Université de Versailles Saint-Quentin-en-Yvelines
Master 2 : Calcul Haute Performance, Simulation

Nom : SMAIL
Prénom : Karim
Numéro étudiant : 21917901
Objet : TD3 Introduction de Modélisation en Biologie

TD3 :
Clusterisation sur matrice de distance

- Afin de pouvoir effectuer une comparaison entre les différentes matrices, on a besoin d'un code pour générer les différents arbres et heatmaps :

Dans le langage R, on a :

```
//Le chemin vers le fichier2
setwd ( "C:/ Users/Desktop/IHPS/IMB/TDs/cm3−2020" )

//La  lecture du fichier
distances <−read . csv2 ( "C:/ Users / u t i l i s a t e u r /Desktop/IHPS/IMB/
TDs/cm3−2020/distancesOK20k . csv " )

//Les différentes variables
var<−c ( " Bacillus_subtilis",
" Bacillus_amyloliquefaciens_FZB42 " ,
"Bacillus_pumilus_SAFR_032" ,
" Bacillus_thuringiensis_BMB171 " ,
"Bacillus_cereus_03BB102" ,
"Bacillus_anthracis_Ames " ,
"Bacillus_coagulans_2_6" ,
" Bacillus_atrophaeus_1942 " ,
"Bacillus_licheniformis_ATCC_14580" ,
"Escherichia_coli_K_12_substr__MG1655" ,
"Pseudomonas_aeruginosa_LESB58" ,
"Rhodobacter_sphaeroides_ATCC_17025" ,
"Streptomyces_flavogriseus_ATCC_33331" ,
"Micrococcus_luteus_NCTC_2665_uid59033" ,
" Lactococcus_lactis_Il1403 " )

//Le choix de la méthode ward.D2
cha<−hclust(dist(t(scale(distances[,var]))), method = "ward.D2")

//La génération  d'un arbre hiérachique
plot ( cha , xlab = "variables", ylab = "Distances", main ="classification Hiérachique")

//Le calcul de la corrélation
obj<−cor ( distances [ ,var] ,  use = "pairwise.complete.obs")
//La géné ration  des  heatmaps
heatmap (obj ,  col=gray ( seq (1,0 ,length.out = 26))
```

1. Comparer la clusterisation des 5 matrices de distances :

En exécutant le code juste avant avec le RStudio en utilisant les fichiers de matrices de distances .csv à différents nombre de bases, on aura les classifications hiérarchiques et les heatmaps suivants :

- Une matrice de distances à 1000 bases :

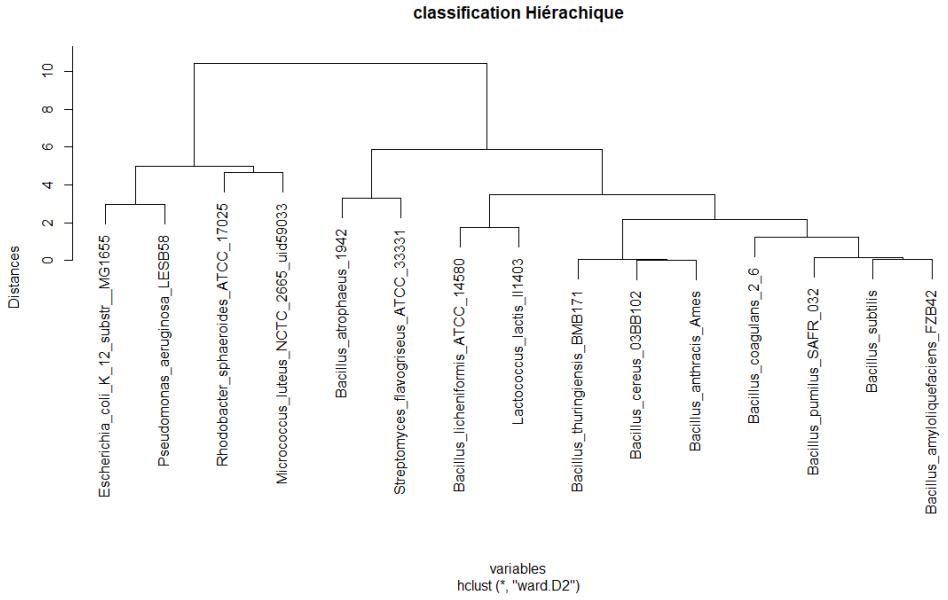

Figure1 – Classification Hiérarchique d'une matrice de distances à 1000 bases

On remarque que les élements de ces ensembles sont serrés, et les distances entre eux sont très petites :
"Bacillus_subtilis", "Bacillus_amyloliquefaciens_FZB42" et "Bacillus_pumilus_SAFR_032"."Bacillus_thuringiensis_BMB171", "Bacillus_cereus_03BB102" et "Bacillus_anthracis_Ames"

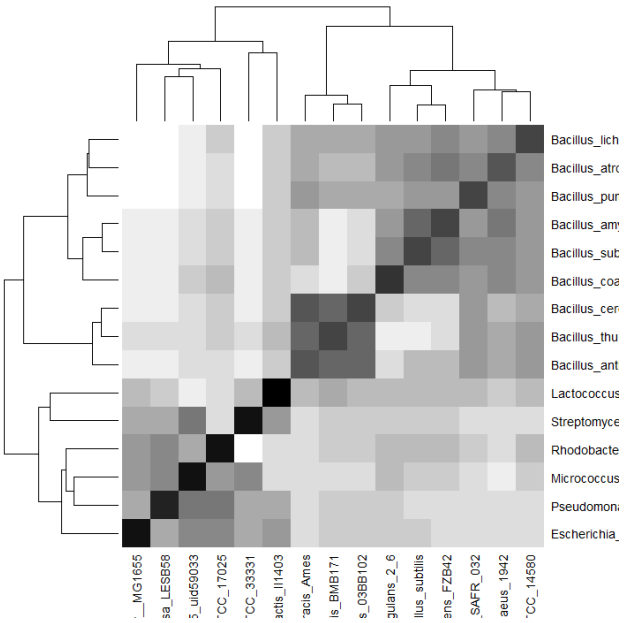

Figure2 – Le heatmap d'une matrice de distances à 1000 bases

On remarque trois carrés en gris foncé, qui sont formés avec les ensembles suivants : "Bacillus_subtilis", "Bacillus_amyloliquefaciens_FZB42" et "Bacillus_pumilus_SAFR_032"."Bacillus_thuringiensis_BMB171", "Bacillus_cereus_03BB102" et "Bacillus_anthracis_Ames". "Escherichia_coli_K_12_substr_MG1655" et "Pseudomonas_aeruginosa_LESB58".

Et un carré a moitié presque blanc avec : "Escherichia_coli_K_12_substr_MG1655", et "Streptomyces_flavogriseus_ATCC_33331".

- Une matrice de distances à 2000 bases :

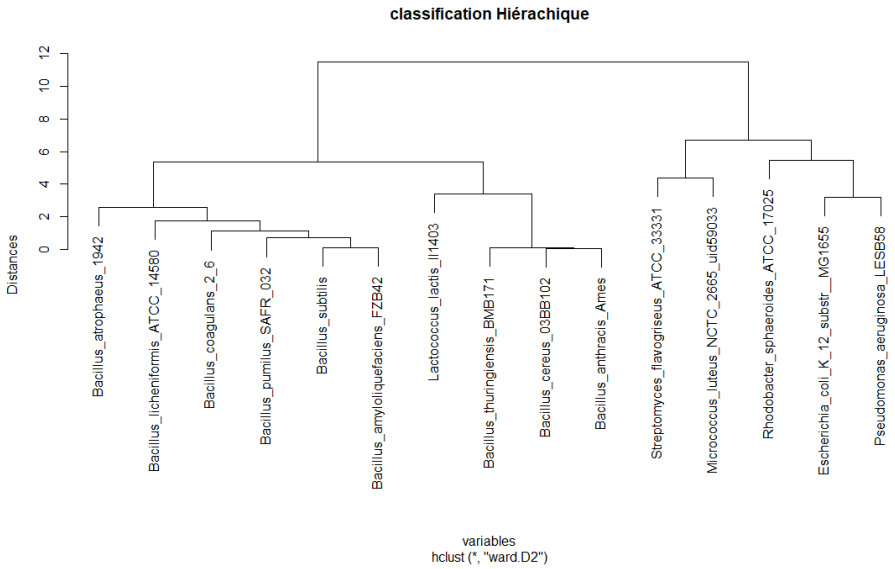

Figure3 – Classification Hiérachique d'une matrice de distances à 2000 bases

On remarque que les élements de ces ensembles sont serrés, et les distances entre eux sont très petites :
"Bacillus_thuringiensis_BMB171", "Bacillus_cereus_03BB102" et "Bacillus_anthracis_Ames"."Bacillus_subtilis" et "Bacillus_amyloliquefaciens_FZB42".

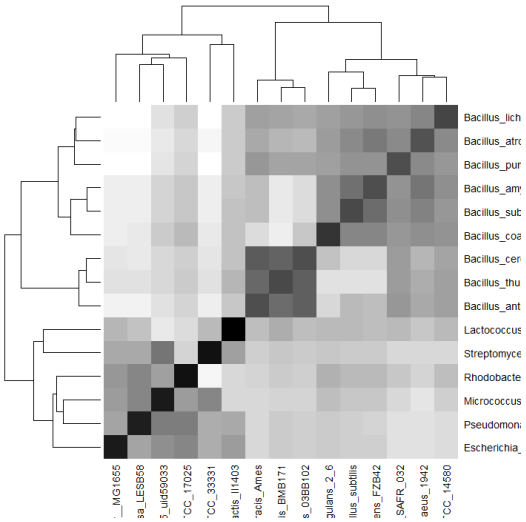

Figure4 – La heatmap d'une matrice de distances à 2000 bases

On remarque trois carrés en gris foncé, qui sont formés avec les ensembles suivants :

"Escherichia_coli_K_12_substr_MG1655", et "Pseudomonas_aeruginosa_LESB58".
"Streptomyces_flavogriseus_ATCC_33331" et "Micrococcus_luteus_NCTC_2665_uid59033".
"Bacillus_thuringiensis_BMB171", "Bacillus_cereus_03BB102" et "Bacillus_anthracis_Ames".

Et un carré a moitié presque blanc avec :
"Lactococcus_lactis_Il1403" et "Pseudomonas_aeruginosa_LESB58"

- Une matrice de distances à 4000 bases :

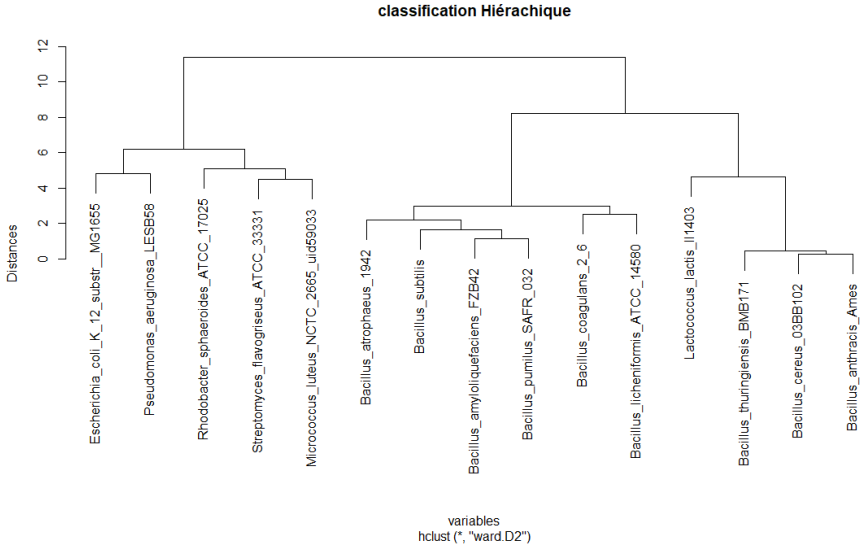

Figure5 – Classification Hiérarchique d'une matrice de distances à 4000 bases

On remarque que les élements de ces ensembles sont serrés, et les distances entre eux sont très petites :
"Bacillus_amyloliquefaciens_FZB42" et "Bacillus_pumilus_SAFR_032".
"Bacillus_thuringiensis_BMB171", "Bacillus_cereus_03BB102" et "Bacillus_anthracis_Ames".

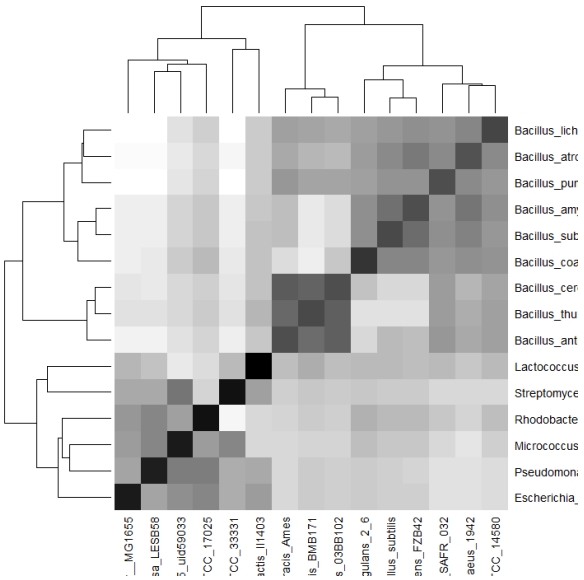

Figure6 – L'heatmap d'une matrice de distances à 4000 bases

On remarque des zones en gris foncé, qui sont formés avec les ensembles suivants :
"Bacillus_amyloliquefaciens_FZB42" et "Bacillus_pumilus_SAFR_032".
"Bacillus_thuringiensis_BMB171", "Bacillus_cereus_03BB102" et "Bacillus_anthracis_Ames".

Et un carré a moitié blanc avec :
"Lactococcus_lactis_Il1403" et "Pseudomonas_aeruginosa_LESB58".

- Une matrice de distances à 10k bases :

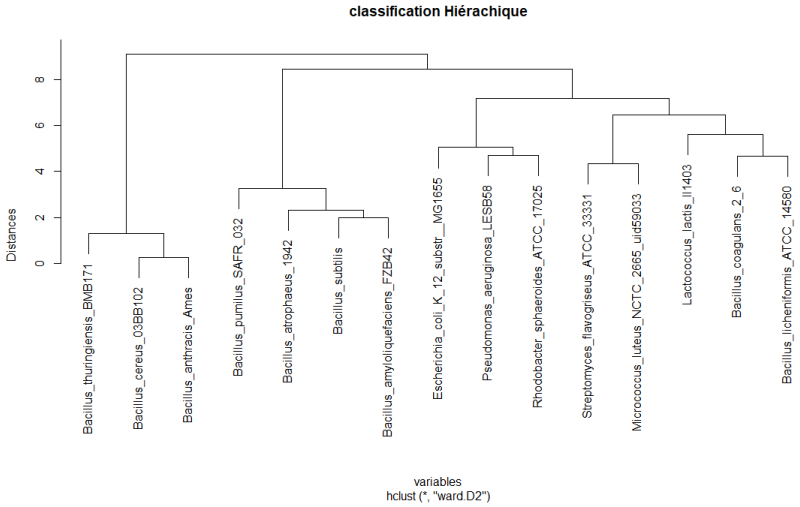

Figure7 – Classification Hiérarchique d'une matrice de distances à 10k bases

On voit que les distances entre les éléments de cet ensemble sont très petites :
"Bacillus_thuringiensis_BMB171", "Bacillus_cereus_03BB102" et "Bacillus_anthracis_Ames".

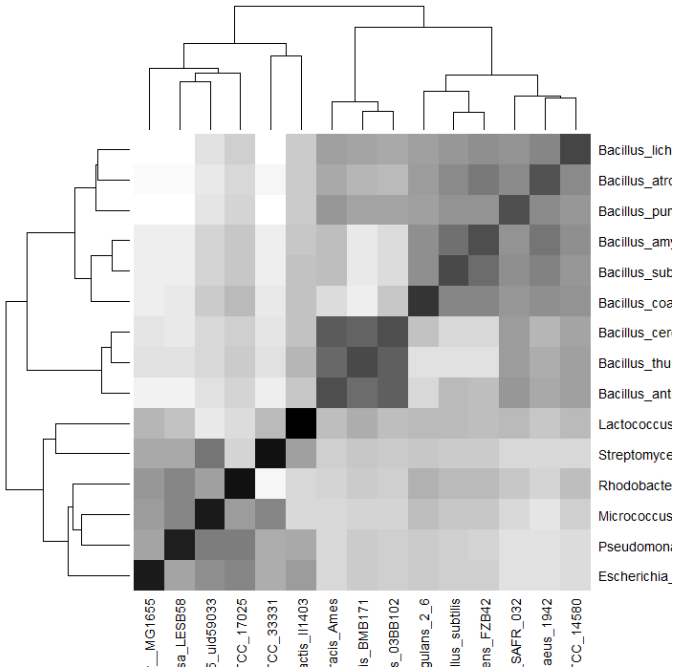

Figure8 – La heatmap d'une matrice de distances à 10k bases

On remarque un carré en gris foncé, qui est formé avec l'ensemble suivant :
"Bacillus_thuringiensis_BMB171", "Bacillus_cereus_03BB102" et "Bacillus_anthracis_Ames".

Et des zones en blanc avec :

"Micrococcus_luteus_NCTC_2665_uid59033" et "Pseudomonas_aeruginosa_LESB58"

- Une matrice de distances à 20k bases :

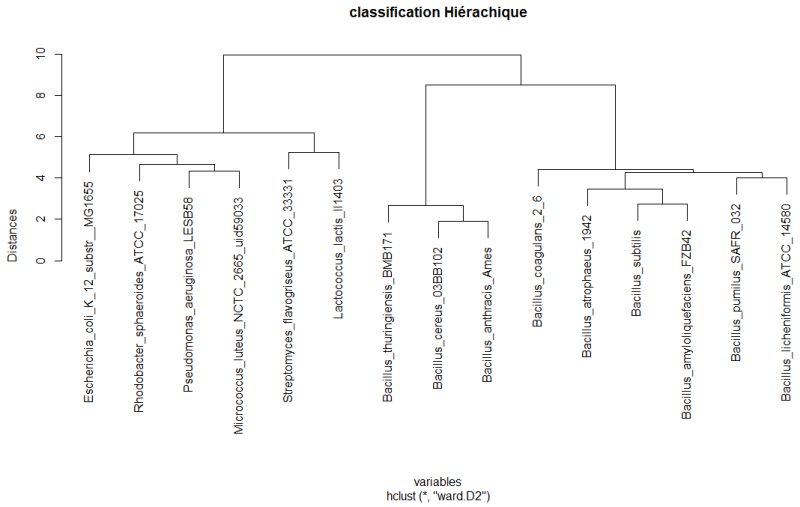

Figure9 – Classification Hiérachique d'une matrice de distances à 20k bases

On remarque que la distance entre les éléments de cet ensemble est très petite : "Bacillus_thuringiensis_BMB171", "Bacillus_cereus_03BB102" et "Bacillus_anthracis_Ames"

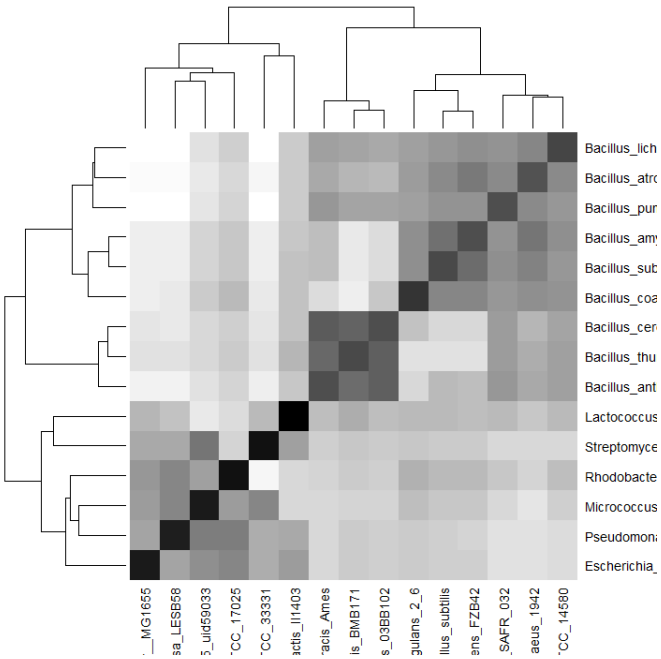

Figure10 – La heatmap d'une matrice de distances à 20k bases

On remarque un carré en gris foncé, qui est formés avec l'ensemble suivant : "Bacillus_thuringiensis_BMB171", "Bacillus_cereus_03BB102" et "Bacillus_anthracis_Ames".
Et des zones presque blanches avec :
"Rhodobacter_sphaeroides_ATCC_17025" et "Streptomyces_flavogriseus_ATCC_33331"

2. **Proposer des hypothèses et conclusions sur les différences observées :**

- ## Une matrice de distances à 1000 bases :

D’après les observations qu’on a eu a partir de l’arbre et la heatmap, on constate que les éléments de chacun des ensembles suivants sont fortement corrélés :

1. "Bacillus_subtilis", "Bacillus_amyloliquefaciens_FZB42" et "Bacillus_pumilus_SAFR_032".
2. "Bacillus_thuringiensis_BMB171", "Bacillus_cereus_03BB102" et "Bacillus_anthracis_Ames".

Alors qu’il ne y a aucune corrélation entre les deux éléments suivants : "Escherichia_coli_K_12_substr_MG1655", et "Streptomyces_flavogriseus_ATCC_33331".

- ## Une matrice de distances à 2000 bases :

D’après nos observations, on constate que les éléments de l’ensemble suivant sont fortementcorrélés :
"Bacillus_thuringiensis_BMB171", "Bacillus_cereus_03BB102" et "Bacillus_anthracis_Ames".

Alors qu’il ne y a aucune corrélation entre les deux éléments suivants :"Lactococcus_lactis_Il1403" et "Pseudomonas_aeruginosa_LESB58".

- ## Une matrice de distances à 4000 bases :

D’après nos observations, on constate que les éléments des deux ensembles suivants sont fortement corrélés :
"Bacillus_amyloliquefaciens_FZB42" et "Bacillus_pumilus_SAFR_032"."Bacillus_thuringiensis_BMB171", "Bacillus_cereus_03BB102" et "Bacillus_anthracis_Ames".

Alors qu’il ne y a aucune corrélation entre les deux éléments suivants :"Lactococcus_lactis_Il1403" et "Pseudomonas_aeruginosa_LESB58".

- ## Une matrice de distances à 10k bases :

D’après les observations obtenues on déduit que les éléments de cet ensemble sont fortement corrélés :
"Bacillus_thuringiensis_BMB171", "Bacillus_cereus_03BB102" et "Bacillus_anthracis_Ames".
Mais ces deux éléments le contraire:
"Micrococcus_luteus_NCTC_2665_uid59033" et "Pseudomonas_aeruginosa_LESB58"

- ## Une matrice de distances à 20k bases :

Grace a l’arbre de classement et la heatmap on déduit que ces éléments sont fortement corrélés :
"Bacillus_thuringiensis_BMB171", "Bacillus_cereus_03BB102" et "Bacillus_anthracis_Ames".

Contrairement a ces deux éléments :
"Rhodobacter_sphaeroides_ATCC_17025" et "Streptomyces_flavogriseus_ATCC_33331".