# Université Claude Bernard - Lyon 1

---

# MADG Année 2016

---

Pr Jean R. Lobry

## NOM Prénom .....

## Première partie

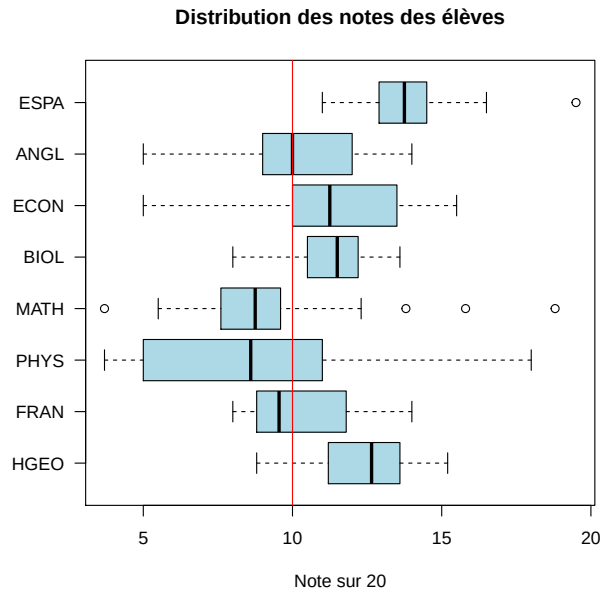
On s'intéresse aux résultats de 22 élèves de seconde dans 8 matières. Les élèves sont notés de 0 à 20, les valeurs élevées correspondent aux bonnes notes. Il faut avoir 10 pour avoir la moyenne. Les données sont les suivantes :

```
library(ade4)
data(seconde)
seconde
```

```
   HGEO FRAN PHYS MATH BIOL ECON ANGL ESPA
1  11.6  8.7  4.5  7.6  9.0  6.5   10 15.5
2  13.6 12.3  6.2  8.5 12.0 11.5    7 12.0
3  13.2 12.1  8.5  6.3 11.6 11.0   13 14.5
4   8.8  8.2  5.0  3.7 10.6 10.0   11 14.5
5  12.7  8.6 10.0  9.3 10.6 14.0    9 13.5
6  12.4 10.0  5.5  9.2 10.5 11.0   10 15.0
7  12.6  9.3  9.5  9.6 10.5 12.5    9 11.0
8  14.4 14.0 15.0 18.8 11.4 15.0   14 14.5
9  14.0 10.6  7.2  8.8 12.2 15.0   12 19.5
10 13.2 13.1 18.0 12.3 12.4  9.5    9 13.0
11 10.6  8.0  4.5  7.1  9.8  8.0   14 16.5
12 13.4  8.8 13.0 13.8 12.6 10.0    9 11.5
13 14.0  9.0  8.7  7.7 11.1 13.5    9 12.6
14  9.7  9.6  4.0  5.5 12.0 12.0    7 14.0
15 15.2 13.8 11.0  8.7  8.0 15.5   14 12.8
16 13.4 11.8 11.0  9.5 13.6 13.5   11 12.9
17 11.4  8.8 11.0  7.2 12.2 12.5   11 13.0
18 11.2  8.8  3.7  8.1 11.8 10.5    5 14.0
19  9.0  8.2  5.7  8.3  8.5  5.0   10 14.0
20 11.4  9.5 10.0 10.2 12.8 13.5   10 13.0
21  9.7  9.6  3.7  9.1  9.5  6.0    8 13.5
22 13.6 10.2 12.0 15.8 13.2 10.0   13 14.5
```

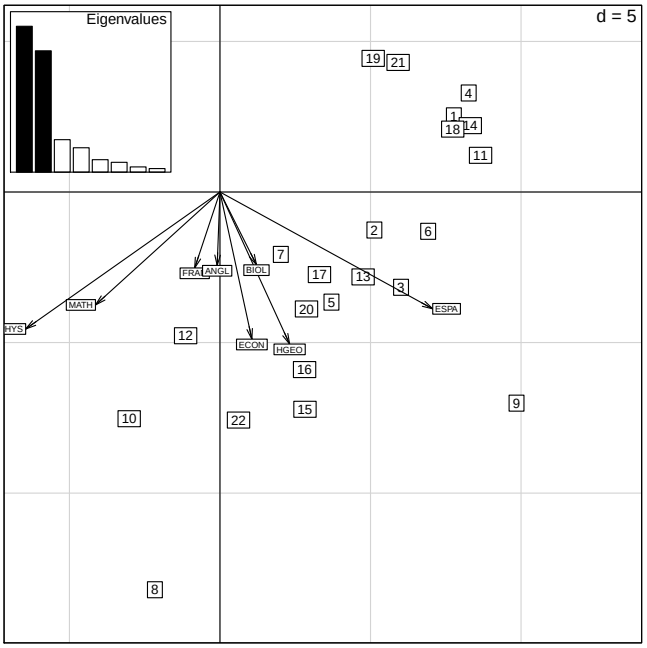
La distribution des notes des élèves est la suivante :

```
boxplot(seconde, horizontal = TRUE, col = "lightblue", las = 1,
  xlab = "Note sur 20", main = "Distribution des notes des élèves")
abline(v = 10, col = "red")
```




On réalise une ACP centrée sur 10 et non réduite pour analyser la structure de ce jeu de données :

```
scatter(dudi.pca(seconde, scann = FALSE, nf = 2, scale = FALSE,
  center = rep(10, 8)), clab.col = 0.5)
```



Au vu du graphe des valeurs propres, combien de facteurs rentiendriez-vous ?

**Réponse :** deux facteurs dominent le graphe des valeurs propres.

Comment interprétez-vous le premier facteur ?

**Réponse :** le premier facteur oppose les matières difficiles (à gauche) aux matières faciles (à droite).

Comment interprétez-vous le deuxième facteur ?

**Réponse :** c'est un effet taille qui oppose les mauvais étudiants (en haut) aux bon étudiants (en bas).

Quel est le numéro du meilleur élève ?

**Réponse :** 8

Quelle est la particularité de l'élève numéro 9 ?

**Réponse :** bilingue espagnol.

Quelles matières conseilleriez-vous à un élève moyen de travailler en priorité pour améliorer ses résultats ?

**Réponse :** la physique puis le français et les maths.

# Seconde partie

On s'intéresse à l'usage du code pour les séquences codantes de *Rickettsia prowazekii*, une bactérie qui d'un point de vue évolutif est proche des mitochondries. On construit la table `tuco` donnant les fréquences absolues des 64 codons pour les 834 séquences codantes :

```
library(seqinr)
choosebank("emglib")
rp <- query("rp", "SP=Rickettsia prowazekii ET T=CDS ET NO K=PARTIAL")
tuco <- matrix(NA, nrow = rp$nelem, ncol = 64)
rownames(tuco) <- getName(rp)
colnames(tuco) <- words()
for(i in seq(1: rp$nelem)){
  print(paste("Calcul pour la séquence", rownames(tuco)[i]))
  tuco[i, ] <- uco(getSequence(rp$req[[i]]))
}
save(tuco, file = "tuco.rda")
```

```
load(url("https://pbil.univ-lyon1.fr/R/donnees/MADG-2016/tuco.rda"))
dim(tuco)
```

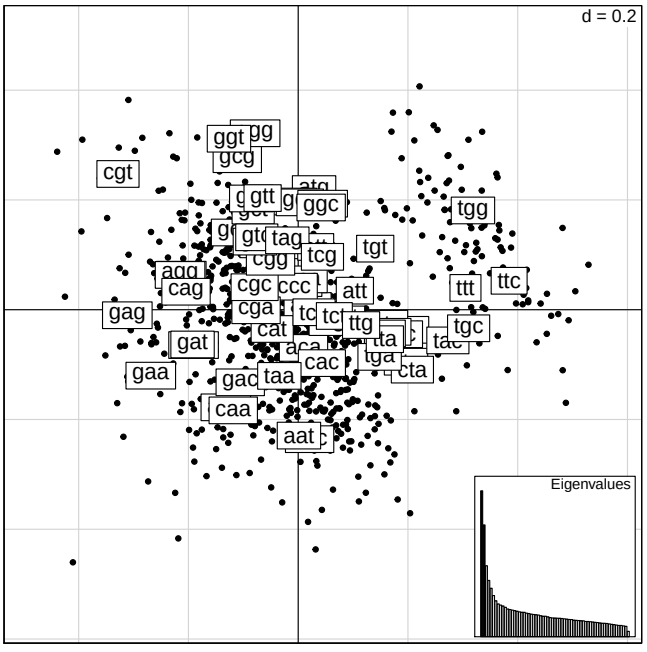
```
[1] 834  64
```

```
head(tuco)
```

```
             aaa aac aag aat aca acc acg act aga agc agg agt ata atc atg att caa
RICPRCG.RP001  20   0   2  20   7   1   0   6  10   1   1   1  18   0   5  16  10
RICPRCG.RP002  12   2   0  10   5   0   1   2   0   3   1   1   6   2   6   6   4
RICPRCG.RP003  24   2   3  12   3   1   0   1   2   2   0   6  19   1   7  15   4
RICPRCG.RP004   9   1   4   4   1   0   2   3   6   0   2   9  19   2   8  20   9
RICPRCG.RP005  29   2   6  13   1   1   2   4   4   2   4   2  10   2   4  13   9
RICPRCG.RP006  53   5  15  50  14   4   2  18   7   7   3  11  32   3  13  37  17
             cac cag cat cca ccc ccg cct cga cgc cgg cgt cta ctc ctg ctt gaa gac
RICPRCG.RP001   1   0   5   5   0   1   3   2   0   0   0   2   2   0   2  16   1
RICPRCG.RP002   0   1   0   2   0   1   3   0   0   0   1   1   2   0   2   9   0
RICPRCG.RP003   0   4   7   4   0   1   2   0   1   0   1   5   0   1   7  13   1
RICPRCG.RP004   0   0   1   7   0   3   5   1   0   0   1   2   0   0   1   1   0
RICPRCG.RP005   1   3   5   3   0   1   3   1   0   1   2   7   2   2   3   9   3
RICPRCG.RP006   2   6  10   9   1   3  14   2   1   1   6   5   0   1  17  36   2
             gag gat gca gcc gcg gct gga ggc ggg ggt gta gtc gtg gtt taa tac tag
RICPRCG.RP001   3  12   5   0   1   8   3   0   0   4   8   1   2   4   1   1   0
RICPRCG.RP002   1  10   1   0   0   0   1   0   0   3   6   0   0   1   0   0   0
RICPRCG.RP003   3  13   4   0   0   4   5   1   2  11   5   1   0   4   0   0   0
RICPRCG.RP004   3   5   7   0   1   2   5   1   0   2   8   0   2  12   0   1   0
RICPRCG.RP005   4  26   3   1   3   4   2   0   0   3   6   0   1   5   0   1   1
RICPRCG.RP006  13  26  13   3   2  16  14   3   2  22  16   2   4  14   1   5   0
             tat tca tcc tcg tct tga tgc tgg tgt tta ttc ttg ttt
RICPRCG.RP001  13   5   1   0  10   0   2   2   4  14   1   3   8
RICPRCG.RP002   2   1   0   2   3   1   0   3   3   4   1   2   3
RICPRCG.RP003   8   8   2   1   4   1   0   0   3  10   1   1  12
RICPRCG.RP004  13   4   0   1  11   1   1   5   2  23   3   3  22
RICPRCG.RP005  14   5   2   2   7   0   1   2   5  19   1   3  21
RICPRCG.RP006  27   6   2   1  12   0   2   4  10  32   4  15  29
```

Le résultat de l'analyse factorielle des correspondances est le suivant.

```
tuco.coa <- dudi.coa(tuco, scannf = FALSE, nf = 5)
scatter(tuco.coa, clab.row = 0, posieig = "bottomright")
```

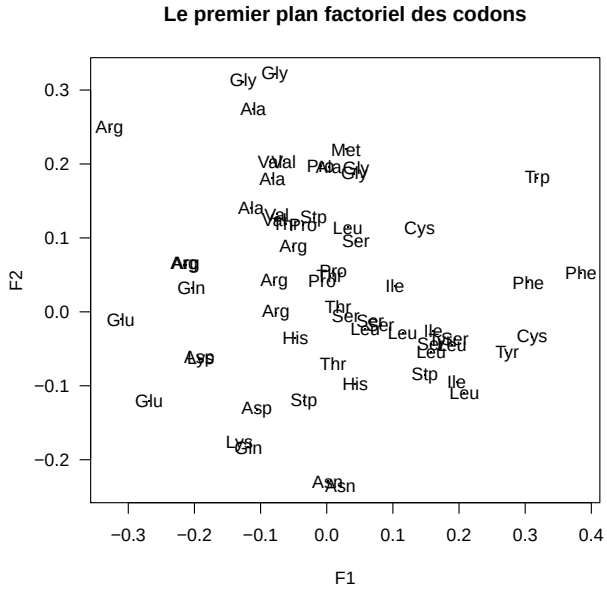



On cherche à décrire et à interpréter les trois premiers facteurs.

## Le premier facteur

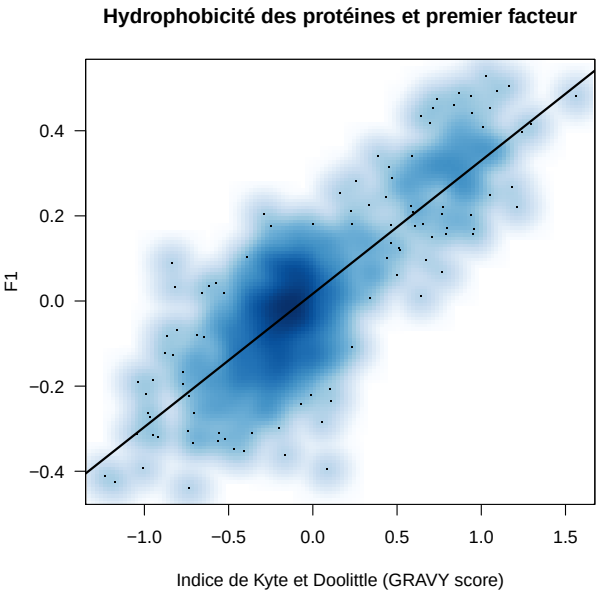
Pour aider l'interprétation on traduit les codons sur le premier plan factoriel :

```
aaname <- sapply(colnames(tuco), function(x) aaa(translate(s2c(x))))
x <- tuco.coa$co[, 1] ; y <- tuco.coa$co[, 2]
plot(x, y, pch = ".", main = "Le premier plan factoriel des codons",
  las = 1, xlab = "F1", ylab = "F2")
text(x, y, aaname)
```




Toujours pour aider l'interprétation on compare les coordonnées factorielles des gènes sur le premier facteur avec l'indice d'hydrophobicité de Kyte et Doolittle :

```
data(EXP)
rtuco <- tuco/rowSums(tuco)
kd <- rtuco %*% EXP$KD
y <- tuco.coa$li[, 1]
smoothScatter(kd, y,
  main = "Hydrophobicité des protéines et premier facteur",
  xlab = "Indice de Kyte et Doolittle (GRAVY score)", ylab = "F1", las = 1)
abline(lm(y~kd), lwd = 2)
```




Décrire le premier facteur.

**Réponse :** un gradient qui oppose les gènes codant pour des protéines hydrophiles à ceux codant pour des protéines hydrophobes.

Quel phénomène biologique pourrait-être à l'origine du premier facteur ? Préciser s'il s'agit d'une pression de sélection ou bien de mutation.

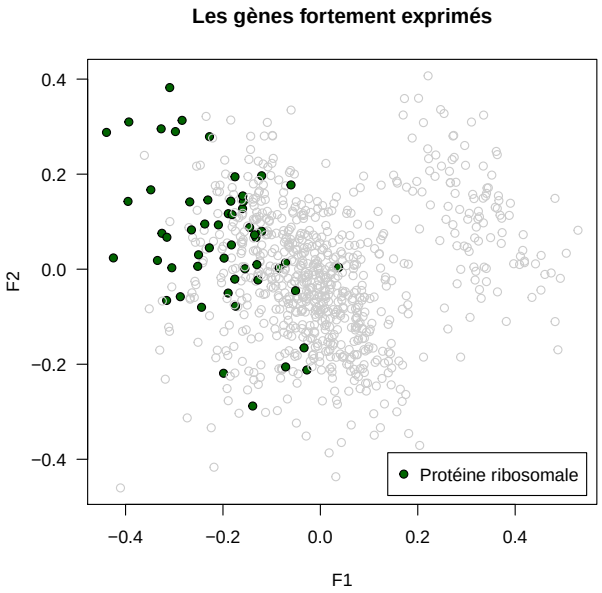
**Réponse :** une pression de sélection pour le maintien de la localisation subcellulaire.

## Le deuxième facteur

Pour aider l'interprétaion de ce facteur on récupère le nom des gènes codant pour des protéines ribosomales, comme proxy des gènes fortement exprimés, et on les représente sur le premier plan factoriel :

```
rib <- query("rib", "rp et k=@ribosomal@protein@")
ribnames <- getName(rib)
save(ribnames, file = "ribnames.rda")
```

```
load(url("https://pbil.univ-lyon1.fr/R/donnees/MADG-2016/ribnames.rda"))
mescouleurs <- ifelse(rownames(tuco) %in% ribnames, "darkgreen", "transparent")
mespourtours <- ifelse(rownames(tuco) %in% ribnames, "black", grey(0.8))
x <- tuco.coa$li[, 1] ; y <- tuco.coa$li[, 2]
plot(x, y, bg = mescouleurs, pch = 21, xlab = "F1", ylab = "F2", las = 1,
   col = mespourtours,
   main = "Les gènes fortement exprimés")
legend("bottomright", inset = 0.02, legend = "Protéine ribosomale",
  pch = 21, pt.bg = "darkgreen")
```




Décrire le deuxième facteur.

**Réponse :** un gradient qui oppose les gènes codant pour des protéines fortement exprimées aux autres.

Quel phénomène biologique pourrait-être à l'origine du deuxième facteur ? Préciser s'il s'agit d'une pression de sélection ou bien de mutation.

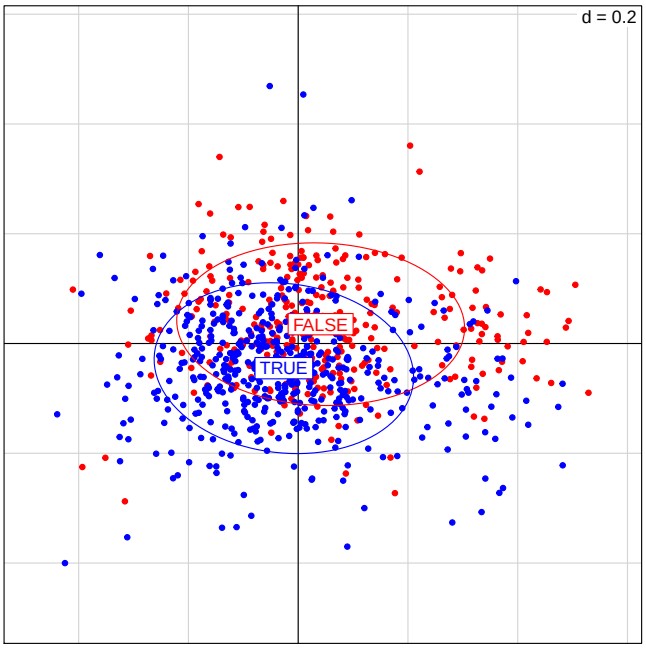
**Réponse :** une pression de sélection pour l'optimisation de l'usage du code pour la traduction.

## Le troisème facteur

Pour aider l'interprétation du troisème facteur on récupère la localisation des gènes sur le brin précoce ou tardif de la réplication, et on représente cette variable illustrative sur le plan F1 × F3 de l'AFC :

```
leading <- logical(rp$nelem)
for(i in seq(1:rp$nelem)){
  print(paste("Traitement de la séquence numéro", i))
  annot <- getAnnot(rp$req[[i]], nbl = 8)
  if(length(grep("leading", annot)) == 1) leading[i] <- TRUE
}
save(leading, file = "leading.rda")
```

```
load(url("https://pbil.univ-lyon1.fr/R/donnees/MADG-2016/leading.rda"))
s.class(tuco.coa$li, as.factor(leading), xax = 1, yax = 3,
  axesell = FALSE, cstar = 0, col = c("red", "blue"))
```




Décrire le troisième facteur.

**Réponse :** l'opposition entre les gènes situés sur le brin précoce ou tardif pour la réplication.

Quel phénomène biologique pourrait-être à l'origine du troisième facteur ? Préciser s'il s'agit d'une pression de sélection ou bien de mutation.

**Réponse :** une pression de mutation asymétrique entre les deux brins de l'ADN.

On remarque qu'il y a plus de gènes sur le brin précoce que sur le brin tardif :

```
 length(leading)
[1] 834
 100*sum(leading)/length(leading)
[1] 61.6307
 100*sum(!leading)/length(leading)
[1] 38.3693
```

Quel phénomène biologique pourrait-être à l'origine de ce déséquilibre ? Préciser s'il s'agit d'une pression de sélection ou bien de mutation.

**Réponse :** une pression de selection pour que les gènes soient transcrits dans le même sens que celui de la progression de la fourche de réplication.

Quelle analyse complémentaire pourrait-on faire pour confirmer cette hypothèse ?

**Réponse :** tester si ce biais est exacerbé pour les gènes fortement exprimés.