TD3. CLUSTERISATION SUR MATRICE DE DISTANCE

MATIERE : INTRODUCTION A LA MODELISATION EN BIOLOGIEPART 02 : ETUDE

Responsable de formation : Mme. Anne-Françoise Batto **&** Mr. Jean-Michel Batto

Etudiante : AICHA LAMMAMRA

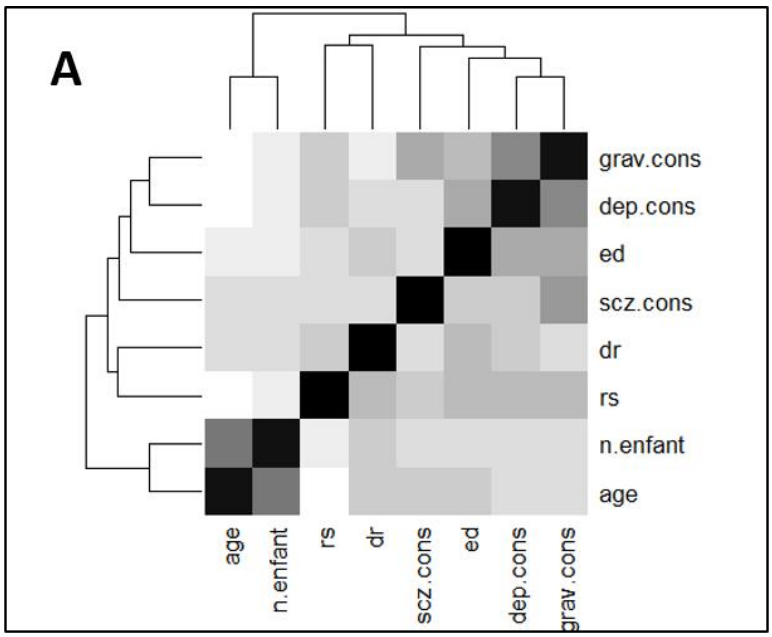
2020/2021

Introduction

Dans le TD n°1 nous avons construit un arbre Phylogénie à partir 15 fichiers de séquence génomique de bactéries. En calculant la distance entre ces fichiers, nous avons obtenu une matrice symétrique positive. Dans ce travail, nous allons comparer la clusterisation sur cinq matrices de distance .Ces matrices ont été obtenus à partir un découpage de 15 génome de différents tailles (1000, 2000,4000, 10k et 20k).

Heatmap

Heatmap est une technique de visualisation de données. Dans R, cette fonction prends comme arguments : «***obj :*** matrice de corrélation à créer» et « ***col :*** palettes des couleurs » (figure 1. A). Le résultat de ***Heatmap*** obtenu est sous forme d'une matrice de corrélation (figure 2 B).



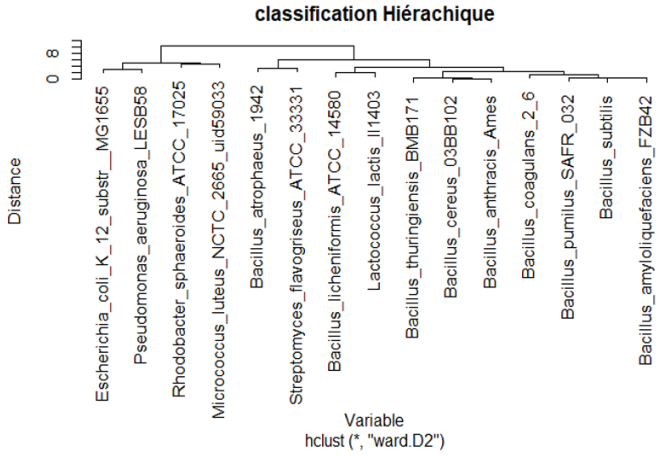

```r
setwd("D:/Documents_ AICHA_LAMMAMRA/master/M2_CHP/Cours/now/GO+R/2021/Cours=TD/CM3-2020")
smp <- read.csv2("D:/Documents_
AICHA_LAMMAMRA/master/M2_CHP/Cours/now/GO+R/2021/Cours=TD/CM3-2020/distancesOK20k.csv")

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")

cha<-hclust(dist(t(scale(smp[,var]))),method = "ward.D2")
plot(cha,xlab = "Variable", ylab="Distance",main = "classification Hiérachique_20K")
obj<- cor(smp[,var],use = "pairwise.complete.obs")
heatmap(obj,col=gray(seq(1,0,length.out=16)))
```

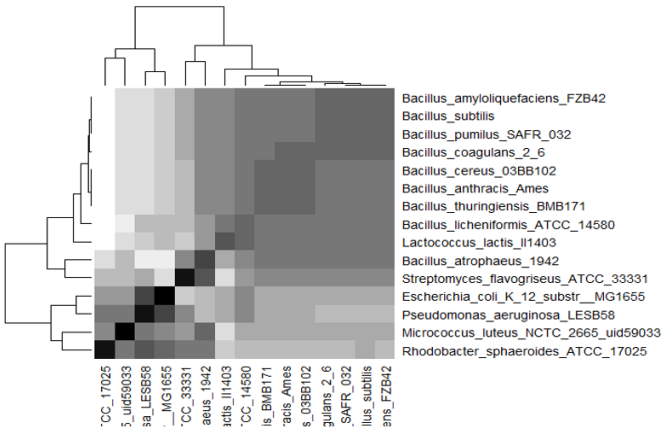
B

Figure 1. Heatmap de clustering hiérarchique de 15 bactéries

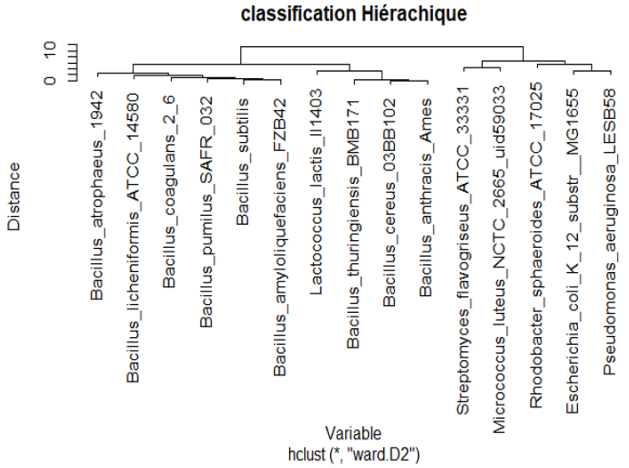
Clusterisation

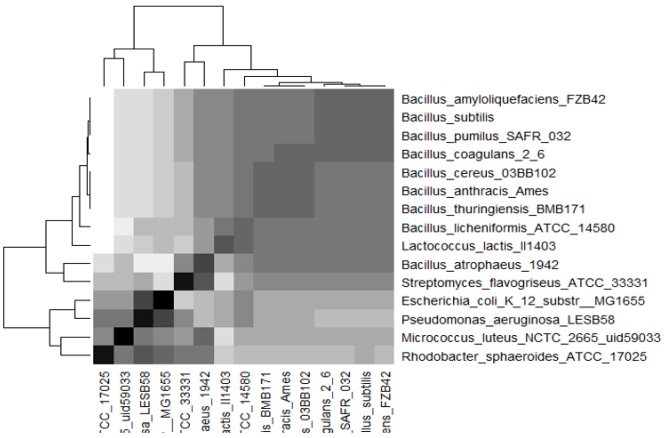
- **En 1000 nucléotides**



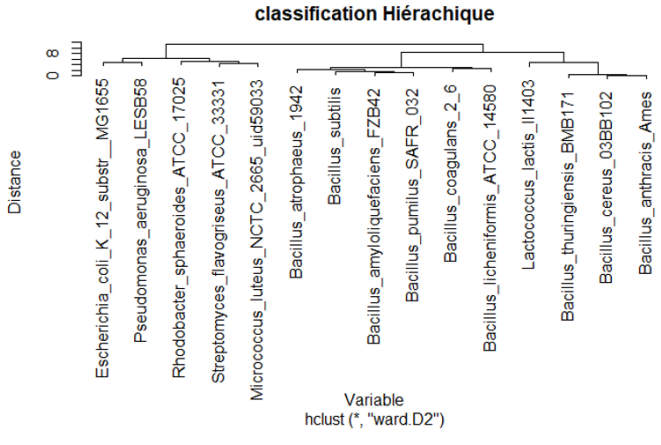


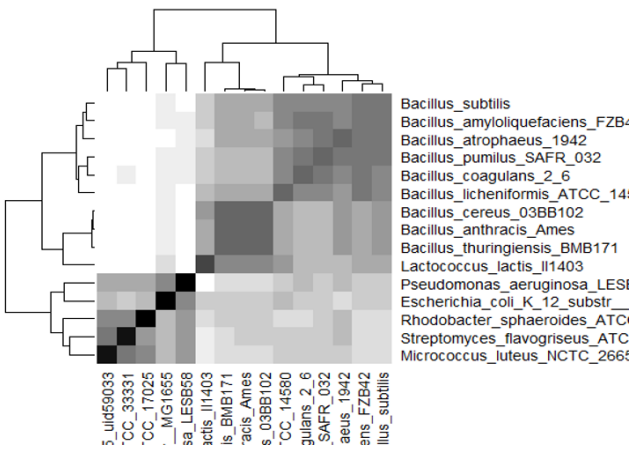
- **En 2000 nucléotides**



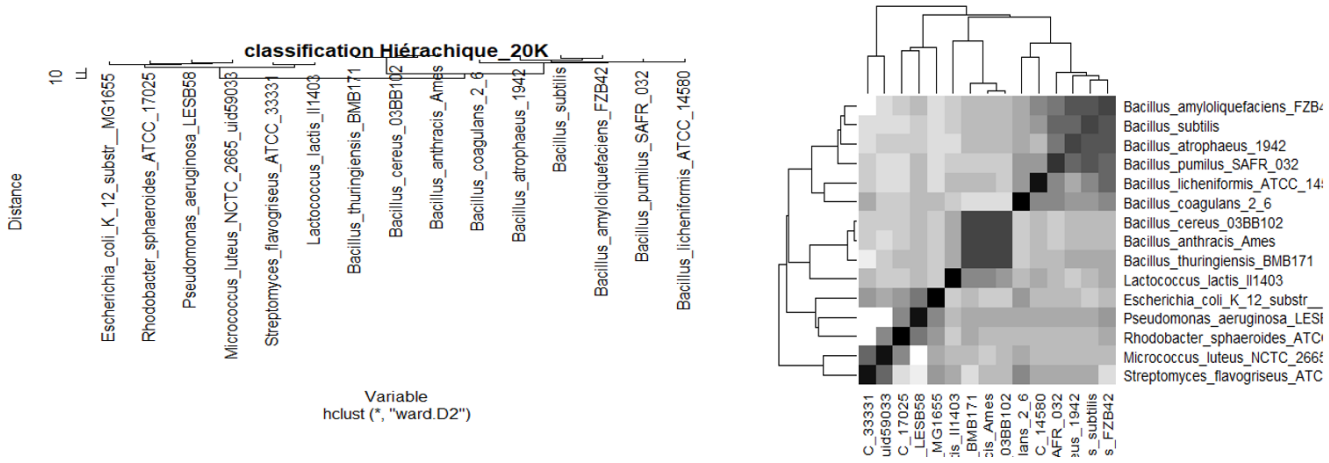


- **En 4000 nucléotides**

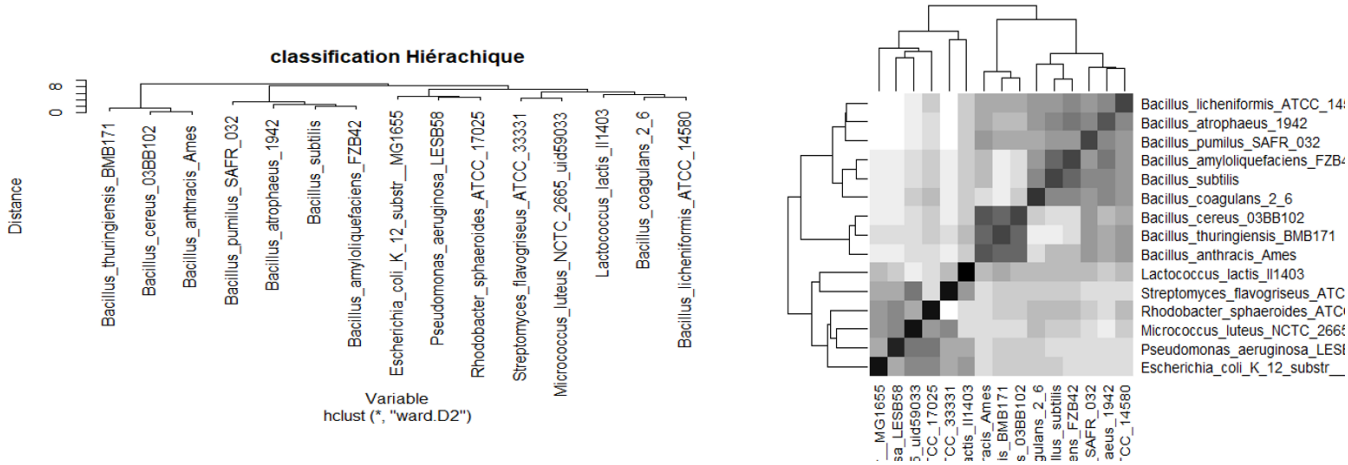




- **En 10000 nucléotides**



- **En 20000 nucléotides**



Interprétations

1. Pour le découpage des séquences de génomique en base de 1000 et de 2000 nucléotides : nous remarquons que tous les variables du groupe *Bacillus* ne sont pas corrélées ni entre eux ni avec les autres variables. Nous avons pu observer aussi que les bactéries *Escherichia_coli_K_12_substr__MG1655, Pseudomonas_aeruginosa_LESB58*, *Rhodobacter_sphaeroides_ATCC_17025 et Micrococcus_luteus_NCTC_2665_uid59033* sont corrélées positivement entres eux.
2. Pour le découpage des séquences de génomique en base de 4000 nucléotides, les trois clusters sont très destins.
   - ***Cluster 1*** : *Escherichia_coli_K_12_substr__MG1655, Pseudomonas_aeruginosa_LESB58*, *Rhodobacter_sphaeroides_ATCC_17025 et Micrococcus_luteus_NCTC_2665_uid59033* .
   - ***Cluster 2*** : *Lactococcus_lactis_Il1403, Bacillus_thuringiensis_BMB171, Bacillus_cereus_03BB102, Bacillus_anthracis_Ames, Bacillus_licheniformis_ATCC_14580*.
   - ***Cluster3*** *:Bacillus_subtilis,Bacillus_amyloliquefaciens_FZB42,Bacillus_atrophaeus_1942,Bacillus_pumilus_SAFR_032,Bacillus_coagulans_2_6,Bacillus_licheniformis_ATCC_14580).*

Les bactéries de chaque cluster sont très proches en distance. Donc, elles sont toutes corrélées positivement entre eux. Par conséquent, ces mesures semblent être indicatrices de la similarité de séquence génomique entre ses bactéries de chaque cluster.

3. Les résultats pour la base 10000 et 20000 sont presque similaire, où nous observons que la couleur de la matrice de corrélation est plus froide donc les bactéries sont différentes en distance entre eux. Nous n'avons pas réussi à trouver des séquences génomiques communes entre les bactéries.

Conclusion

Après présentation des cinq matrices de découpage de différentes tailles, nous avons eu des résultats différents ainsi la méthode de base 4000 nucléotides est la plus robuste pour faire la clusterisation.