Stage L3 - Rémi Legrand

Encadrement : Sylvain Moinard, Eric Coissac

Janvier 2022

1 Problématique

L'ADN environnemental constitue une source de données extrêmement vaste sur la biodiversité et apparaît comme complémentaire des observations écologiques classiques. Son étude repose sur les codes-barres moléculaires (le métabarcoding) qui permettent de retrouver les espèces desquelles provient l'information génétique retrouvée dans l'environnement. Cette technique repose sur plusieurs étapes de manipulation induisant chacune des biais expérimentaux, notamment lors de l'amplification de l'ADN collecté par Réaction de Polymérisation en Chaîne (PCR) qui induit un grand nombre d'artefacts moléculaires. L'application directe des estimateurs classiques de biodiversité sur les données de métabarcoding conduit à des résultats aberrants. Il est donc nécessaire de traiter ces données pour les rendre exploitables.

Comment détecter les erreurs PCR dans des données de métabarcoding pour améliorer l'estimation de la biodiversité du *Addo Elephant National Park* en Afrique du Sud ?

Tu exploiteras les liens entre les séquences lues en fin de PCR (similarité génétique, abondances...) pour en déduire des informations sur le contenu réel des échantillons étudiés et proposer une estimation plus fiable de la biodiversité végétale du parc.

2 Un peu de lecture pour commencer...

Afin d'appréhender le contexte de l'étude, je te renvoie à la lecture de l'introduction de mon rapport de stage de M2 (pages 5 à 9). Il est important de comprendre comment et pourquoi on mesure la biodiversité ainsi que l'usage de l'ADN environnemental pour répondre à des problématiques écologiques. L'étape cruciale du processus étant pour nous l'amplification par PCR, des rappels peuvent être utiles (voir le rapport également).

Les données à étudier sont issues du projet Protea du programme de collaboration avec l'Afrique du Sud Hubert Curien et proviennent d'un parc naturel, l'*Addo Elephant National Park*. Des excréments de 13 espèces d'herbivores ont été collectés pour déterminer leurs régimes alimentaires et décrire la flore naturelle de ce milieu. Mon rapport (section 2.1.2) expose brièvement l'origine des données ainsi que leur pré-traitement.

Si tu souhaites consulter des sources citées dans mon rapport, n'hésite pas à me demander (il n'est pas toujours évident de récupérer des articles de recherche directement).

3 Missions

Voici une trame de l'étude que je te propose de mener. Le sujet reste volontairement très libre et toutes tes idées sont les bienvenues pour analyser le jeu de données. Je te demande de rédiger une petite synthèse de ton travail, par exemple dans un fichier rmd (R Markdown) avec ton code, tes figures, tes interprétations...

3.1 Prise en main des données

Les données sur lesquelles tu vas travailler sont un sous-ensemble des données du projet Protea : 1011 variants génétiques ont été inclus dans le jeu de données. Certains correspondent à des plantes identifiées, d'autres non.

Les fichiers contenant les données à exploiter sont nommés *data-afrique-sud-Remi.txt*, *motus-afrique-sud-Remi.txt*, *samples-afrique-sud-Remi.txt*. Ils sont ouvrables avec le petit script R *script-stageL3-Remi.R*. Le premier décrit les résultats obtenus pour toutes les PCRs, séquence par séquence. Le deuxième est un résumé séquence par séquence de ces résultats. Le troisième contient des informations sur les différentes PCR qui ont été réalisées.

Commence par regarder ce que contiennent les données : nombre d'observations, espèces animales présentes... Tu peux représenter quelques valeurs : nombre d'échantillons par animal, nombre de lectures par échantillon... A ta guise.

Les 1 011 variants variants de la base de données ont été retenus sur un critère de similarité génétique. **Etudie la distance entre chacun des variants.** Représentation avec une *heatmap*. Comment interpréter ces proximités (ou non) entre les variants ?

Une première difficulté est que certaines espèces (réelles) ont des séquences très proches les unes des autres. A partir de la base de référence fournie (*Plant reference collection sequence information.xlsx*), **peut-on distinguer des espèces réelles de probables erreurs PCR ?** Annote la base de données "motus" avec l'espèce correspondante, quand elle est identifiée.

Utilisation de la librairie de référence en ligne NCBI BLAST (je te montrerai comment l'utiliser) : fais quelques tests : ces résultats sont-ils cohérents avec notre base de référence ?

3.2 Indicateurs de biodiversité

D'après l'ouvrage très complet d'Eric Marcon, on décide d'analyser une famille d'indices : les nombres de Hill, qui correspondent à un nombre équivalent d'espèces présentes dans l'écosystème étudié. Pour un nombre $q \geq 0$, ce nombre est défini par :

$$ {}^qD = \left(\sum_{s=1}^{S} p_s^q \right)^{\frac{1}{1-q}} $$

où les p_s sont les proportions relatives de chacune des S espèces ($\sum_s = 1^S p_s = 1$).

Code une fonction *nombre-hill(motus, q)* qui renvoie la biodiversité estimée selon la valeur du paramètre retenu q pour la communauté résumée dans *motus*.

Applique cette fonction :
— A l'ensemble de la base de données
— A chaque échantillon individuellement

Représentation graphique : "Hill = f(q)".

Cette question se décline à plein de possibilités : par exemple, diversité par animal, par saison... En fonction du temps disponible, on pourra regarder certaines configurations, mais plutôt après avoir avancé dans la partie suivante du sujet.

3.3 Probabilité de dérivation d'un autre variant

On veut se doter d'une méthode de détection des erreurs PCR : en particulier, on veut distinguer les vraies espèces (qui ne sont pas forcément toutes dans les bases de référence) des séquences qui ont été créées par mutation au cours de l'amplification. D'autre part, on veut attribuer correctement les variants supposés erronés à des espèces réelles.

Concrètement, j'aimerais que tu construises une matrice de taille 1011×1011 où la la valeur de la ligne i et de la colonne j est **la probabilité que le variant i appartienne à l'espèce j**. Par défaut, la matrice est la matrice identité :

$$\begin{pmatrix} 1 & 0 & & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ & & \dots & & \\ 0 & & \dots & 0 & 1 \end{pmatrix}$$

Ici chaque variant est considéré comme une espèce à part entière.

On peut considérer que les espèces présentes dans la base de référence (identifiées plus haut) ne sont pas issues de mutations. Leur ligne est donc inchangée dans la matrice : $p_{i,i} = 1, p_{i,j\neq i} = 0$. (question subsidiaire : est-ce vrai ?). **Modifie ta fonction *nombre-hill* en une fonction *nombre-hill(motus, matrice-proba, q)***.

Par ailleurs, si on sait que le variant i est dérivé d'un seul variant k, on a : $p_{i,k} = 1, p_{i,j\neq k} = 0$. Mais que faire quand le variant peut avoir des origines diverses ?

L'idée principale du stage est de construire une ou plusieurs matrices qui permettraient d'estimer plus précisément la biodiversité réelle. C'est un problème ouvert, plusieurs critères peuvent être envisagés pour y parvenir : en fonction de la proximité génétique, des abondances relatives... Pas plus de précisions dans ce sujet écrit pour te laisser trouver quelques idées mais on en discutera quand tu en seras à ce point du sujet pour choisir ensemble des pistes intéressantes à explorer.

Dans tous les cas, il est intéressant de regarder à chaque essai comment évolue le spectre de Hill selon la correction apportée.