Application des outils à des données réelles

Indice de Diversité de Hill

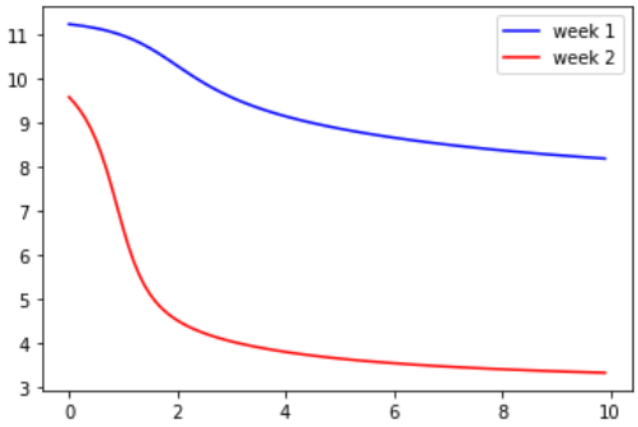
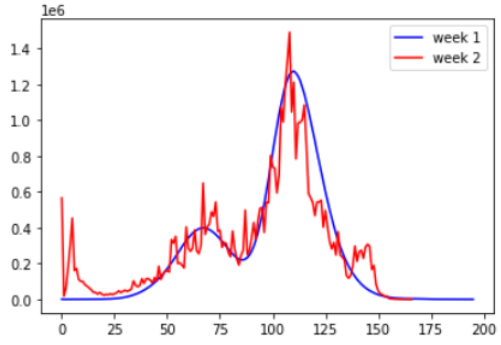


FIGURE 1 – Evolution de l'indice de diversité de Hill en fonction de la valeur de α

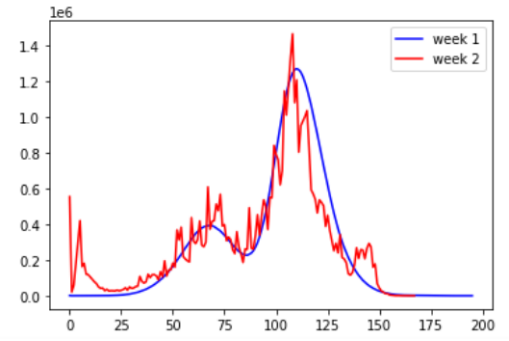
Nous observons que le répertoire de la deuxième semaine possède des valeurs de diversité de Hill inférieure à celles du répertoire de la première semaine. On peut donc supposer qu'il y a plus de diversité dans le répertoire de la semaine 2.

Pairwise Distance Distribution

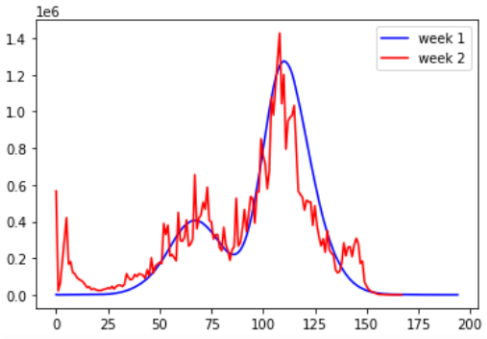
Le calcul du vecteur de Pairwise Distance Distribution requiert beaucoup de temps d'éxécution, nous avons donc fait le choix d'échantilloner les répertoires et calculer la Pairwise Distance Distribution sur ces échantillons. Nous avons choisi des échantillons de 10000 séquences pour obtenir un temps de calcul raisonnable et en se basant dur les paramètres du package R SumRep.
Nous avons effectué 3 tests en échantillonant aléatoirement 10000 séquences dans les répertoires. On obtient ainsi des distributions similaires selon l'échantillon considéré.
Sur la courbe du répertoire de la semaine 2, on remarque un pic à 0 et un pic à environ 10 ce qui montre que beaucoup de séquence sont similaires dans ce répertoire.



Echantillon 1



Echantillon 2



Echantillon 3

FIGURE 2 – Pairwise Distance Distribution pour 3 échantillons diffŕents

Mise en Réseau

La mise en réseau d'un répertoire d'une très grande taille nécéssite un temps de calcul conséquent. Nous avons donc fait le choix déchantilloner aléatoirement 5000 séquences pour chaque répertoire, afin de réduire le temps d'éxécution.
Chaque nœud représente une séquence et la taille de ce nœud est proportionnelle au nombre de séquences identiques à celle-ci. Un arc relie deux nœuds si la distance de Levenshtein entre les deux séquences est inférieure à un seuil.
Nous avons effectué deux tests pour deux valeurs de seuils différentes : 20 et 40.
Concernant les graphes avec une valeur seuil de 20, on remarque que pour le répertoire de la semaine 1 les nœuds sont de petite taille et sont dispersés. On observe tout de même des petits groupes de séquences. Dans le graphe du répertoire de la semaine 2 on observe des nœuds de grande taille.
Concernant les graphes avec une valeur seuil de 40, pour le répertoire de la semaine 1 on remarque que les nœuds sont toujours de petite taille cependant la taille des clusters est plus grande. Pour le graphe du répertoire de la semaine 2, on observe des nœuds de grande taille et également des clusters de plus grande taille.

Seuil à 20

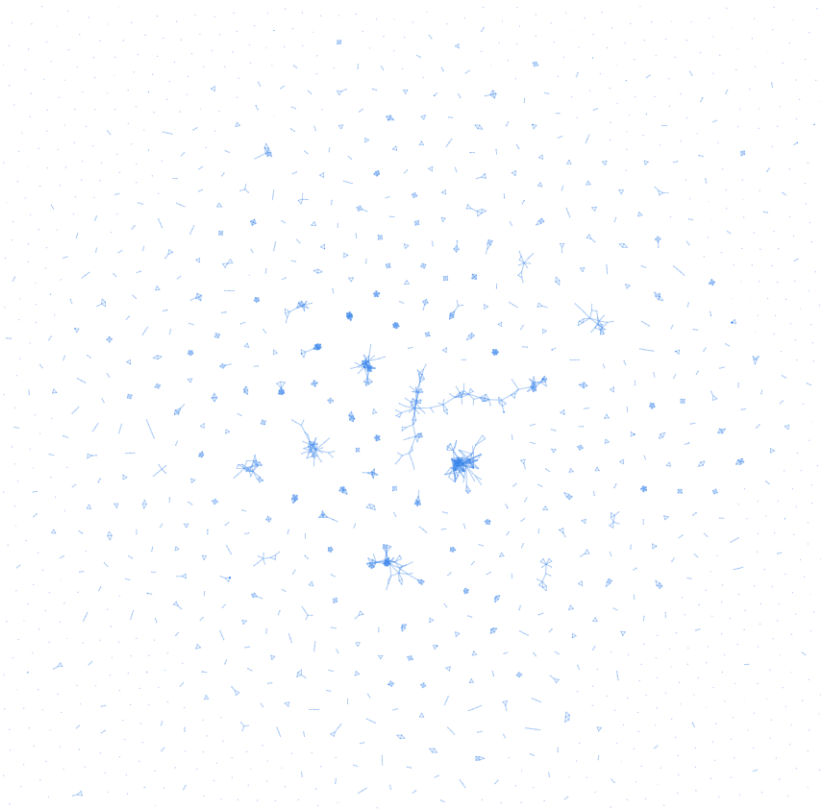


FIGURE 3 – Réseau des séquences du répertoire de la semaine 1 sur un échantillon de 5000 séquences et une valeur seuil fixée à 20

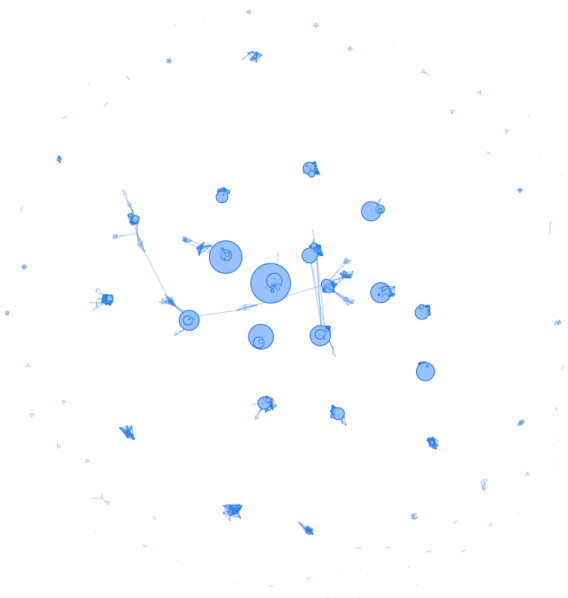


FIGURE 4 – Réseau des séquences du répertoire de la semaine 2 sur un échantillon de 5000 séquences et une valeur seuil fixée à 20

Seuil à 40

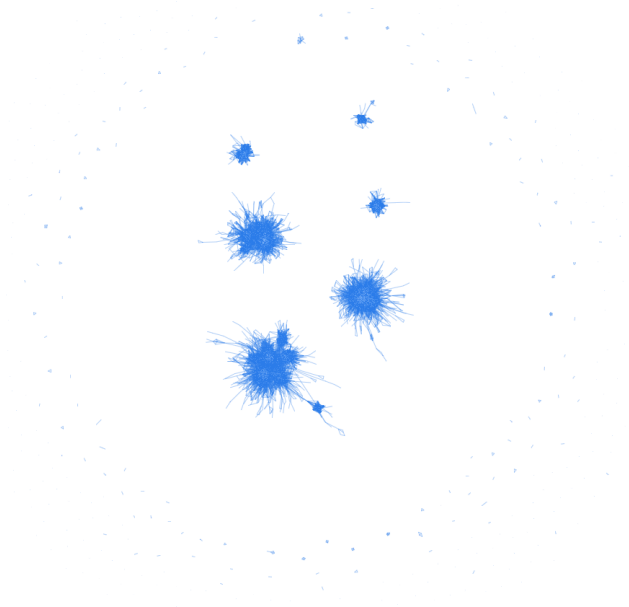


FIGURE 5 – Réseau des séquences du répertoire de la semaine 1 sur un échantillon de 5000 séquences et une valeur seuil fixée à 40

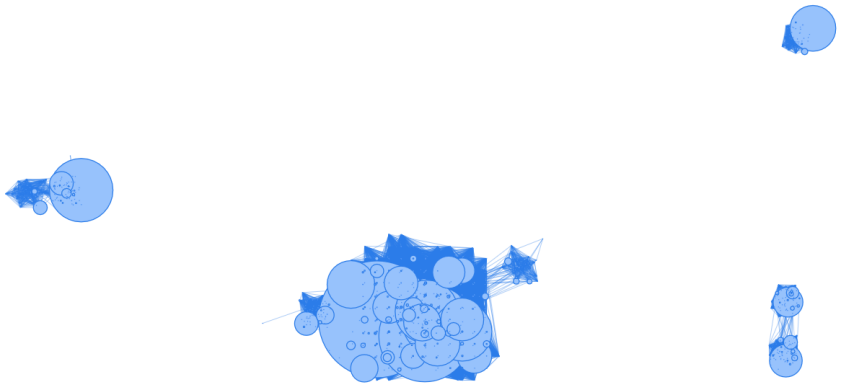


FIGURE 6 – Réseau des séquences du répertoire de la semaine 2 sur un échantillon de 5000 séquences et une valeur seuil fixée à 40

Coefficient de corrélation de Pearson

	week1	week2
week1	1.000000	0.866621
week2	0.866621	1.000000

FIGURE 7 – Coefficient de corrélation de Pearson entre les vecteurs de diversité de Hill des deux répertoires

Jensen-Shannon Divergence

Le calcul de la Jensen-Shannon Divergence s'effectue avec les vecteurs de Pairwise Distance Distribution. On approxime donc toujours les résultats en échantillonant 10000 séquences pour chaque répertoire.
Nous obtenons une valeur très proche de 0, ce qui montre que les distributions sont très similaires. En effet, sur la figure 2, on voit que les courbes suivent la même allure. La faible différence s'explique par l'irrégularité de la distribution du répertoire de la semaine 2.

```
JSD value :  0.03653448184683998
```

FIGURE 8 – Valeur JSD entre les deux répertoires

Pourcentage de Similarité

Similarité de 0%. Vérifier si Erreur.