Métataxonomie des bactéries du fromage

Antoine Branca

18/11/2021

1. Introduction

1.0 Instructions Compte-Rendu

Un compte-rendu du TP par binôme (ou trinôme si nombre impair) vous est demandé pour le **01/12 20h00 dernier délai**. Il est à déposer dans l'espace dédié sous e-campus. Bien mettre le nom des 2 binômes sur le nom du fichier. Il vous est demandé de rédiger 5 pages maximum avec 3 figures maximum sous format d'article (en pdf format LateX ou word avec marge et police par défaut). Vous devez répondre à la question : **Quel(s) facteur(s) influe(nt) le plus sur la composition et la diversité des communautés bactériennes des croûtes de fromages ?**

1.1 Contexte

Une étude a cherché à étudier la composition microbienne des croûtes de fromage en relation avec le type de fabrication. Les auteurs ont étudié la composition microbienne de 24 croûtes de fromages différents dont les données sont stockées dans le fichier **MetadataCheese.csv** déposé sur ecampus :

ID	RindType	Moisture	pH	NaCl	Pasteurized	Country	Region	Milk
4524482	natural	34.91	6.55	0.03	N	USA	Connecticut	Cow
4524483	washed	58.54	6.47	0.19	N	France	Burgundy-Champagne	Cow
4524484	bloomy	31.45	6.56	0.27	N	France	Savoie	Goat
4524485	bloomy	62.55	6.31	0.17	N	France	Normandy	Cow
4524486	washed	55.75	6.71	0.14	Y	Ireland	Co._Cork	Cow
4524487	washed	26.91	7.36	0.12	N	France	Doubs_county	Cow
4524488	natural	37.16	7.50	0.22	N	England	Nottinghamshire	Cow
4524489	bloomy	58.62	6.54	0.10	N	England	Somerset	Goat
4524490	natural	33.00	6.69	0.04	N	USA	Vermont	Cow
4524491	natural	16.80	6.13	0.15	Y	USA	Vermont	Cow
4524493	natural	37.24	5.44	0.09	N	USA	Massachusetts	Cow
4524494	washed	61.71	7.42	0.09	Y	USA	California	Cow
4524495	washed	47.75	7.02	0.08	N	USA	Virginia	Cow
4524496	washed	50.60	7.75	0.12	N	France	Auvergne	Cow
4524497	natural	42.47	6.80	0.15	Y	France	Auvergne	Cow
4524498	natural	10.91	7.24	0.09	Y	Spain	Catalunya	Goat
4524499	washed	40.71	8.26	0.13	N	USA	Vermont	Cow
4524500	washed	34.36	7.86	0.12	Y	Switzerland	Tufertschwil	Cow
4524501	natural	43.59	7.31	0.12	N	USA	Vermont	Goat
4524502	washed	40.00	7.03	0.19	N	France	Rhone-Alpes	Cow
4524504	bloomy	67.37	6.74	0.06	Y	Italy	Piedmont	Sheep
4524505	washed	60.26	7.18	0.07	N	USA	Wisconsin	Cow

La croûte de fromage est un biofilm de microorganismes, c'est-à -dire une communauté de bactéries et de

champignons qui forme une matrice. Les composants principaux du lait sont la caséine, les acides gras animaux et le lactose. Pour faire du fromage, on ajoute de la chymosine ou présure qui va cliver la caséine par hydrolyse pour former alors des agrégats de protéines appelés micelles. Ces dernières s'aggrégeront entre elles pour former une structure plus ou moins ferme appelée le caillé. C'est ce caillé qui va alors être salé et affiné pour fabriquer du fromage. Les composés du fromage vont être métabolisés par les différents microorganismes. La dégradation du caillé et des peptides résultants va notamment engendrer la libération d'ammoniac et ramollir le fromage au cours du temps. Le lactose va rapidement être dégradé en acide lactique ce qui a pour conséquence d'acidifier le fromage. Le pH remonte ensuite sous l'action de la protéolyse. Enfin les acides gras vont être oxydés par les divers microorganismes pour récupérer de l'énergie (β-oxydation des acides gras). Cette voie métabolique est la plus importante pour générer des composés aromatiques et volatils typiques de chaque fromage. Le choix du lait (vache, brebis, chèvre ou pasteurisé, lait cru) va influer principalement sur les microorganismes présents et la quantité realtive de protéines/lipides/sucres. Chaque type de fabrication de fromage va sélectionner différentes communautés de microorganismes.

1.2 Matériel à disposition

MG-RAST (https://www.mg-rast.org/) est un pipeline d'assemblage et d'annotation de données métagénomiques. L'assemblage et l'annotation sont des étapes coûteuses en temps et qui nécessitent des infrastructures informatiques conséquentes. Nous allons donc utiliser des données de métagénomiques de l'étude produites sous MGRAST pour pouvoir lier les données taxonomiques et fonctionnelles aux différentes particularités de chaque fromage. Les données correspondent aux nombres de séquences qui se sont alignés à une annotation. Pour la métaxonomie, il s'agit de séquences de gènes marqueurs (16S bactérien ou ITS champignons) et pour la métagénomique ce sont tous les gènes ou leur annotation. Ici nous allons nous intéresser uniquement aux bactéries. Tout d'abord, il va falloir installer un certain nombre de packages pour pouvoir utiliser le package matR qui permet de directement interroger la base de données MG-RAST depuis R. Il vous faudra tout d'abord vérifier que ces paquets linux sont présent pour utiliser devtools dans R:

```
sudo apt-get install libxml2-dev
sudo apt-get install libcurl4-openssl-dev
```

Ensuite dans R il faut installer :

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager", quietly = TRUE)
BiocManager::install("qvalue")
list.of.packages <-c('devtools','RJSONIO','ecodist','gplots','scatterplot3d','usethis', 'httr', 'rcmdch
new.packages <- list.of.packages[!(list.of.packages %in% installed.packages()[,"Package"])]
if(length(new.packages)) install.packages(new.packages)
library(devtools)
install_github(repo='MG-RAST/matR',quiet=T)
```

Une fois matR installé, il est alors possible d'aller récupérer les données sous MG-RAST.

2. Analyse des données microbiennes

2.1 Estimation de la diversité α

Dans le cadre de ce TP, nous allons nous intéresser aux microrganismes. Nous allons ici récupérer des OTUs pour Operational Taxonomic Units. Ils s'agit simplement de taxons déduits à partir de similarité de séquences. Voici la liste des commandes qui permettent de récupérer les données :

```
library(matR)
```

```
## Loading required package: MGRASTer
```

```
## MGRASTer (0.9 02e288)
```

```
## Loading required package: BIOM.utils
```

```
## BIOM.utils (0.9 dbcb27)

## matR: metagenomics analysis tools for R (0.9.1)
```

```
#List des accessions associées à l'étude mgp3362 (2 échantillons absents des métadonnées)
list_mgp3362<-metadata("mgp3362")$mgp3362
list_mgp3362<-list_mgp3362[c(-11,-22)]
#Récupération des données taxonomiques (request='organism') depuis les hits
#de la base de données RDP (source='RDP') au niveau de l'ordre (group_level='order')
#avec une evalue de 1e-15 (evalue=15)
biom_phylum<-biomRequest(list_mgp3362,request="organism",source="RDP",
                         group_level="order",evalue=15,wait=TRUE)
```

```
##   start stop requested                                ticket
## 1     1   22      TRUE d7f38782-2d50-4f6d-b1a2-09c27f14445e
##                               file
## 1 /tmp/Rtmp7fnLNn/file694b82e34f2
```

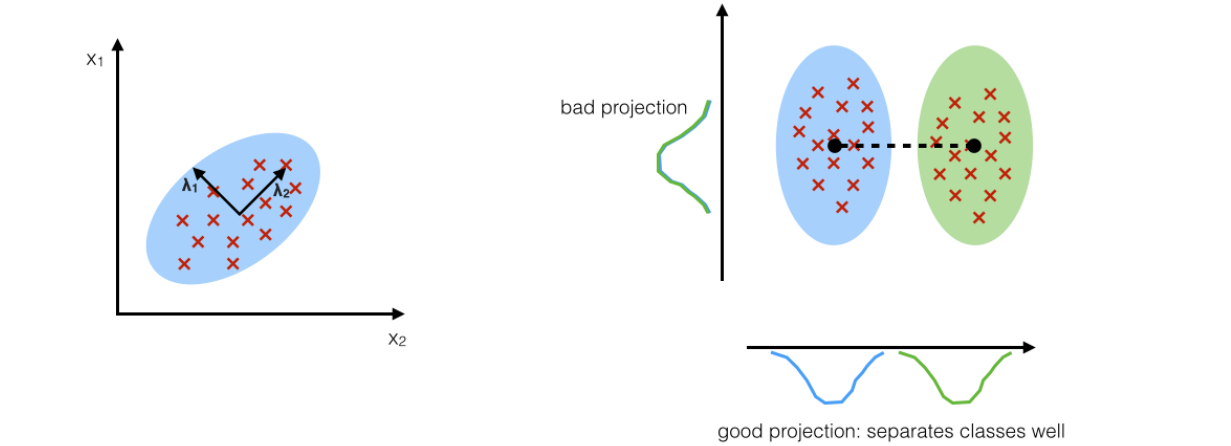
```
#Transformation en matrice
phylum_matrix<-as.matrix(biom_phylum)
```

Dans un premier temps, nous allons voir si les communautés microbiennes ont été bien échantillonnées. Pour ce faire, on effectue des courbes de raréfactions. Il s'agit de rééchantillonner les données de comptage des OTUs aléatoirement avec une taille d'échantillonnage de plus en plus grande. On trace ensuite la courbe correspondant au nombre d'OTUs en ordonnées et à la quantité de séquences en abscisse. **Faites ces graphiques pour les échantillons et les interprétez.**
Les courbes de raréfaction permettent d'avoir un aperçu de la diversité α de chaque communauté notamment en terme du nombre d'OTUs. Une autre façon de faire est l'indice de diversité de Shannon $H' = \sum_{i=1}^{R} p_i * ln(p_i)$ où R est le nombre d'OTUs, et p_i la fréquence de l'OTU i. **Calculez cet indice pour chaque communauté et testez si un des facteurs mesurés permet d'expliquer la diversité observée.**

2.2 Analyse des facteurs structurants la communauté

Les différences de physico-chimie entre les fromages peuvent jouer sur la diversité microbienne mais c'est surtout sur la composition même des communautés qu'elle devrait jouer. À partir des données nous allons donc voir s'il y a une association entre les paramètres des fromages produits et la composition microbienne de la croûte.
Construisez trois ACP des abondances microbiennes à partir de l'objet *phylum_matrix* **en faisant apparaître le type de croûte, le lait et la pasteurisation sur vos figures.** Pouvez-vous discriminer des types fromages à partir des communautés de microorganismes ?

L'ACP bien qu'utile pour explorer les données ne permet pas directement de répondre à cette dernière question. En effet, l'ACP sert à représenter le maximum de variation du jeu de données dans un espace dimensionnel réduit. Ici on veut savoir si certains microorganismes discriminent entre tel et tel fromage. On peut alors utiliser l'analyse discriminante qui cherche à maximiser la discrimination entre groupes.



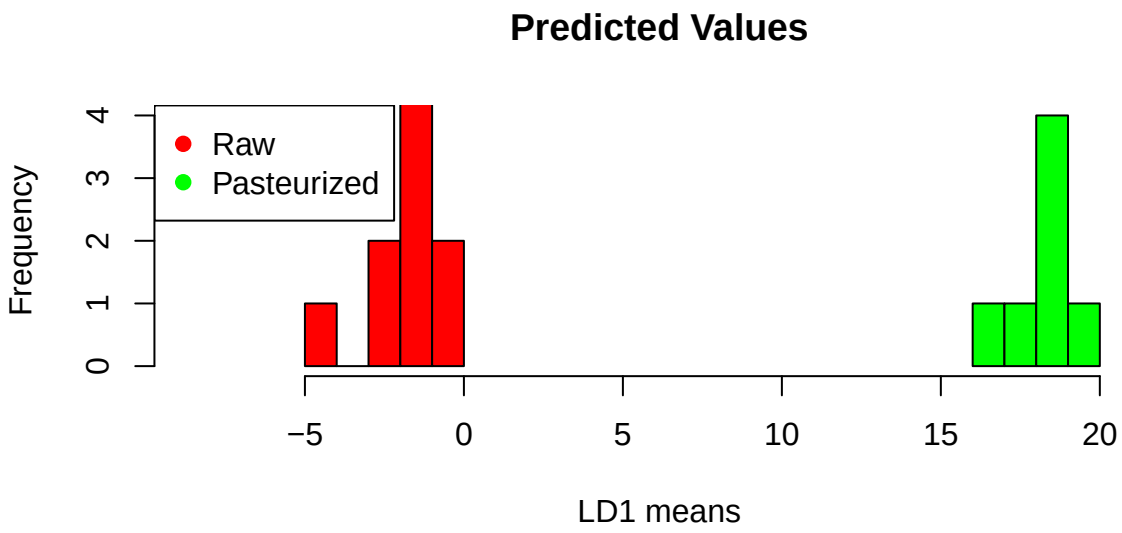
On va alors ajuster une droite de régression qui passe à distance égale des valeurs des groupes que l'on souhaite discrimner. La fonction *lda()* du package MASS permet de faire cela. Voici un exemple pour discriminer les fromages au lait cru et les fromages pasteurisés.

```
#Chargement du package MASS
library(MASS)
## NB la matrice de données doit être transformée
LDA<-lda(x=t(phylum_matrix),grouping=metadata$Pasteurized)
```

```
## Warning in lda.default(x, grouping, ...): variables are collinear
```

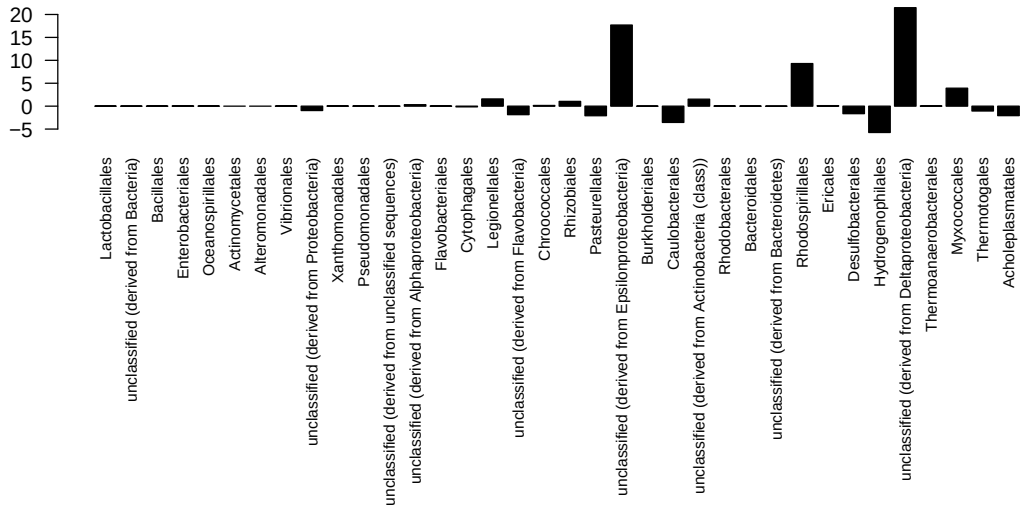
On va ensuite voir si notre droite discriminante sépare bien nos deux groupes en appliquant la régression à chacune de nos valeurs observées :

```
## Calcul des valeurs pour chaque groupe
LDA1_RawMilk<-colSums(apply(phylum_matrix[,metadata$Pasteurized=='N'],2,
                            function(x){LDA$scaling*x}))
LDA1_Pasteurized<-colSums(apply(phylum_matrix[,metadata$Pasteurized=='Y'],2,
                                function(x){LDA$scaling*x}))
#Représentation sous forme d'histogramme
hist(LDA1_Pasteurized,xlim=c(min(LDA1_RawMilk)-4,max(LDA1_Pasteurized)+1),col='green',
     xlab='LD1 means',ylab='Frequency',main='Predicted Values')
hist(LDA1_RawMilk,col='red',add=T)
legend('topleft',pch=19,col=c('red','green'),legend = c('Raw','Pasteurized'))
```



On voit que les 2 fromages sont bien discriminés en appliquant les coefficients de notre fonction discriminante, il n'y a pas de chevauchement des valeurs. On observe également plus de variance pour les laits crus comme on pouvait le suspecter. Nous pouvons maintenant regarder les coefficients les plus élevés qui nous permettent de voir quels taxons discriminent nos fromages.

```
par(mar=c(15,5,5,5))
barplot(LDA$scaling[,1],names.arg = rownames(LDA$scaling),las=2,
        col='black',cex.names = 0.76)
```

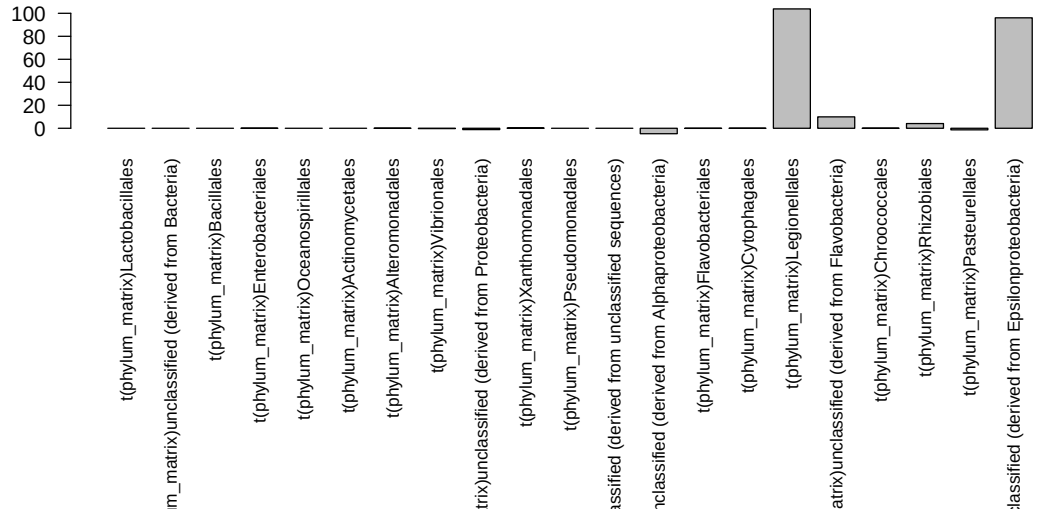


Faites maintenant la même analyse discriminante sur les variables catégorielles du type de croûte et de lait.

Les analyses précédentes ont été faites sur des données non normalisées, sur les comptes bruts. Une façon de normaliser les données est d'utiliser les valeurs des axes ACP au lieu d'utiliser les valeurs de chaque taxon. On peut aussi travailler sur la composition ou bien sur une normalisation plus classique (fonction *scale()*).

Pour associer les variables quantitatives au variations des OTUs, nous pouvons faire de simple régressions linéaires multiples. Par exemple pour le pH :

```
## On ajuste le modèle
LM<-lm(metadata$pH ~ t(phylum_matrix))
## On regarde les coefficients
par(mar=c(15,5,5,5))
barplot(LM$coefficients[!is.na(LM$coefficients)][-1],las=2,cex.names = 0.8)
```



On observe beaucoup de coefficients non estimés car beaucoup de taxons ne sont présents que dans un échantillon. **Il faut alors veiller à les retirer de l'analyse. Faites maintenant une régression linéaire sur les variables NaCL et Humidité**.

À partir de ces analyses, qu'en déduisez vous sur les taxons bactériens responsables de la diversité des fromages ?

Si vous avez le temps, recommencez l'analyse au niveau de la famille ou du genre bactérien.