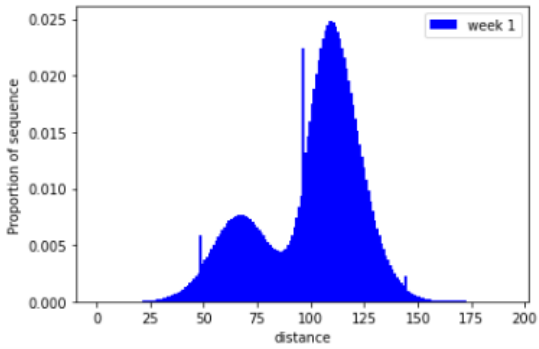
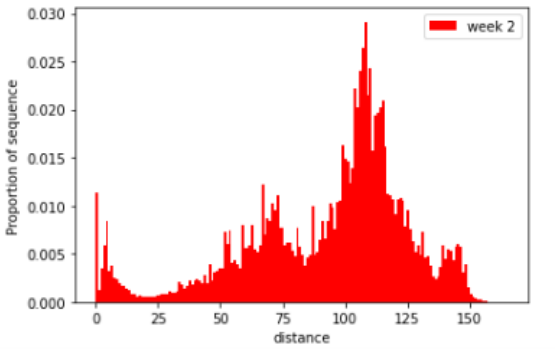
Comparaison des fonctions Pairwise Distance Distribution (PDD)

Pairwise Distance Distribution V1

La première fonction de la fonction calculant la PDD consiste à considérer toutes les séquences du répertoire et calculer la distance de Levenshtein pour chaque paires de séquences. Le problème est que si le répertoire contient beaucoup de séquences le temps de calcul sera conséquent. En effet, si on considère n séquences on doit alors calculer $\frac{n(n-1)}{2}$ valeurs de distances.
Nous avons donc fait le choix d'échantilloner le répertoire considéré à 10000 séquences pour atteindre un temps d'éxécution tournant autour des 10 minutes pour un répertoire.



Semaine 1



Semaine 2

FIGURE 1 – Pairwise Distance Distribution sur un échantillon de 10000 séquences.

```
Execution Time :  23.05892814795176 minutes
```

FIGURE 2 – Temps d'éxécution de la fonction pour les deux répertoires

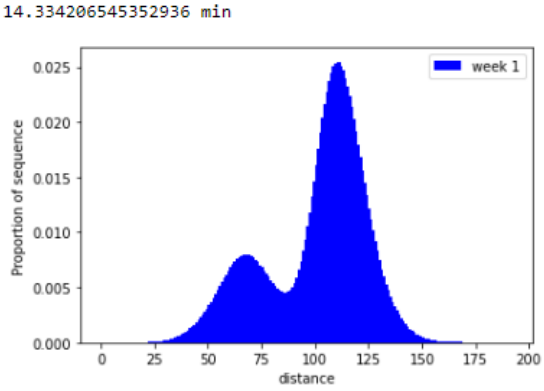
Pairwise Distance Distribution V2

Nous avons codé une deuxième version de la fonction permettant de calculer la PDD dans le but de réduire le temps d'éxécution. On remarque que la première version de la fonction donne des résultats précis, même en échantillonant à 10000 séquences. Or, on cherche seulement à identifier la tendance de la courbe de répartition. On peut donc penser à approximer encore plus les résultats pour diminuer le temps d'éxécution.

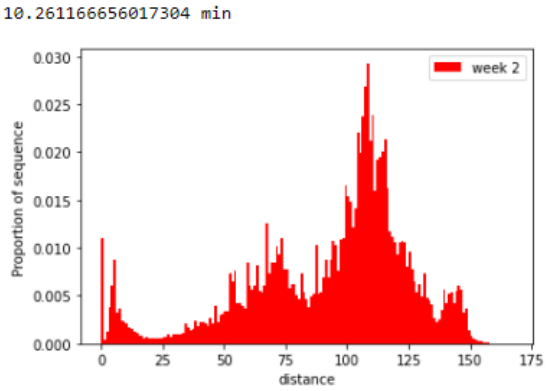
Nous avons fait le choix de partitionner le répertoire considéré en échantillons de 1000 séquences et calculer les PDD de chacun de ces échantillons. (Figure 3)

On peut approximer encore plus les résultats en effectuant un premier échantillonage pour considérer 10000 séquences et ensuite partitionner cet échantillon en sous-échantillons de 1000 séquences et calculer la PDD de chacun de ces sous-échantillons. (Figure 4)

On remarque que les courbes de distributions sont similaires pour les trois approximations.

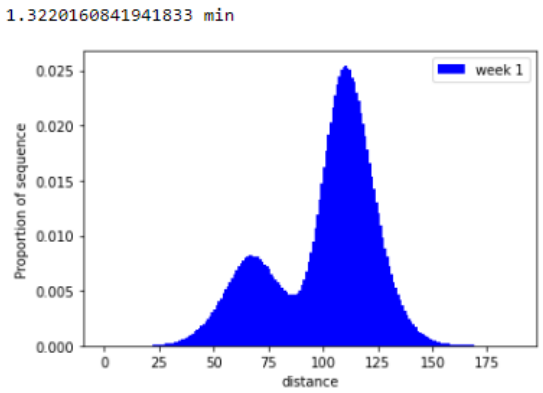


Semaine 1

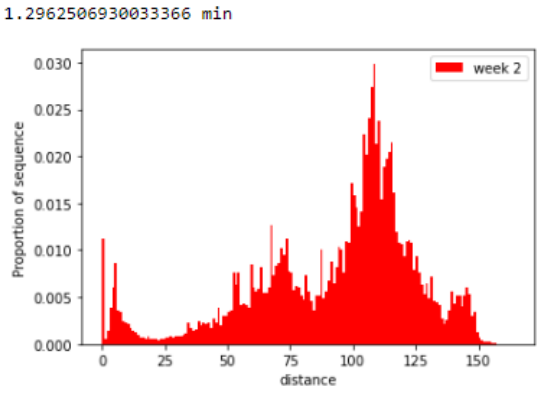


Semaine 2

FIGURE 3 – Temps d'éxécutions et Pairwise Distance Distribution en partitionnant les répertoires en échantillons de 1000 séquences.



Semaine 1



Semaine 2

FIGURE 4 – Temps d'éxécutions et Pairwise Distance Distribution sur un échantillon de 10000 séquences partitionné en sous-échantillons de 1000 séquences.