Exercices avec le logiciel R

Épreuve MADG

J.R. Lobry

Première session - automne 2025

Tous documents et ordinateurs autorisés. Échanges strictement interdits. Répondre directement sur le sujet. Le PDF du sujet est disponible à l'URL donnée en pied-de-page, mais vous n'avez normalment pas besoin de reproduire les analyses.

Numéro d'intercalaire :

On s'intéresse ici à la relation entre l'usage du code et le niveau d'expression des gènes chez deux espèces : *Borrelia garinii*, un procaryote, et *Quercus robur*, le chêne pédonculé. On utilisera comme *proxy* du niveau d'expression les gènes des protéines ribosomales dont ont sait qu'ils sont fortement exprimés. Pour toutes les demandes d'interprétation, on précisera la nature sélective ou mutationnelle du phénomène invoqué.

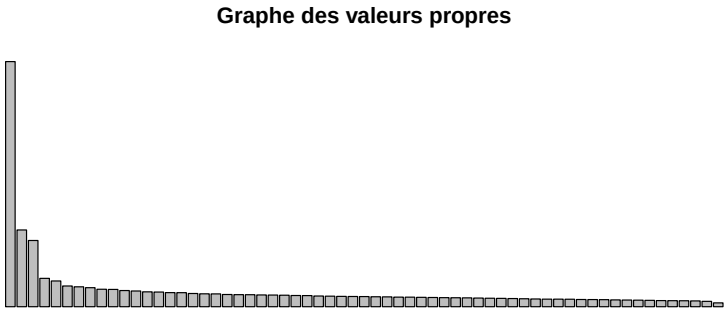
1 *Borrelia garinii*

```
load(url("https://esb.univ-lyon1.fr/donnees/pps067.rda"))
```

Les données[1] ont été compilées par Anamaria Necşulea (M2 - EEB - 2005). Il s'agit d'une table de contingence dans laquelle ont été ventilés 280294 codons en croisant les 64 modalités possibles pour la nature du codon et les 832 modalités possibles pour la nature du gène. On soumet cette table de contingence à une analyse factorielle des correspondances (AFC). Au vu du graphe des valeurs propres, combien de facteurs seriez-vous tentés de vouloir interpréter ?

```
library(ade4)
afcbg <- dudi.coa(pps067$codons, scannf = FALSE, nf = 3)
barplot(afcbg$eig, main = "Graphe des valeurs propres", yaxt = "n")
```

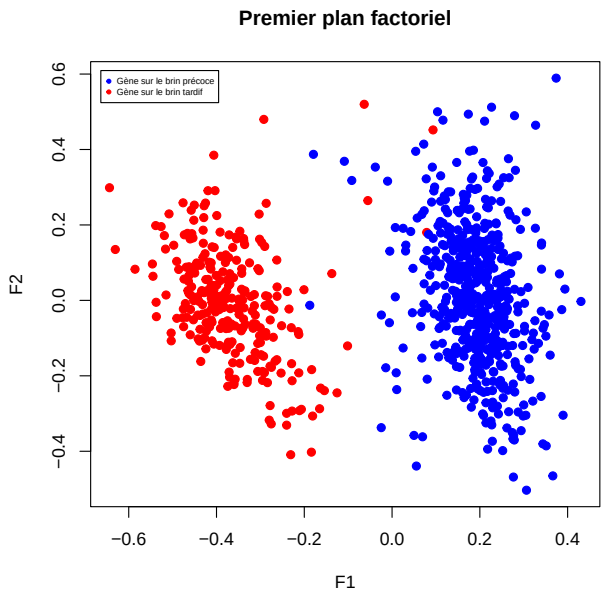
[1] https://esb.univ-lyon1.fr/pdf/pps067.pdf



Réponse:

On représente le premier plan factoriel en portant comme information supplémentaire le sens d'orientation de la transcription des gènes par rapport au sens de progression de la fourche de réplication. Quel phénomène biologique est mis en évidence par le premier facteur de l'AFC ?

```
x <- afcbg$li[, 1] ; y <- afcbg$li[, 2]
plot(x,y, pch = 19, xlab = "F1", ylab = "F2",
  col = ifelse(pps067$info$strand == "leading", "blue", "red"),
  main = "Premier plan factoriel")
legend("topleft", inset = 0.02, pch = 19, col = c("blue", "red"), cex = 0.5,
  legend = c("Gène sur le brin précoce", "Gène sur le brin tardif"))
```



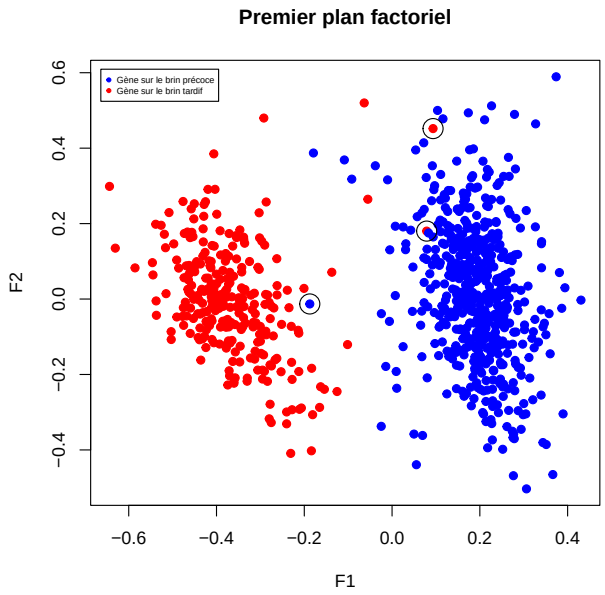
Réponse:

On remarque qu'il n'y a que 279 (34 %) gènes sur le brin tardif contre 553 (66 %) sur le brin précoce. Quelle interprétation donneriez-vous pour expliquer ce déséquilibre ?

Réponse:

Certains gènes, comme par exemple ceux mis en évidence par un cercle dans le graphique ci-après, ne semblent pas être à leur place. Quelle hypothèse pourriez-vous faire quant à leur histoire évolutive récente ?

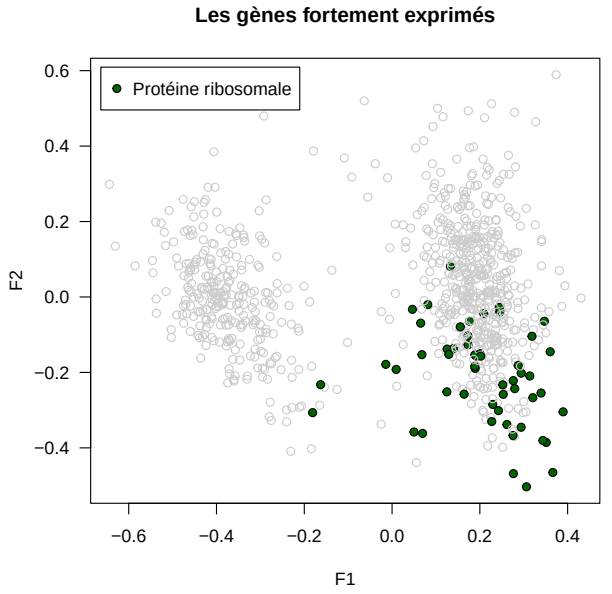
```
x <- afcbg$li[, 1] ; y <- afcbg$li[, 2]
plot(x,y, pch = 19, xlab = "F1", ylab = "F2",
  col = ifelse(pps067$info$strand == "leading", "blue", "red"),
  main = "Premier plan factoriel")
legend("topleft", inset = 0.02, pch = 19, col = c("blue", "red"), cex = 0.5,
  legend = c("Gène sur le brin précoce", "Gène sur le brin tardif"))
ii <- c(430L, 435L, 617L)
points(x[ii],y[ii], pch = 1, cex = 2.5)
```



Réponse:

On porte comme information supplémentaire sur le premier plan factoriel les gènes codant pour des protéines ribosomales. Quelle interprétation biologique pourriez-vous donner au deuxième facteur de l'AFC ?

```
isrib <- pps067$info$Acc %in% pps067$rib
mybg <- ifelse(isrib, "darkgreen", "transparent")
mycol <- ifelse(isrib, "black", grey(0.8))
plot(x, y, bg = mybg, pch = 21, xlab = "F1", ylab = "F2", las = 1,
col = mycol,
main = "Les gènes fortement exprimés")
legend("topleft", inset = 0.02, legend = "Protéine ribosomale", pch = 21, pt.bg = "darkgreen")
```



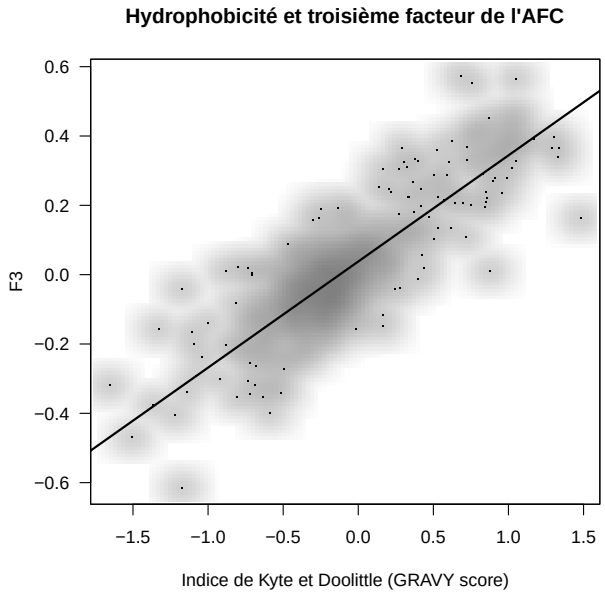
Réponse:

On remarque pour les gènes codant pour des protéines ribosomales qu'il n'y en a que 2 (4 %) sur le brin tardif contre 51 (96 %) sur le brin précoce. Comment interpréteriez-vous cette exacerbation de la tendance générale ?

Réponse:

On confronte le troisième facteur de l'AFC à l'indice d'hydrophobicité de Kyte et Doolittle [1]. Quelle interprétation biologique pourriez-vous donner au troisième facteur de l'AFC ?

```
rtuco <- with(pps067, as.matrix(codons/rowSums(codons)))
library(seqinr) ; data(EXP) ; kd <- rtuco %*% EXP$KD
y <- afcbg$li[,3]
smoothScatter(kd, y,
main = "Hydrophobicité et troisième facteur de l'AFC",
xlab = "Indice de Kyte et Doolittle (GRAVY score)", ylab = "F3", las = 1,
  colramp = colorRampPalette(c("white", grey(0.5))))
abline(lm(y~kd), lwd = 2)
```



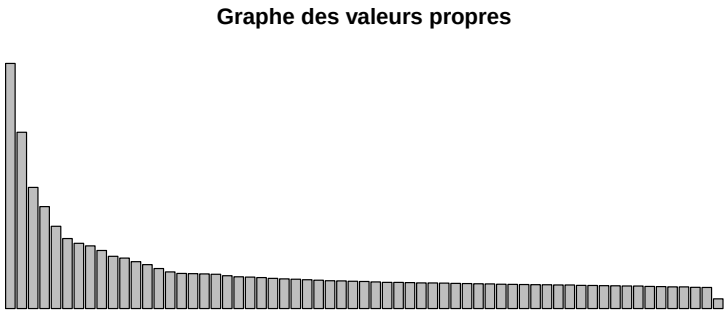
Réponse:

2 *Quercus robur*

```
load(url("https://esb.univ-lyon1.fr/donnees/QuercusCodonUsage/tuco.Rda"))
```

Les données[2] sont issues de la « *treegenesdb* ». Il s'agit d'une table de contingence dans laquelle ont été ventilés 10102937 codons en croisant les 64 modalités possibles pour la nature du codon et les 25808 modalités possibles pour la nature du gène. On soumet cette table de contingence à une analyse factorielle des correspondances (AFC). Au vu du graphe des valeurs propres, combien de facteurs seriez-vous tentés de vouloir interpréter ?

```
afcqr <- dudi.coa(tuco, scannf = FALSE)
barplot(afcqr$eig, main = "Graphe des valeurs propres", yaxt = "n")
```

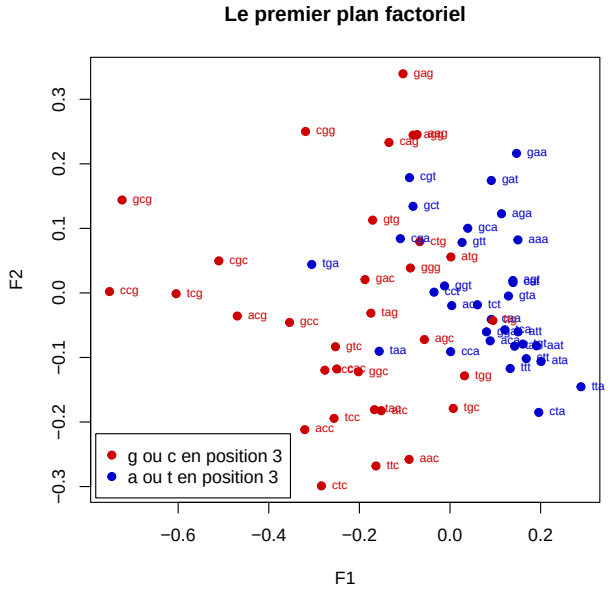


Réponse:

On représente le premier plan factoriel pour les codons en portant comme information supplémentaire s'ils possèdent une base G ou C en *troisième position*. Quelle interprétation biologique pourriez-vous proposer pour le premier facteur de l'AFC ?

```
mycol <- ifelse(substr(colnames(tuco), 3, 3) %in% c("g", "c"), "red3", "blue3")
x <- afcqr$co[ , 1] ; y <- afcqr$co[, 2]
plot(x, y, pch = 19, col = mycol, xlab = "F1", ylab = "F2",
  main = "Le premier plan factoriel")
text(x, y, colnames(tuco), pos = 4, col = mycol, xpd = NA, cex = 0.7)
legend("bottomleft", inset = 0.01, col = c("red3", "blue3"), pch = 19,
legend = c("g ou c en position 3", "a ou t en position 3"))
```

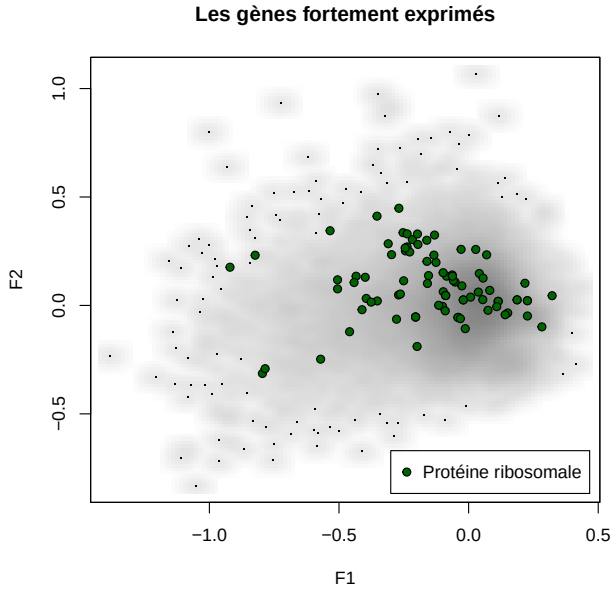
[2]Source : `https://treegenesdb.org/org/Quercus-robur`



Réponse:

On représente les gènes sur le premier plan factoriel en portant comme information supplémentaire les gènes codant pour des protéines ribosomales. Pouvez-vous proposer une interprétation pour le deuxième facteur de l'AFC ?

```
load(url("https://esb.univ-lyon1.fr/donnees/QuercusCodonUsage/annotRib.Rda"))
x <- afcqr$li[ , 1] ; y <- afcqr$li[, 2]
smoothScatter(x, y, main = "Les gènes fortement exprimés", xlab = "F1", ylab = "F2",
  colramp = colorRampPalette(c("white", grey(0.5))))
irib <- which(rownames(tuco) %in% annot.rib$Query.Sequence)
points(x[irib], y[irib], pch = 21, bg = "darkgreen")
legend("bottomright", inset = 0.02, legend = "Protéine ribosomale", pch = 21, pt.bg = "darkgreen")
```



Réponse:

3 Synthèse

DISCUTEZ de la relation entre l'usage du code et le niveau d'expression des gènes au vu des résultats chez *Borrelia garinii* et *Quercus robur*.

Réponse:

References

[1] J. Kyte and R.F. Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, 157:105–132, 1982.