuyant sur les méthodes d'inférences par similarité et régression 3
1.4.4 Autres outils et méthodes 3
1.5 Objectifs 4

2 Matériel et Méthodes **4**

2.1 Outils et méthodes évaluées 4
2.1.1 SparCC 4
2.1.2 REBACCA 6
2.1.3 SPIEC-EASI 6
2.2 Moyens techniques à disposition 7
2.3 Données simulées 7
2.3.1 Modèle de simulation des données 7
2.3.2 Schéma de simulation 8

3 Résultats et Discussion **9**

3.1 Temps d'exécution 9
3.2 Limites en mémoire 10
3.3 Performances des outils pour la reconstruction des réseaux 11
3.3.1 Cas particulier : Petit réseaux denses 12
3.3.2 Réseaux de densité variable 14

4 Conclusion et perspectives **18**

5 Références **20**

Table des figures

1 Exemple de compositionnalité 3
2 Exemple de corrélations directe et indirecte 6
3 Temps d'exécution de SparCC, REBACCA et SPIEC-EASI 10
4 Petit réseaux denses : F-measure 12
5 Petit réseaux denses : Précision 13
6 Petit réseaux denses : Sensibilité 13
7 Petit réseaux denses : Vrai positifs 14
8 Réseaux de densité variable : F-measure 15
9 Réseaux de densité variable : Précision 16
10 Réseaux de densité variable : Sensibilité 17
11 Impact de la densité des réseaux et des réplicats sur le temps d'exécution II

Liste des tableaux

1 Paramètres des simulations utilisés 9

1 Introduction

1.1 Une définition de la métagénomique

La métagénomique consiste à caractériser expérimentalement la globalité d'un écosystème microbien, sans isoler au préalable les différents microorganismes qui le compose. Il est déjà établi que moins de 1 % des microorganismes de la biosphère sont cultivables *in vitro* [1, 2]. De plus, l'analyse par isolation des différents constituants d'un écosystème engendre la perte des relations écologiques entre ces microorganismes. Résultant en une interprétation partielle, voire erronée des dynamiques relationnelles de cet écosystème. Cependant les approches de métagénomiques permettent de s'affranchir de ces limites en permettant d'appréhender un écosystème dans sa globalité.

1.2 Problématique

L'analyse des données métagénomiques soulève de nombreuses questions méthodologiques. Au-delà de la constitution d'un répertoire d'espèces ou de gènes et de l'étude fonctionnelle de la composition de ce répertoire, les analyses s'orientent de plus en plus vers des approches comparatives (données spatialisées, séries temporelles, prise en compte de covariables, ...) et des études d'interactions, ou plus précisément des études d'associations au sein d'un écosystème [3].

Plusieurs méthodes statistiques sont apparues ces dernières années pour détecter des co-occurrences significatives entre espèces d'un même écosystème. Ces méthodes supposent que ces co-occurrences sont indicatives d'interactions écologiques (mutualisme, parasitisme, ...) entre différentes espèces [3]. Elles s'appliquent donc à révéler des interactions écologiques par reconstruction de réseaux de co-occurrences via des procédures statistiques.

1.3 Réseaux de co-occurrences en métagénomique

Les méthodes d'inférence de réseaux de co-occurrences utilisent l'abondance de différents taxa dans différentes conditions expérimentales et/ou réplicats pour générer une matrice de similarité. Cette matrice de similarité indique pour chaque couple de taxa, une mesure de la force de leur interaction. Une fois la matrice obtenue, la significativité des interactions est évaluée. Les interactions significatives sont ensuite représentées en tant qu'arêtes entre les nœuds (taxa) dans une graphe de co-occurrence.

L'abondance de différents taxa dans différentes conditions expérimentales et/ou échantillons est représentée sous forme de table d'abondances. Les échantillons sont en colonnes et les OTUs (taxa) sont en lignes. Un OTU (Unité Taxonomique Opérationnelle) est une pseudo-espèce, elle est créée par regroupement des séquences nucléotidiques ayant une forte similarité (en général $> 97\%$). Ainsi, les individus d'OTUs différents ne proviennent en général pas de la même

espèce.

1.4 Méthodes d'inférences et leurs problèmes

Plusieurs méthodes existent afin d'estimer les interactions entre les différents taxa [4].

1.4.1 Par similarité et par régressions

Par similarités

Le principe est de calculer un score de similarité (une corrélation) pour *chaque paire de taxa possible*. Si ce score est significatif alors il y a une corrélation significative entre ces 2 taxa. La *sparsité* des données impacte grandement les résultats de ce type de méthode.

Des métriques simple telles que le coefficient de Pearson, le coefficient de Spearman ou encore l'indice de similarité de Bray-Curtis permettent de connaître la similarité entre les abondances d'une paire de taxa. Cependant les données de métagénomiques sont **sparses**, c'est-à-dire qu'il y a beaucoup de 0 dans la table d'abondances. En effet, ces métriques accordent de l'importance à des co-absences de taxa (la comparaison de deux 0 donnera un score très élevé). Cela est problématique car un taxon rare sera souvent en très faible abondance et sera donc très difficile à détecter.

Par régressions

L'abondance d'un taxon est estimée par l'abondance combinée de plusieurs autres taxa. Cela permet d'obtenir pour un taxon, les interactions plusieurs autres taxa. Les problèmes à pallier pour ces méthodes sont, le nombre élevé de faux positifs, l'*overfitting* des données et le manque de puissance qui est du à la dimensionnalité des données (beaucoup d'OTUs mais peu d'échantillons).

1.4.2 Problèmes des méthodes d'inférences par similarité et régression

En plus des problèmes de sparsité, d'overfitting et de dimensionnalité des données, un autre problème majeur des données de métagénomiques est à souligner : la compositionnalité. Les profondeurs de séquençage des données de métagénomique peuvent être différentes d'un échantillon à l'autre. Ainsi les abondances utilisées pour l'inférence des réseaux peuvent être biaisés. Par exemple, un nombre élevé de 0 dans une échantillon suite à une faible profondeur de séquençage pour cet échantillon. De plus, les OTUs étant comparés sur différents échantillons, une normalisation est nécessaire pour rendre les abondances comparables d'un échantillon à l'autre. De ce fait, les abondances absolues sont converties en abondances relatives, les rendant dépendantes l'une des autres. Une espèce ayant été comptée le même nombre de fois dans deux échantillons ayant des profondeurs de séquençage différentes aura une proportion différente dans ces échantillons. En particularité, si l'abondance d'un OTU croît et que

tous les autres conservent leur abondance, les OTUs ayant conservé leur abondance d'origine vont tous décroître simultanément du point de vue de l'*abondance relative*.

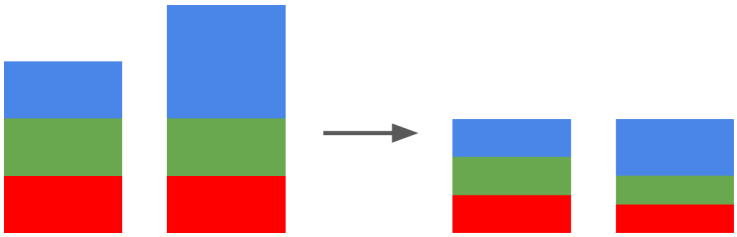


FIGURE 1 – Exemple de compositionnalité. On a ici, deux communautés de 3 OTUs, les abondances absolues sont représentées à gauche de la figure. Une fois que les abondances sont en relatifs (à droite) les OTUs vert et rouge semblent être en moindre abondance alors que ces OTUs sont en abondances identiques dans les deux communautés.

1.4.3 Outils s'appuyant sur les méthodes d'inférences par similarité et régression

La procédure ReBoot de l'outil CoNet [5] réalise des bootstraps afin pallier le problème du manque d'indépendances entre les abondances de taxa. L'outil détecte les relations écologiques en combinant les informations obtenues par différentes métriques.

Des outils adaptés pour les données compositionnelles issues d'écosystèmes microbiens existent également. L'outil SparCC [6] de Friedman *et al.* et l'outil REBACCA [7] de Ban *et al.* réalisent une transformation log-ratio de leurs données (méthode proposée par Aitchison [8]) pour traiter ce problème.

Les méthodes citées précédemment détectent des corrélations directes et indirectes entre différentes espèces. L'outil SPIEC-EASI [11] à la spécificité de ne détecter que les corrélations directes, pour cela les abondances des différentes espèces sont représentés sous la forme de modèle graphique gaussien.

1.4.4 Autres outils et méthodes

Local Similarity Analysis

L'outil LSA [9] est un des outils optimisés pour détecter des associations sensibles au facteur temps. Cette méthode compare les décalages dans les compositions des OTUs dans des séries temporelles.

Théorie des matrices aléatoires

Deng *et al.* ont mit à disposition MENA [10], un framework permettant de reconstituer des réseaux d'associations écologiques en se basant la théorie des matrices aléatoires (RMT), cette méthode se veut robuste aux bruits et a de ne pas nécessiter un seuillage arbitraire.

Une évaluation exhaustive des performances et limites de la plupart de ces méthodes et outils a récemment été réalisée par Weiss *et al.* [12], différents schéma de simulations de données ont été développés, notamment des copules, séries temporelles et écologiques. Les performances de ces outils face aux difficultés spécifiques des données de métagénomique telles que la compositionnalité et la sparsité des données ont aussi été mesurées. Il a été montré que ces différentes méthodes ont des performances très différentes en termes de sensibilité et précision.

1.5 Objectifs

Les objectifs de ce stage sont de réaliser une comparaison des méthodes à fort ancrage statistiques pour inférer des réseaux de co-occurrences à partir de données métagénomiques. Pour cela, un choix d'outils à tester doit être réalisé grâce à un recensement des méthodes dans la littérature. L'évaluation des performances et des limites (temps de calcul, mémoire, passage à l'échelle) de chaque outil doit être réalisé avec un schéma de simulation reflétant la complexité des données de métagénomiques. Le but de ce stage est d'élaborer des recommandations pour chaque outil testé.

2 Matériel et Méthodes

On va évaluer les outils SparCC, REBACCA et SPIEC-EASI.

2.1 Outils et méthodes évaluées

2.1.1 SparCC

L'outil SparCC (*Spar*se *C*orrelations for *C*ompositional data) a été développé afin d'identifier des corrélations entre taxa au sein de communautés écologiques. Pour chaque taxa i, on suppose qu'il existe une abondance de base w_i et on cherche des corrélations entre les abondances de base (en échelle logarithmique) :

$$(w_1, \ldots, w_p) \sim e^{\mathcal{N}(\mu, \mathbf{\Sigma})}$$

Avec $\Sigma = (\sigma_{ij})$ tel que

$$\begin{cases} \sigma_{ij} = \sigma_i^2 & \text{si} \quad i = j \\ \sigma_{ij} = \sigma_i \sigma_j \rho_{ij} & \text{si} \quad i \neq j \end{cases}$$

Pour chaque paire de taxa, il calcule une quantité t_{ij} robuste aux effets de compositionnalité :

$$t_{ij} \equiv Var\left[log\frac{x_i}{x_j}\right] = Var\left[log\frac{w_i}{w_j}\right]$$

x_i et x_j étant les fractions réelles des OTU i et OTU j, elles sont estimées à partir des comptages dans une approche bayésienne en utilisant un à priori de Dirichlet $D(1, \ldots, 1)$. Numériquement, cela revient à remplacer les comptages $(n_1, \ldots, n_p)$ par les pseudo-comptages $(n_1+1, \ldots, n_p+1)$ et à estimer les fractions à partir des pseudo-comptages. La régularisation via les pseudo-comptages permet notamment d'éviter les $log(0)$. Le ratio des fractions réelles est équivalent au ratio des abondances de bases w_i et w_j.

La quantité t_{ij} est liée à la variance des log-ratios des abondances de base des deux OTUs ainsi que leur corrélation via les formules :

$$\begin{aligned} t_{ij} &\equiv Var\left[log\frac{w_i}{w_j}\right] \\ t_{ij} &\equiv Var\big[logw_i\big] + Var\big[logw_j\big] - 2Cov\big[logw_i, logw_j\big] \\ t_{ij} &\equiv \sigma_i^2 + \sigma_j^2 - 2\rho_{ij}\sigma_i\sigma_j \end{aligned} \tag{1}$$

L'estimation de la matrice de variances-covariances de base $\Sigma = (\sigma_{ij})$ permet d'obtenir les variances de bases σ_i^2 et σ_j^2, et surtout, les corrélations ρ_{ij} nécessaires pour la reconstruction des réseaux. Cependant, il est impossible d'inférer directement ces covariances de base car il y a plus de variables que d'équations.

Pour contrer ce problème, SparCC calcule une approximation des variances de base σ_i^2 en supposant que le réseau à reconstruire est large (grand nombre d'OTUs) et que la majorité des taxa du réseau sont faiblement corrélés entre eux (hypothèse de sparsité avec $\rho_{ij} \simeq 0$). Il suffit d'injecter ces σ_i^2 et σ_j^2 dans les équations (1) pour estimer t_{ij}.

L'inférence du réseau entier se fait ensuite par un nombre fixé d'itération. À chaque itération, les corrélations sont estimées et les paires d'OTUs fortement corrélés sont écartées. Les approximations précédentes supposent d'avoir au moins 4 OTUs.

SparCC est implémenté en Python et l'inférence d'un réseau est réalisé de cette façon :

1. **Calcul des corrélations** : Pour un nombre fixé d'itérations, la corrélation de chaque paire non écartée d'OTUs est calculée à chaque itération. La corrélation finale est la médiane de la distribution des corrélations.
2. **Génération des bootstraps** [1] : Les données sont rééchantillonnés plusieurs fois, pour pouvoir calculer une p-value par la suite, et le même procédé de calcul des corrélations est appliqué à chaque jeu.
3. **Calcul des p-values** : Les p-values sont calculées à partir de la distribution des valeurs de bootstraps. Elles correspondent à la proportion des corrélations *bootstraps* étant au moins aussi grande que la corrélation calculée sur les données d'origine.

1. Création de nouveaux échantillons par mélange (tirages avec remise) des valeurs de l'échantillon d'origine

2.1.2 REBACCA

L'outil REBACCA (*R*egularized *e*stimation of the b*a*sis *c*ovariance based on *c*ompositional d*a*ta) identifie aussi des corrélations entre taxa au sein de communautés écologiques. Tout comme SparCC, il estime les covariances de base des paires d'OTUs afin d'obtenir leurs corrélations.

Pour pallier le problème de compositionnalité des données, REBACCA réalise aussi une transformation log-ratio. Mais contrairement à SparCC, pour éviter les $log(0)$, REBACCA rajoute une valeur équivalente à 1/10$^{\text{ième}}$ du minimum des valeurs non nulles aux valeurs nulles au lieu de faire un pseudo-comptage de l'ensemble des valeurs.

REBACCA construit un système de $\frac{D(D-1)}{2} - D$ équations linéaires avec $\frac{D(D-1)}{2}$ variables à partir des log-ratios (D étant le nombre total d'OTUs). Le nombre de variables étant inférieure au nombre d'équations, il est impossible de résoudre le système directement. Ainsi, pour le résoudre, REBACCA utilise la méthode LASSO en supposant que le réseau est sparse : *Chaque OTU interagit avec moins de* $\frac{D}{4}$ *autres OTUs*.

La régularisation LASSO permet de rendre le système stable et de le résoudre efficacement.

2.1.3 SPIEC-EASI

L'outil SPIEC-EASI (*Sp*arse *I*nverse *C*ovariance Estimation for *E*cological *Ass*ociation *I*nference) permet d'inférer des réseaux d'interactions écologiques microbiennes à partir de données de séquençage 16S (détails dans l'annexe). Les corrélations détectées par cet outil sont des interactions directes entre deux OTUs. En effet, les outils tels que SparCC et REBACCA ne font pas la distinction entre corrélations directes et indirectes.

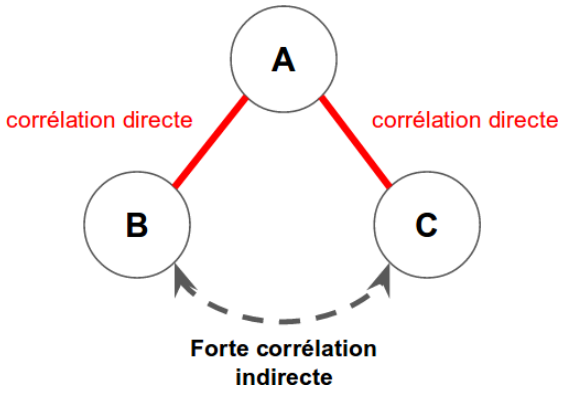


FIGURE 2 – Exemple de corrélations directe et indirecte. L'OTU A interagit avec les OTUs B et C (corrélations directes). Les OTUs B et C sont fortement corrélés, via leur dépendance à l'OTU A, mais ne sont pas en interaction directe (corrélation indirecte).

Tout comme les deux précédents outils, le problème de compositionnalité est levé par une transformation des données, ici un log-ratio centré. Deux méthodes sont ensuite implémentées pour l'estimation d'un réseau d'interactions sparse.

L'algorithme MB (**M**einshausen and **B**ühlmann) qui consiste en une sélection de voisinages

(par régression) et glasso (Graphical LASSO) qui permet d'estimer directement l'inverse de la matrice de variance-covariance.

La sparsité du réseau est inférée par sous-échantillonnages aléatoires et constitue une mesure de la stabilité des interactions. L'outil produit directement une matrice de covariances régularisée qui correspond à la matrice de similarité (son inverse correspond à la matrice d'adjacence du réseau d'interactions).

2.2 Moyens techniques à disposition

Les différents outils testés ont été intégrés à la plateforme MIGALE. Les traitements de fichiers ont été réalisés d'une part via le serveur d'entrée de MIGALE et d'autre part via mon poste de travail personnel (système d'exploitation Ubuntu 16.04 LTS).

Les calculs ont été réalisés par multithread (4 threads) et ont été réparti sur l'ensemble des 53 nœuds des clusters de calcul de la plateforme MIGALE, représentant ainsi un total de 628 processeurs de générations différent dont la fréquence varie de 2,2 à 2,8 GHz. Chaque processeur est équipé de multi-processeurs Intel ou AMD. Les ressources du cluster sont exploitables via la couche logicielle Sun Grid Engine (SGE) avec un accès en ligne de commande.

2.3 Données simulées

2.3.1 Modèle de simulation des données

Le modèle utilisé pour simuler des données de composition de communautés reflétant des corrélations entre taxa est le même que celui utilisé par les 3 outils évalués SparCC, REBACCA et SPIEC-EASI pour l'inférence.

On suppose qu'on a n communautés (indexées par $i = 1 \dots n$) et p espèces (indexées par $j = 1 \dots p$). Il peut s'exprimer de façon hiérarchique comme suit :

- Les abondances dans la communauté i sont gouvernées par des abondances de bases Z_i qui sont échantillonnées suivant une loi normale multivariée $\mathcal{N}(\mu, \mathbf{\Sigma})$:

$$Z_i \sim \mathcal{N}(\mu, \mathbf{\Sigma})$$

- L'abondance relative (ou proportion) p_{ij} du taxa j dans l'échantillon i se déduit de Z_i par une transformation logistique :

$$p_{ij} = \frac{e^{Z_{ij}}}{\sum_{k=1} e^{Z_{ik}}}$$

- Les comptages $\mathbf{n}_i = (n_{i1}, \dots, n_{ip})$ des différents taxa dans l'échantillon i suivent alors une loi multinomiale de paramètre N_i (profondeur de séquençage) et $\mathbf{p_i} = (p_{i1}, \dots, p_{ip})$.

$$\mathbf{n}_i \sim \mathcal{M}(N_i, \mathbf{p_i})$$

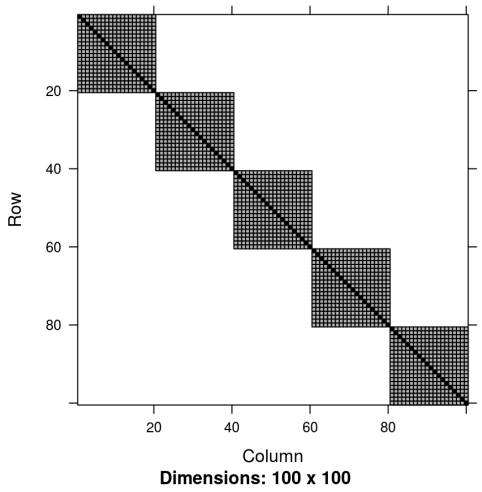
2.3.2 Schéma de simulation

La matrice que SparCC et REBACCA estime est la matrice de variance-covariance des abondances de base Σ, quant à SPIEC-EASI, il s'agit de son inverse Ω. On génère cette matrice à estimer en faisant varier les paramètres :

- p : Le nombre d'OTUs qui permet d'évaluer les performances des outils à reconstruire des réseaux de taille variable, ainsi que leurs passages à l'échelle.
- μ : La gamme des abondances des OTUs qui permet d'évaluer la capacité des outils à détecter des taxa rares. Ici μ correspond à l'ordre de grandeur sur lequel les abondances varient.
- ρ : La force moyenne des corrélations qui permet d'évaluer la puissance des outils pour détecter des interactions faibles entre taxa.
- k : La matrice générée est diagonale par blocs avec k blocs. Le nombre de blocs permet d'évaluer la robustesse des outils aux données non sparses. Le nombre de blocs dépend de p.

Puis on génère, suivant notre modèle de simulation, une table de comptage avec un nombre n d'échantillons, chaque échantillon a 3000 reads de profondeur de séquençage.

Pour illustrer cette matrice, voici un exemple avec 100 OTUs ($\mathbf{p} = 100$), les OTUs sont réparti aléatoirement dans 5 blocs ($\mathbf{k} = 5$) de taille identique (25 OTUs par blocs). L'abondance de base moyenne est la même pour tous les OTUs ($\mu = 0$). Et finalement, les OTUs d'un même bloc sont tous corrélés avec la même corrélation de $\rho = 0.5$.



Un ensemble de 4752 simulations sont effectuées pour chaque outil et afin de connaître la variabilité des résultats, chaque jeu de simulation est répété 5 fois (23760 simulations au total pour un outil). Les paramètres des simulations sont résumé dans le tableau ci-dessous :

Chaque simulation a été réalisé en multithread (4 threads). Pour SparCC 100 bootstraps ont été effectués pour l'obtention des p-values. Le LASSO de REBACCA a été effectué avec 100

Paramètres	Valeurs				
p	5, 20, 50, 100, 200				
n	10, 20, 30, 40, 50, 100, 200, 500				
mu	0, 1, 2, 3, 4, 5				
rho	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9				
blocks	p=5	p=20	p=50	p=100	p=200
	1	1, 2	2, 5	2, 5, 10	2, 5, 10, 50

TABLE 1 – Paramètres des simulations utilisés

bootstraps également (correspondant à $nbootstraps = 50$). Quant à SPIEC-EASI, 100 sous-échantillonnages ont été effectué pour la sélection de λ avec la méthode *stars* ; le reste des paramètres reste fixé aux valeurs par défaut.

3 Résultats et Discussion

3.1 Temps d'exécution

Les temps d'exécution sont faibles pour de petits réseaux (moins de 4 minutes pour $p = 5 \dots 50$) avec les outils REBACCA et SPIEC-EASI, REBACCA ayant les meilleurs temps (une ou deux minutes voire quelques secondes). Les gammes des abondances μ et le nombre d'échantillons n n'impactent pas le temps d'exécution de ces 3 outils. Contrairement à SparCC dont le temps d'exécution double entre $n = 100$ (5 minutes) et $n = 500$ (plus de 10 minutes).

Pour de plus grands réseaux ($p = 100 \dots 500$), les temps d'exécution de la méthode *glasso* de SPIEC-EASI sont considérablement allongés lorsque la gamme des abondances est très large ($\mu = 3 \dots 5$) et que le nombre d'échantillons est faible. Cela est dû à un manque de puissance de l'outil pour détecter correctement les corrélations car le nombre d'échantillons disponibles n est très faible pour un nombre d'espèces p très élevé (cas de simulations "underpowered"). SparCC réalise les meilleurs temps pour les large réseaux (moins de 20 minutes pour 200 OTUs). Les temps d'exécution de REBACCA sont très impactés par le nombre d'espèces dans le cas de large réseaux, avec moins de 8 minutes pour 100 OTUs pour près de 3 heures avec le double d'OTUs.

Les forces des interactions (corrélations) des différentes espèces du réseau et la sparsité de celui-ci n'impactent pas les temps d'exécution des 3 outils. Si on s'intéresse aux temps d'exécution pour chaque réplique (excepté la première qui n'a pas été réalisée en multithread) on observe que la variabilité de ces temps est minime d'une réplique à l'autre pour les 3 outils aussi (figure en annexe).

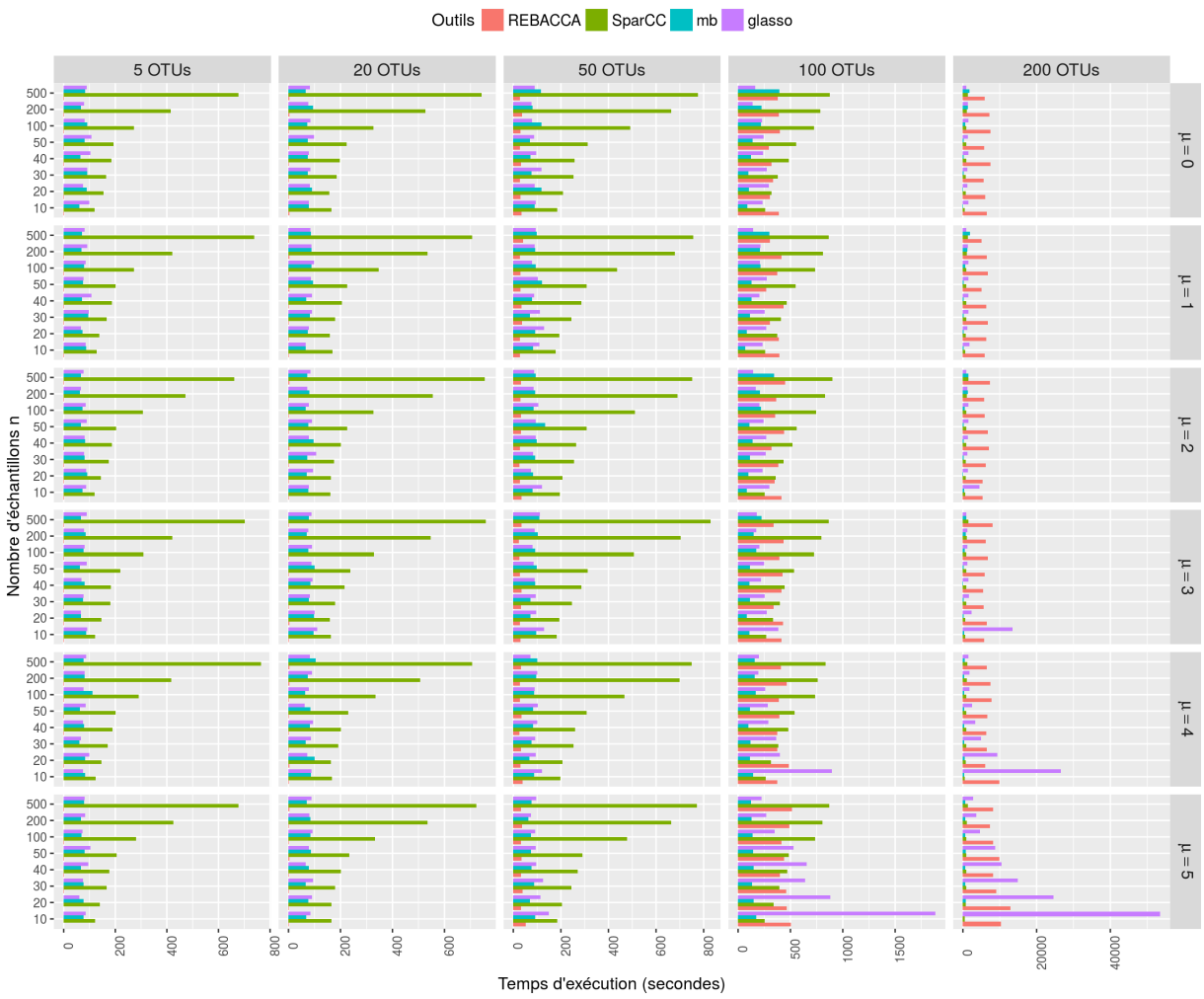


FIGURE 3 – Temps d'exécution des outils SparCC, REBACCA et des méthodes glasso et MB de l'outil SPIEC-EASI. Les outils REBACCA et SPIEC-EASI ont des temps d'exécution faibles pour de petit réseaux. La méthode glasso de SPIEC-EASI est confronté à un manque de puissance pour détecter les corrélations lorsque le nombre d'échantillon est trop faible pour un réseau large, ce qui impacte considérablement son temps d'exécution (plus de 13 heures pour la simulation la plus difficile). SparCC qui est légèrement plus lent par rapport aux autres outils sur des petit réseaux, possède les meilleurs temps pour de large réseaux.

3.2 Limites en mémoire

La limite en mémoire des outils a été déterminé en faisant varier le nombre p d'OTUs dans les tables de comptages.

Pour les outils SparCC et REBACCA, des tables de comptages ayant jusqu'à 2000 OTUs ont été testés. L'outil SparCC utilise un algorithme optimisé en temps jusqu'à 1500 OTUs, au-delà, un algorithme plus lent mais optimisé en mémoire est utilisé. À l'inverse, REBACCA est dans l'impossibilité de calculer des corrélations au-delà de 250 OTUs, cela est dû à une implémentation sous R très coûteuse en mémoire (complexité en $O(D^4)$ où D est le nombre d'OTUs).

Concernant SPIEC-EASI, la méthode Graphical LASSO ne provoque pas de problèmes de mémoire, cependant cette limite n'a pas été testée pour des tables de comptages de plus de 200 OTUs. La méthode MB rencontre par contre de nombreux problèmes, 438 simulations n'ont pas abouties à cause de problèmes de mémoire sur 23760 du schéma de simulation :

- Pour les simulations les plus difficiles — maximum d'OTUs $p = 200$, minimum d'échantillons $n = 10$ et gamme d'abondance la plus étendue $\mu = 5$ — aucune des 180 simulations n'a aboutie (36 simulations par répétitions).
- Pour une gamme d'abondance moins étendue ($\mu = 4$), 150 simulations correspondants à 30 simulations par répliques n'ont pas abouties.
- Des cas de difficulté modérée : $p = 100$ avec $n = 10$ ou $p = 200$ avec $n = 20$ voire $p = 100$ avec $n = 20$ ont put aboutir dans certaines répliques et ce avec une gamme d'abondance étendue $\mu = 2 \dots 5$.

Ainsi, il semblerait que le manque de puissance nécessaire pour détecter les taxa rares est compensé par une utilisation étendue de la mémoire.

3.3 Performances des outils pour la reconstruction des réseaux

Les performances des différents outils ont été évaluées avec 4 mesures dont la valeur varie dans $[0, 1]$:

		Vrai réseau	
		Interaction	Pas d'interaction
Présumé par l'outil	Interaction	**V**rai **P**ositif	**F**aux **P**ositif
	Pas d'interaction	**F**aux **N**égatif	**V**rai **N**égatif
		Positifs	**N**égatifs

- **Sensibilité** qui correspond à la proportion des interactions du réseau ayant été détecté par l'outil :

$$\frac{VP}{P}$$

- **F-measure** qui combine la sensibilité et la précision. Une F-measure de 1 correspond à une sensibilité et précision parfaite alors qu'une F-measure signifie que l'une des métriques est à 0 :

$$\frac{VP}{VP + FN}$$

- **Précision** qui correspond à la proportion de véritables interactions détectées parmi ce qui a été défini comme des interactions par l'outil :

$$\frac{VP}{VP + FP}$$

- **Spécificité** qui correspond à la proportion de "non interactions" ayant été correctement défini par l'outil (absence de corrélations pour les "non interactions") :

$$\frac{VN}{N}$$

3.3.1 Cas particulier : Petit réseaux denses

Les simulations avec $p = 5$ correspondent à des matrices Σ et Ω "pleines", c'est-à-dire que chaque OTU est en interaction avec tout les autres OTUs du réseau. Dans ce cas particulier, l'hypothèse de sparsité – Un OTU interagit avec moins de $\frac{1}{4}$ des OTUs du réseau – est violée.

Les F-measure obtenues pour ces simulations sont assez faibles (figure 3), avec moins de 0.25 pour les outils REBACCA et SPIEC-EASI. SparCC atteint néanmoins 0.5 pour des forces de corrélations moyennes de 0.6. Cela s'explique par une très bonne précision (environ 0.75, figure 5) mais une sensibilité mauvaise (moins de 0.5, figure 6). Pour la force de corrélation la plus forte (à 0.9), les performances de SparCC rejoignent celles des autres outils (moins de 0.25). Cette baisse de la F-measure peut être dû à la violation des conditions pour l'estimation des corrélations de SparCC (les corrélations doivent être faibles).

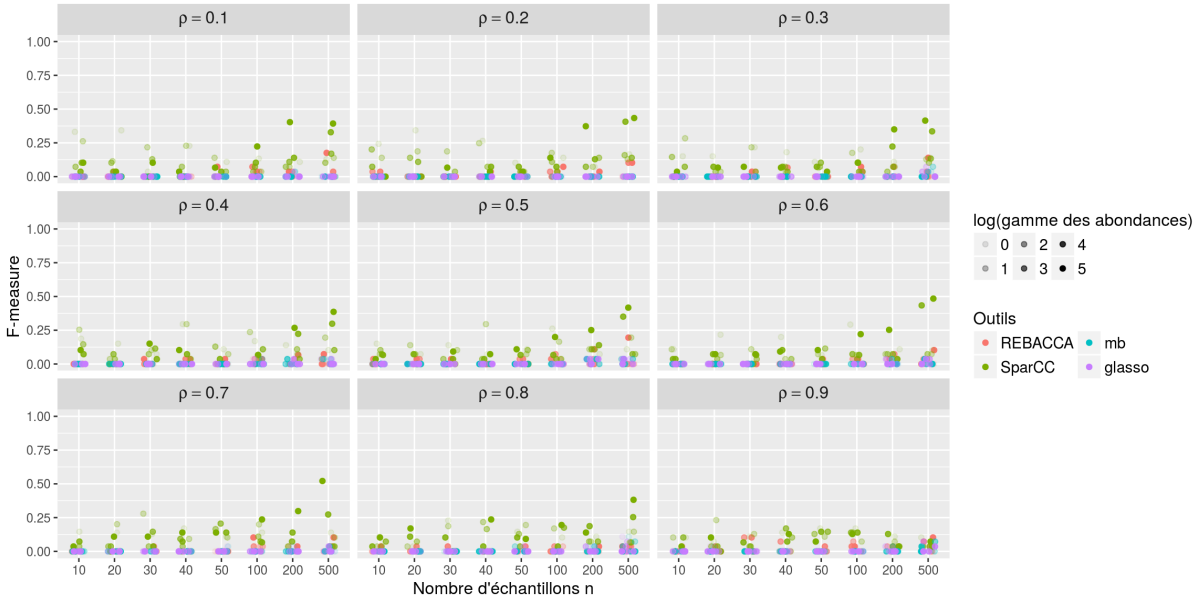


FIGURE 4 – F-measure des outils SparCC, REBACCA et SPIEC-EASI pour des petit réseaux denses. Tous les outils ont de très mauvaises F-measures pour ce type de petit réseaux denses. Les méthodes de SPIEC-EASI sont particulièrement mauvais avec une F-measure de 0 ce qui pourrait vouloir dire que ces méthodes ont une précision ou sensibilité nulle pour ce type de simulation. On obtient les meilleurs résultats avec SparCC, cependant ces F-measures ne dépassent pas les 0.5 et ce même avec un nombre élevé d'échantillons.

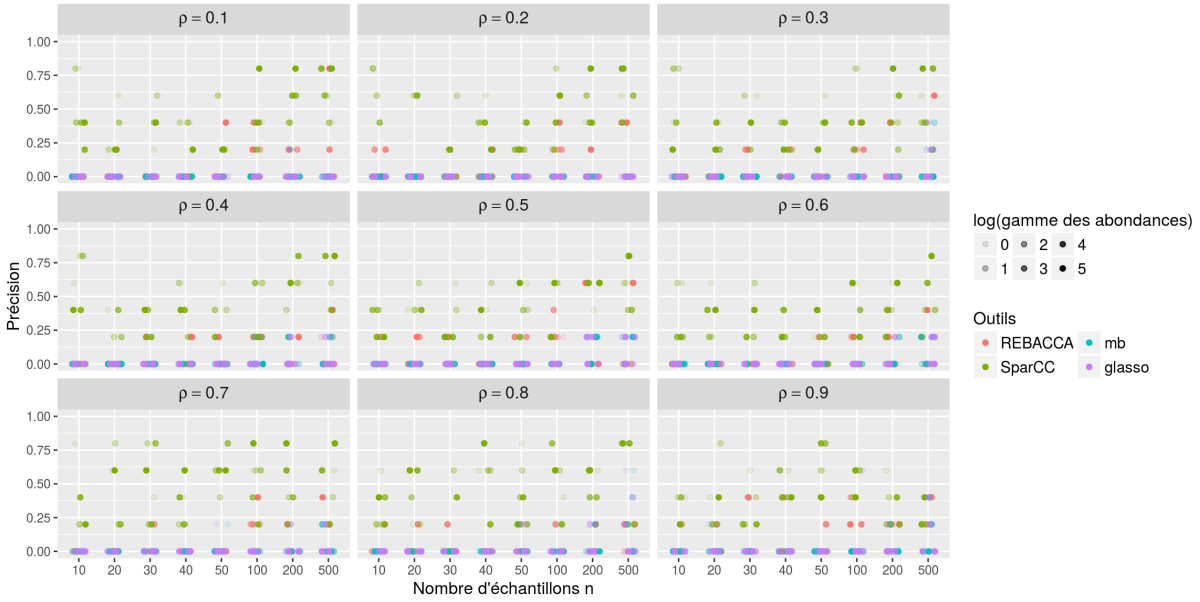


FIGURE 5 – Précision des outils SparCC, REBACCA et SPIEC-EASI pour des petit réseaux denses. Globalement, les méthodes de SPIEC-EASI ne détectent aucunes interactions du réseau (précision nulle). Cependant cette précision peut augmenter légèrement lorsque les interactions sont facilement détectable (grande force de corrélation $\rho = 0.9$ et plus grande puissance $n = 200 \dots 500$). L'hypothèse de sparsité de REBACCA est violée avec ce type de réseaux, les résultats de REBACCA sont inconsistants, cela est dû au fait que l'outil arrive à détecter de vrais interactions au hasard. SparCC possède les meilleurs résultats avec une précision supérieure à 0.75.

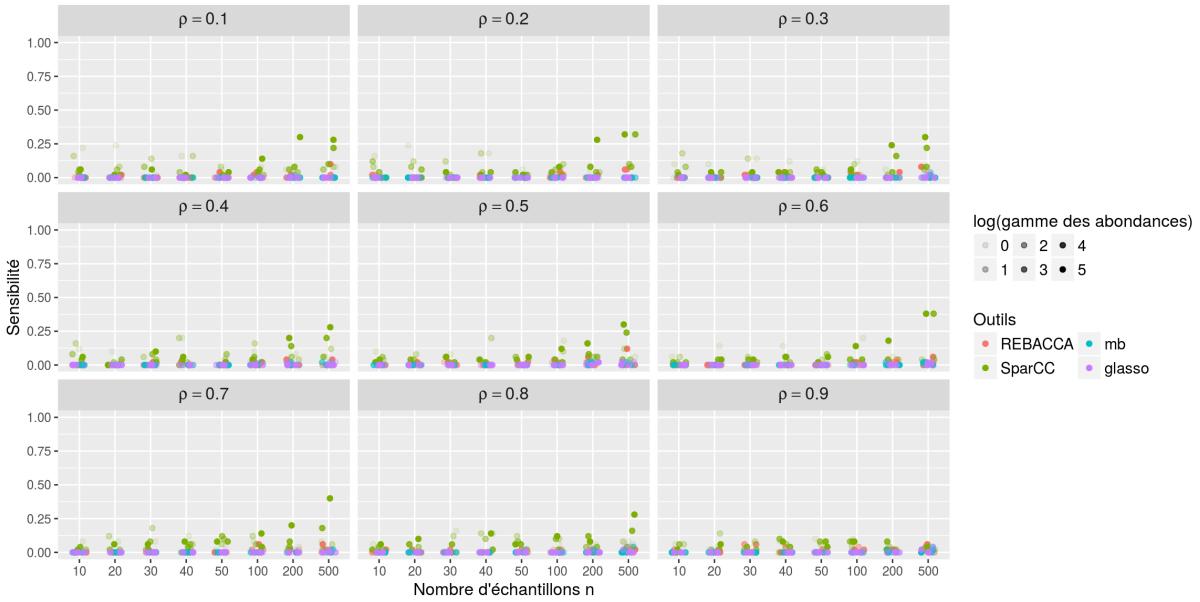


FIGURE 6 – Sensibilité des outils SparCC, REBACCA et SPIEC-EASI pour des petit réseaux denses. Tous les outils ont de très mauvaises précisions pour ce type de petit réseaux denses. Les méthodes de SPIEC-EASI sont particulièrement mauvais avec une sensibilité de 0 ce qui veut dire que SPIEC-EASI est incapable de reconstruire ce type de réseaux (car il ne détecte aucune des interactions de ceux-ci). On obtient les meilleurs résultats avec SparCC, cependant ces sensibilités ne dépassent pas les 0.5 et ce même avec un nombre élevé d'échantillons.

Moins de la moitié des corrélations ont été détectés pour les réseaux de 5 OTUs (10 corrélations possibles, figure 7). Globalement, aucun outils ne peut détecter correctement les corrélations de

ce réseau très dense et de très petite taille, et ce même si le nombre d'échantillon est très élevé.

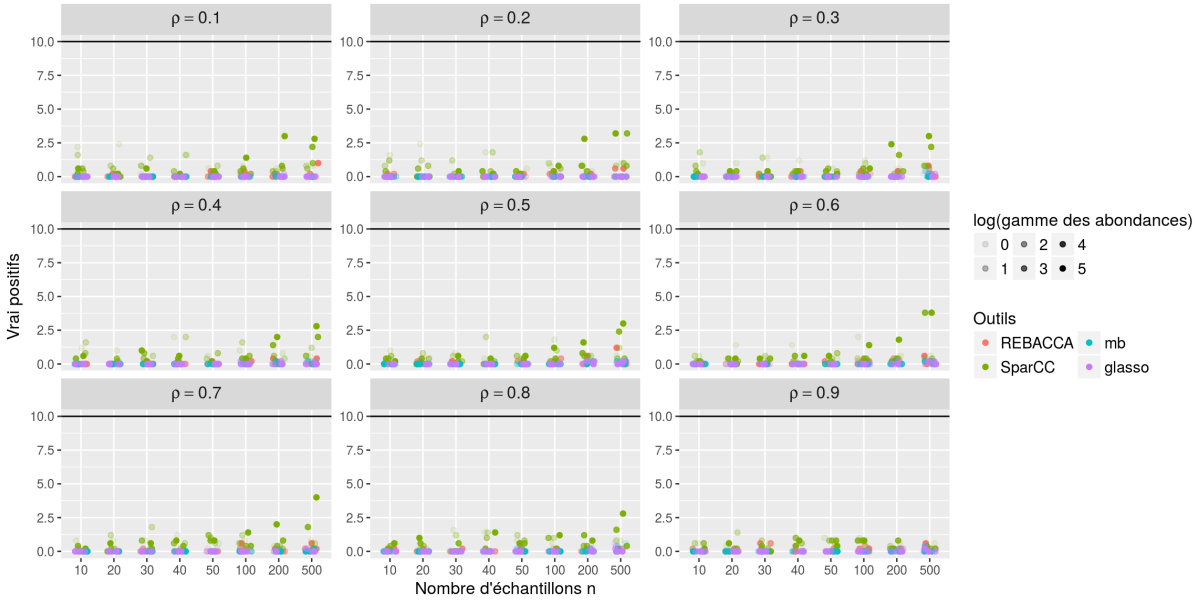


FIGURE 7 – Nombre de vrai positifs obtenus avec les outils SparCC, REBACCA et SPIEC-EASI pour des petit réseaux denses. Le nombre de vrai positifs possible (10) est représenté par la barre noire. Moins de la moitié des interactions ont été correctement détectées.

3.3.2 Réseaux de densité variable

L'effet de la sparsité des réseaux est paramétré dans nos simulations par le nombre de blocs dans les matrices de variance-covariance Σ et de précision Ω. Plus ce nombre est élevé et plus le réseau est sparse.

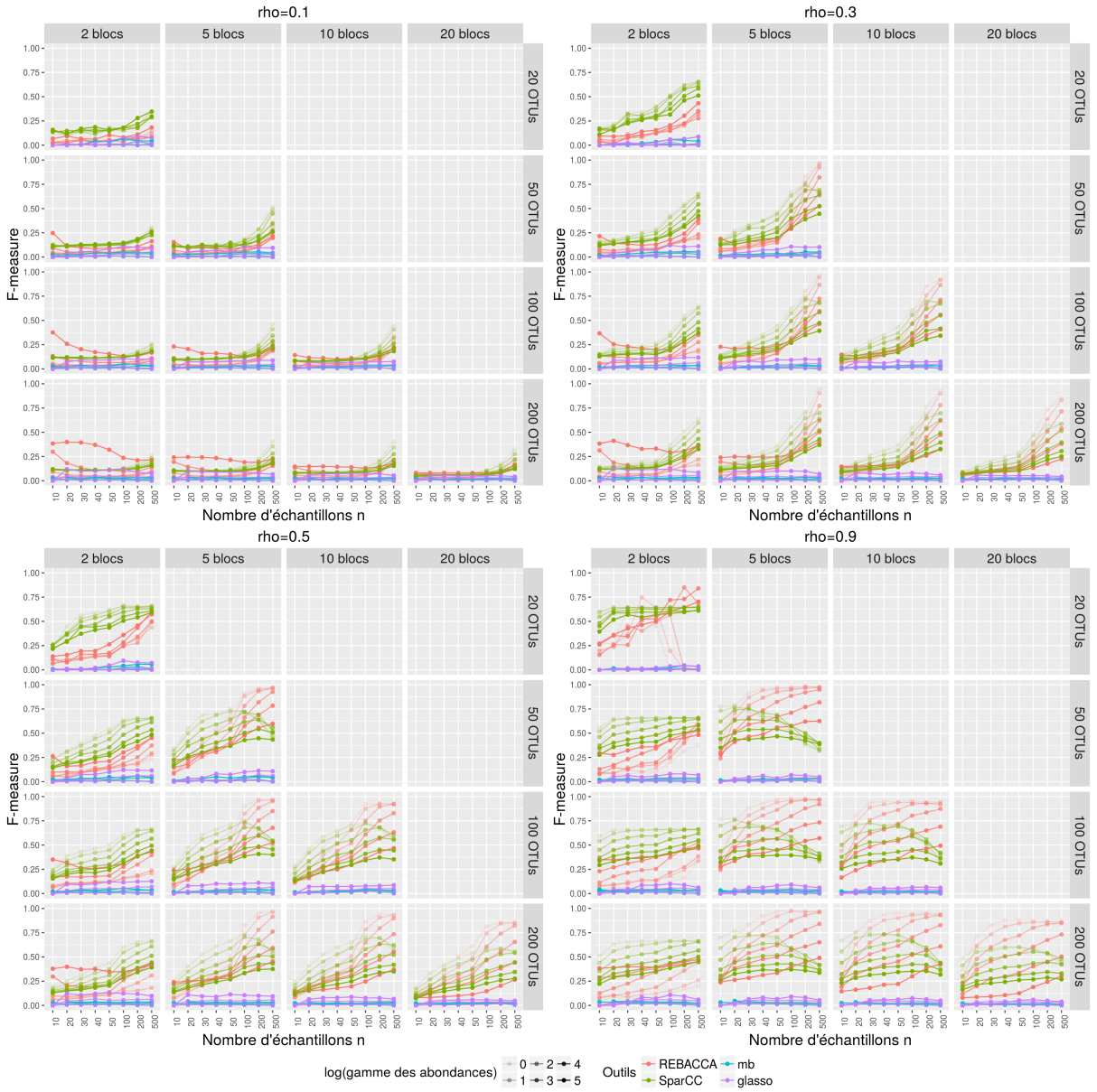


FIGURE 8 – F-measure des outils SparCC, REBACCA et SPIEC-EASI pour les simulations avec $p = 20 \ldots 200$. Lorsque l'hypothèse de sparsité de REBACCA n'est pas violée (simulations ayant au moins 5 blocs), l'outil possède les meilleures F-measures. Pour des forces de corrélations très faibles ($\rho = 0.1$), les 3 outils sont tous mauvais. Pour $rho = 0.3$ les performances de SparCC et REBACCA sont semblables. Les méthodes de SPIEC-EASI sont mauvais dans chaque simulation.

Lorsque la force des corrélations moyenne est très faible ($\rho = 0.1$), tous les outils ont des F-measures en-dessous de 0.5. Les corrélations de cette force sont très difficiles à détecter et cela se traduit par des sensibilités très basses. De plus parmi les faible nombre de corrélations détecté, très peu correspondent à de véritable interactions (faible précision aussi).

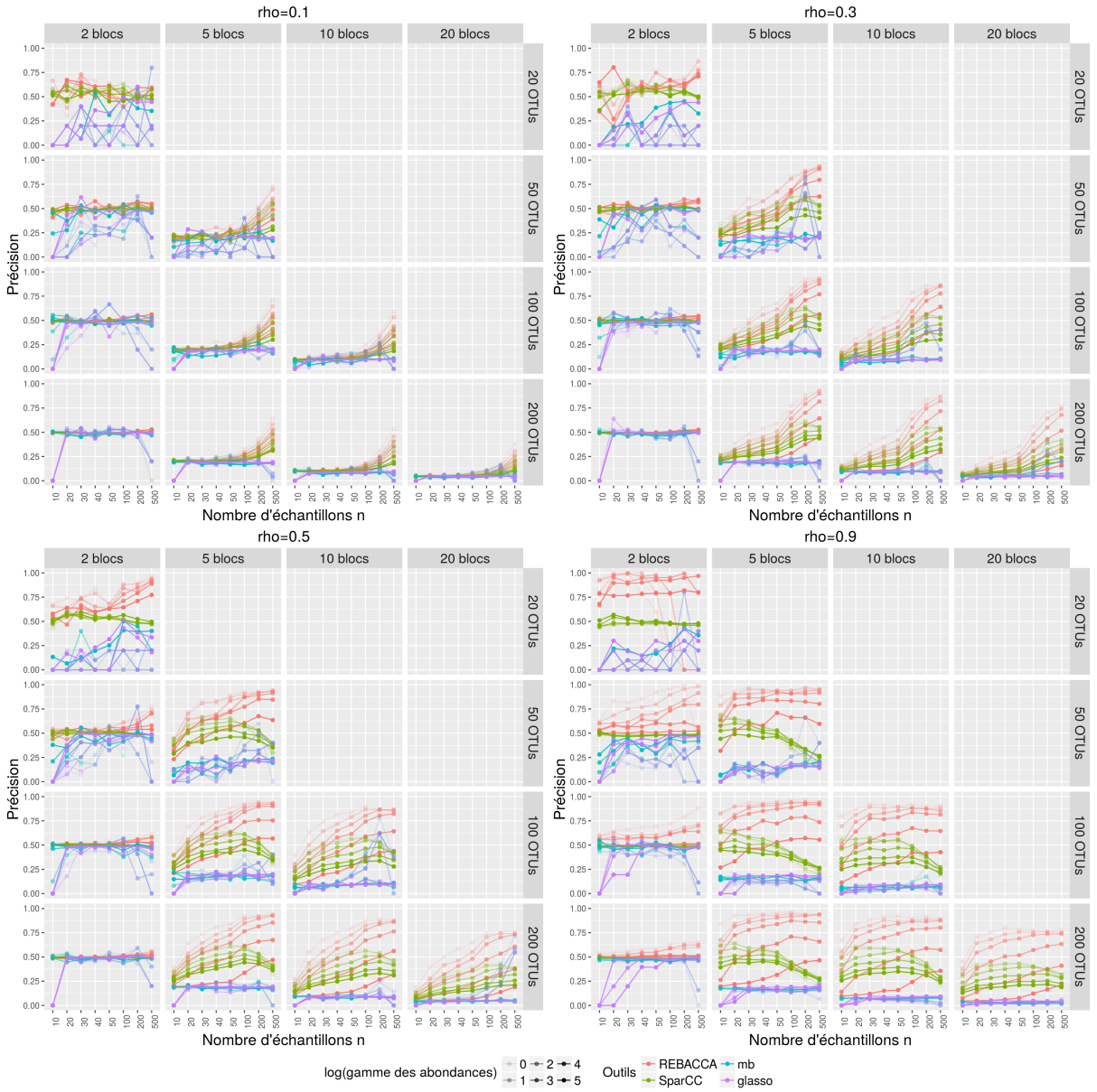


FIGURE 9 – Précision des outils SparCC, REBACCA et SPIEC-EASI pour les simulations avec $p = 20 \ldots 200$. Lorsque l'hypothèse de sparsité de REBACCA n'est pas violée (simulations ayant au moins 5 blocs), l'outil possède les meilleures précisions. Pour des forces de corrélations très faibles ($\rho = 0.1$), les 3 outils sont tous mauvais. Lorsque la densité du réseau augmente (2 blocs), les résultats des méthodes de SPIEC-EASI sont aléatoires.

L'outil REBACCA a une très bonne précision, avec des valeurs supérieures à 0.75. Cette précision diminue lorsque les corrélations à détecter sont faibles, mais un nombre élevé d'échantillons peut augmenter la puissance de l'outil (simulations avec $\rho = 0.3$ et $n = 200 \ldots 500$). La violation de l'hypothèse de sparsité de REBACCA impacte les performances de l'outil en sensibilité, en effet toutes les simulations avec moins de 5 blocs ont des sensibilités inférieures à 0.5.

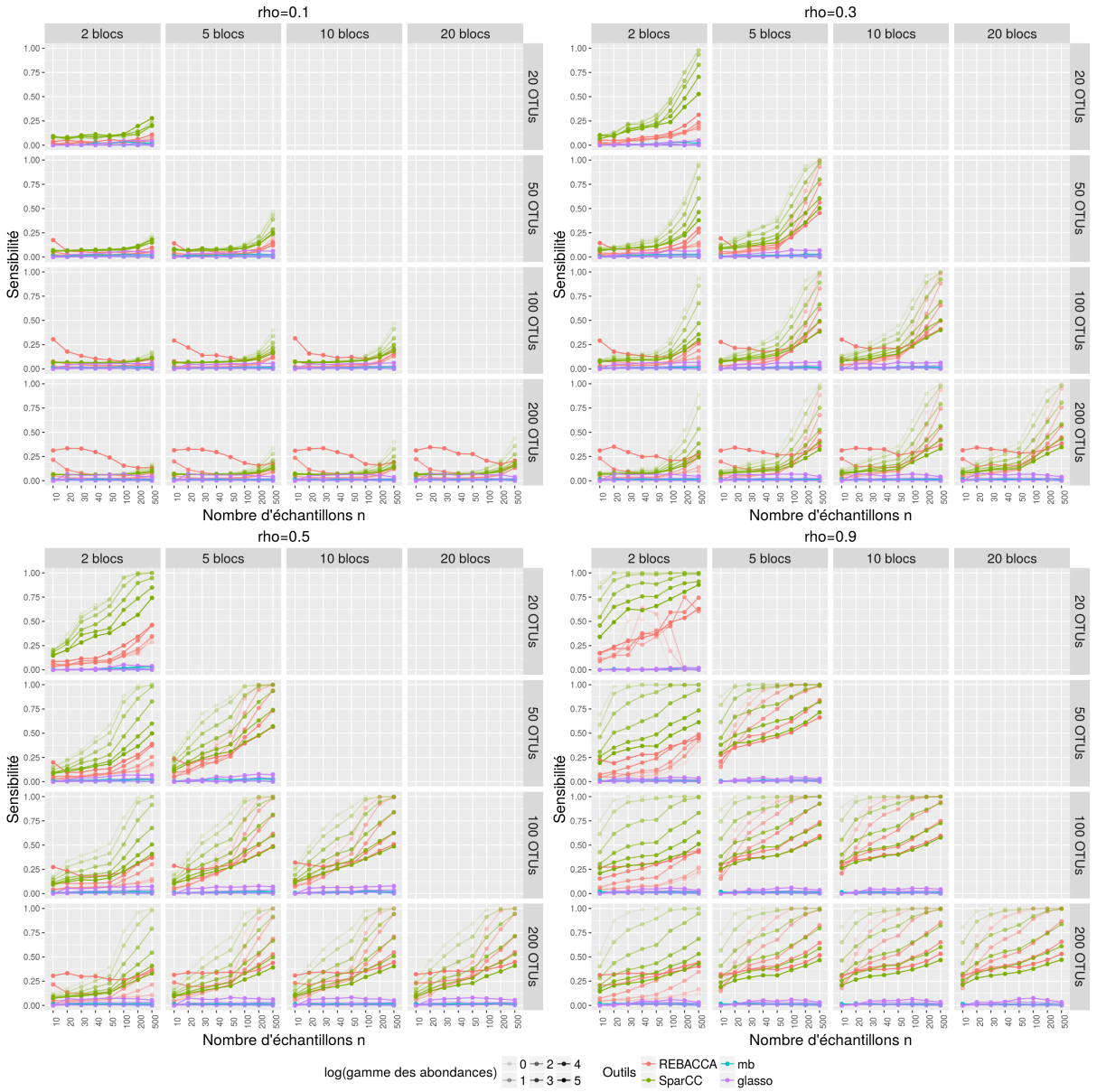


FIGURE 10 – Sensibilité des outils SparCC, REBACCA et SPIEC-EASI pour les simulations avec $p = 20 \ldots 200$. L'outil SparCC possède les meilleures sensibilités. Lorsque l'hypothèse de sparsité de REBACCA n'est pas violée (simulations ayant au moins 5 blocs), l'outil REBACCA a des performances semblablent à SparCC. La sensibilité de ces 2 outils est très dépendantes de la force des corrélations. À $\rho = 0.3$ ces corrélations sont trop faibles être détectées par ces outils mais le nombre d'échantillons peut pallier à ce problème. Les résultats des méthodes de SPIEC-EASI sont par contre très mauvais avec une sensibilité proche de 0.

L'outil SparCC réalise les meilleures performances en sensibilité et ces performances ne sont dépendantes que de la force des corrélations.

Les méthodes glasso et MB sont très peu performants sur ces schéma de simulations. Au maximum, 50% des interactions détectées par ces outils sont détectées correctement et cette précision diminue d'autant plus lorsque le réseau est sparse. De plus, leurs sensibilités sont trop proches de 0.

4 Conclusion et perspectives

L'outil REBACCA privilégie la précision et la spécificité au détriment de la sensibilité. Si la gamme des abondances n'est pas trop étendue ($\mu = 0 \dots 2$), que le réseau est très sparse et que le nombre d'échantillons est élevé ($n = 100 \dots 500$), on obtient une sensibilité proche de 1. Communément, environ une centaine d'échantillons sont disponibles dans les données de métagénomiques mais le nombre d'OTUs excède fréquemment les milliers. Or REBACCA est dans l'impossibilité de reconstruire des réseaux de plus de 250 OTUs. Afin de contourner cette limite, des méthodes de filtrage des OTUs rares peut être envisagé. Ces méthodes sont déjà utilisé pour des jeux de données de très grande taille afin de limiter les temps de calculs nécessaires pour le reconstruction des réseaux et la visualisation de ceux-ci. Cependant la suppression d'OTUs du jeu de donnée entraîne une déstabilisation de la structure du réseau reconstruit. Pour pallier ce problème, l'agglomération d'OTUs en méta-OTUs peut être envisagé. Par exemple, les OTUs pourraient être regroupé selon leur assignation taxonomique. Un pathogène d'intérêt pourrait correspondre à plusieurs OTUs dont seule l'assignation taxonomique au niveau *Espèce* diffère. Ainsi il serait intéressant de regrouper ces OTUs au niveau *Genre* pour avoir une seule méta-OTU représentante du pathogène.

De ce fait, il aurait été intéressant de d'évaluer les outils sur des réseaux plus large et de tester les différents filtres mentionnés précédemment.

L'outil SparCC, contrairement à REBACCA, privilégie la sensibilité au détriment de la précision. Dans le cas de corrélations faibles voire moyenne ($\rho = 0.3 \dots 0.5$), si la gamme des abondances n'est pas trop étendue ($\mu = 0 \dots 2$), la sensibilité est proche de 1 pour un nombre d'échantillons élevé ($n = 100 \dots 500$) et ce quelle que soit la sparsité du réseau (excepté les réseaux de densité maximale).

Globalement, REBACCA reconstruit des réseaux en minimisant les "fausses" interactions cependant les réseaux seront souvent partiels. Alors que SparCC essaye de capturer la totalité des interactions existantes mais le nombre de "fausses" interactions est donc très supérieur.

REBACCA étant sensible à la violation de l'hypothèse de sparsité

Le modèle proposé par les auteurs de SPIEC-EASI est un modèle copules, celui-ci n'utilise pas de multinomiale pour simuler les profondeurs de séquençage contrairement à notre modèle qui reprend le modèle d'inférence. Afin d'obtenir des performances comparables à ceux obtenus par les auteurs de SPIEC-EASI dans leur publication, il aurait été nécessaire de simuler sous leur modèle copules. Cette différence joue vraisemblablement sur les performances des méthodes de SPIEC-EASI.

Afin d'évaluer les performances de SPIEC-EASI il aurait fallu essayer d'autres modèles de simulation des données en plus de celui présentement utilisé. En effet, de nombreux modèles différents existent et le croisement des informations apportées aurait permit une vision plus globale des performances des outils. Notamment, un modèle simulant des relations écologiques entre OTUs aurait pu être envisagé. Aucun des modèles utilisés pour évaluer les performances

des 3 outils ne gère l'influence de co-variables (variabilité dû à la différence de conditions expérimentales), ainsi il aurait été intéressant d'évaluer ce point.

5 Références

[1] J. T. Staley and a. A. Konopka, "Measurement of in Situ Activities of Nonphotosynthetic Microorganisms in Aquatic and Terrestrial Habitats," *Annual Review of Microbiology*, vol. 39, no. 1, pp. 321–346, 1985.

[2] W. Wade, "Unculturable bacteria—the uncharacterized organisms that cause oral infections," *Journal of the Royal Society of Medicine*, vol. 95, pp. 81–83, Feb. 2002.

[3] G. Lima-Mendez, K. Faust, N. Henry, J. Decelle, S. Colin, F. Carcillo, S. Chaffron, J. C. Ignacio-Espinosa, S. Roux, F. Vincent, L. Bittner, Y. Darzi, J. Wang, S. Audic, L. Berline, G. Bontempi, A. M. Cabello, L. Coppola, F. M. Cornejo-Castillo, F. d'Ovidio, L. D. Meester, I. Ferrera, M.-J. Garet-Delmas, L. Guidi, E. Lara, S. Pesant, M. Royo-Llonch, G. Salazar, P. Sánchez, M. Sebastian, C. Souffreau, C. Dimier, M. Picheral, S. Searson, S. Kandels-Lewis, T. O. Coordinators, G. Gorsky, F. Not, H. Ogata, S. Speich, L. Stemmann, J. Weissenbach, P. Wincker, S. G. Acinas, S. Sunagawa, P. Bork, M. B. Sullivan, E. Karsenti, C. Bowler, C. d. Vargas, and J. Raes, "Determinants of community structure in the global plankton interactome," *Science*, vol. 348, p. 1262073, May 2015.

[4] K. Faust and J. Raes, "Microbial interactions : from networks to models," *Nature Reviews Microbiology*, vol. 10, pp. 538–550, Aug. 2012.

[5] K. Faust, J. F. Sathirapongsasuti, J. Izard, N. Segata, D. Gevers, J. Raes, and C. Huttenhower, "Microbial Co-occurrence Relationships in the Human Microbiome," *PLOS Computational Biology*, vol. 8, p. e1002606, July 2012.

[6] J. Friedman and E. J. Alm, "Inferring Correlation Networks from Genomic Survey Data," *PLOS Computational Biology*, vol. 8, p. e1002687, Sept. 2012.

[7] Y. Ban, L. An, and H. Jiang, "Investigating microbial co-occurrence patterns based on metagenomic compositional data," *Bioinformatics*, vol. 31, pp. 3322–3329, Oct. 2015.

[8] J. Aitchison, "A new approach to null correlations of proportions," *Journal of the International Association for Mathematical Geology*, vol. 13, pp. 175–189, Apr. 1981.

[9] L. C. Xia, D. Ai, J. Cram, J. A. Fuhrman, and F. Sun, "Efficient statistical significance approximation for local similarity analysis of high-throughput time series data," *Bioinformatics*, vol. 29, pp. 230–237, Jan. 2013.

[10] Y. Deng, Y.-H. Jiang, Y. Yang, Z. He, F. Luo, and J. Zhou, "Molecular ecological network analyses," *BMC Bioinformatics*, vol. 13, p. 113, 2012.

[11] Z. D. Kurtz, C. L. Müller, E. R. Miraldi, D. R. Littman, M. J. Blaser, and R. A. Bonneau, "Sparse and Compositionally Robust Inference of Microbial Ecological Networks," *PLOS Computational Biology*, vol. 11, p. e1004226, May 2015.

[12] S. Weiss, W. Van Treuren, C. Lozupone, K. Faust, J. Friedman, Y. Deng, L. C. Xia, Z. Z. Xu, L. Ursell, E. J. Alm, A. Birmingham, J. A. Cram, J. A. Fuhrman, J. Raes, F. Sun, J. Zhou, and R. Knight, "Correlation detection strategies in microbial data sets vary widely in sensitivity and precision," *The ISME Journal*, vol. 10, no. 7, pp. 1669–1681, 2016.

Annexes

Approches en métagénomique

Deux approches de séquençage sont utilisées pour quantifier l'information génique des espèces présentent dans différents échantillons.

L'approche de séquençage par gène marqueur fait appel à une spécificité du ribosome. Cet organite ubiquitaire comporte une unité 16S chez les procaryotes et 18S chez les eucaryotes. La séquence codante comporte des régions variables et conservées. À partir des régions conservées, des amorces sont crées permettant l'amplification des régions variables qui sont spécifiques de chaque espèce. L'assignation taxonomique se réalise par comparaison des amplicons existants dans diverses banques. Il est important de noter que la variabilité des copies d'ADN ribosomal 16S ou 18S d'une espèce à l'autre, ainsi que l'amplification importante d'une voire plusieurs régions variables entraîne des biais dans l'estimation de l'abondance de chaque espèce.

À l'inverse, l'approche *shotgun* s'intéresse à la totalité de l'ADN extrait. Ainsi l'assignation taxonomique est plus précise que la méthode par amplicons (assignation taxonomique obtenue jusqu'à l'espèce contre genre). De plus, les fonctions des espèces de l'écosystème peuvent être révélées grâce à l'information des gènes des séquences obtenues. Cependant un génome de référence est nécessaire pour l'assignation taxonomique et cette approche est plus coûteuse.

Informations supplémentaires sur les temps d'exécution

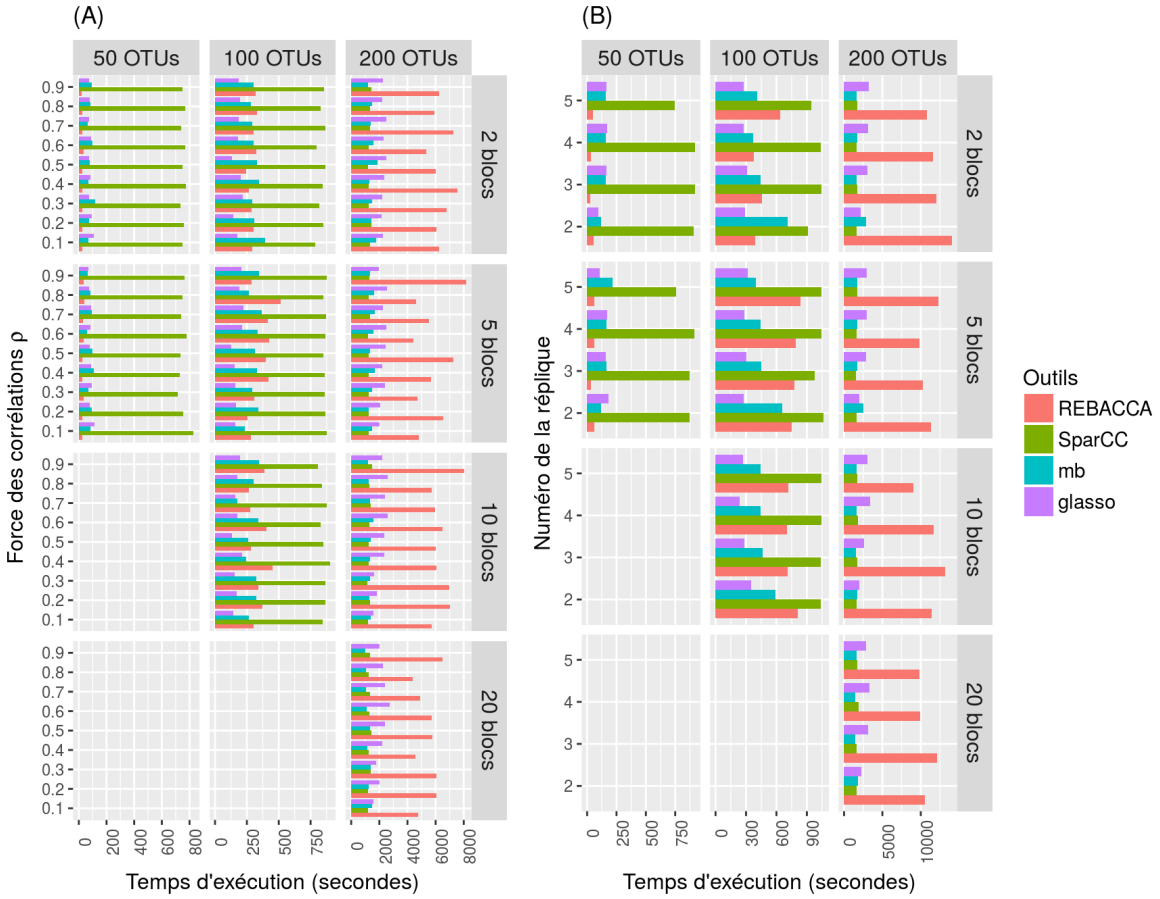


FIGURE 11 – Temps d'exécution des outils SparCC, REBACCA et des méthodes glasso et MB de l'outil SPIEC-EASI. A : Impact des forces des corrélations entre les taxa du réseau et la sparsité de celui-ci. B : Variabilité des temps d'exécution pour chaque réplique (4 threads) excepté la réplique 1 dont les simulations n'ont pas été réalisé en multithreading.

Résumé

La métagénomique consiste à caractériser expérimentalement la globalité d'un écosystème microbien, sans isoler au préalable les différents microorganismes qui le compose. Il est déjà établi que moins de 1 % des microorganismes de la biosphère sont cultivables. De plus, l'analyse par isolation des différents constituants d'un écosystème engendre la perte des relations écologiques entre ces microorganismes. Résultant en une interprétation partielle, voire erronée des dynamiques relationnelles de cet écosystème. Les approches de métagénomiques permettent de s'affranchir de ces limites en permettant d'appréhender un écosystème dans sa globalité. L'analyse des données métagénomiques soulève de nombreuses questions méthodologiques. Au-delà de la constitution d'un répertoire d'espèces ou de gènes et de l'étude fonctionnelle de la composition de ce répertoire, les analyses s'orientent de plus en plus vers des approches comparatives et des études d'interactions, ou plus précisément des études d'associations au sein d'un écosystème. Plusieurs méthodes statistiques sont apparues ces dernières années pour détecter des co-occurrences significatives entre espèces d'un même écosystème. Ces méthodes supposent que ces co-occurrences sont indicatives d'interactions écologiques (mutualisme, parasitisme, ...) entre différentes espèces. SparCC, REBACCA et SPIEC-EASI sont des outils récemment développés pour la reconstruction de réseaux microbiens. Nous présentons une comparaison des principales méthodes de reconstruction utilisées et les performances et limites de ces outils ont été évalués sur des données simulées.

Abstract

Metagenomics consists of experimentally characterizing a microbial ecosystem as a whole without prior isolation of the different microorganisms composing it. Many microorganisms are not culturable and separate analyses of each microorganism result in a warped understanding of the ecosystem, as they overlook close relationships between these microorganisms (mutualism, parasitism). Metagenomics enable us to apprehend a microbial ecosystem globally. Metagenomics data raise many methodological questions, as studies are increasingly moving beyond the mere constitution of a catalogue of species or genes and towards more complex analyses accounting for spatial data, time series and covariates. In particular, it is not clear how best to perform interaction studies and, more precisely, how to detect associations within the ecosystem. In recent years, several statistical methods were developed to detect significant cooccurrences between species, in different ecosystems and under different experimental conditions. These methods assume that cooccurences are indicative of biological interactions between species and interactions are thus revealed by reconstructing the cooccurrence network. SparCC, REBACCA and SPIEC-EASI are recently developed tools for the problem of network reconstruction in microbial ecology. We present a benchmark on the main reconstruction methods. Accuracy and running time was assessed on simulated metagenomic data.