Fiche TD avec le logiciel : tdr623

Pratique de l'AFC inter-intra sur données génomique

J.R. Lobry

Analyse de la variabilité de l'usage du code du point de vue synonyme et non synonyme pour 739 génomes bactériens. Les AFC inter et intra permettent de décomposer une problématique en sous problématiques plus faciles à interpréter.

Table des matières

1 Les données

Les données sont extraites de [1]. Les importer dans R avec :

```
load(url("http://pbil.univ-lyon1.fr/R/donnees/afcinin.rda"))
class(afcinin)
[1] "list"
names(afcinin)
[1] "count"  "topt"   "domain" "gc"     "aa"
```

Il s'agit donc d'une liste avec 5 éléments. Examinons le contenu de tous ces éléments.

1.1 count

```
class(afcinin$count)
[1] "data.frame"
dim(afcinin$count)
[1] 739  61
afcinin$count[1:5, 1:5]
                              aaa  aac  aag  aat  aca
ACHROMOBACTER DENITRIFICANS   216  417  691  149  134
ACHROMOBACTER XYLOSOXIDANS    349  807 1225  283  169
ACIDIANUS AMBIVALENS          756  330  625  519  301
ACIDITHIOBACILLUS FERROOXIDANS 732 897 1252  662  270
ACINETOBACTER BAUMANNII      3442 1233 1429 2590 1271
```

C'est le jeu de données à analyser. Il comporte 739 lignes, une ligne par espèce bactérienne (au sens large : archae + bacteria). Il comporte 61 colonnes, une colonne par codon. Il n'y a pas 64 colonnes ici parce que les trois codons stop (taa, tag et tga) n'ont pas été comptés. Toutes les espèces analysées ici utilisent le code génétique standard pour coder leur protéines. Une entrée (i, j) de la table donne le nombre de codons de type j que l'on a compté dans le génome de l'espèce i. Par exemple :

```
afcinin$count["ACINETOBACTER BAUMANNII", "aat"]
[1] 2590
```

on a compté 2590 codons aat dans le génome de *Acinetobacter baumannii*. Calculer le nombre total d'observation :

```
[1] 559514732
```

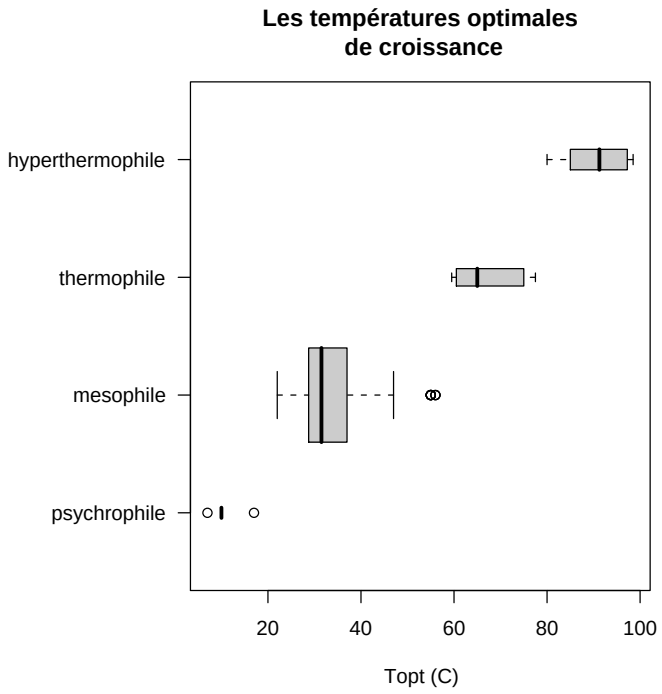
On a donc compté en tout 559,514,732 codons.

1.2 topt

```
class(afcinin$topt)
[1] "data.frame"
dim(afcinin$topt)
[1] 739   2
head(afcinin$topt)
                               topt        typephile
ACHROMOBACTER DENITRIFICANS   30.00        mesophile
ACHROMOBACTER XYLOSOXIDANS    30.00        mesophile
ACIDIANUS AMBIVALENS          80.00 hyperthermophile
ACIDITHIOBACILLUS FERROOXIDANS 31.25       mesophile
ACINETOBACTER BAUMANNII       30.00        mesophile
ACINETOBACTER CALCOACETICUS   32.00        mesophile
```

Cette table contient la température optimale de croissance (`topt`) pour les espèces étudiées. La colonne `typephile` est une variable qualitative ordonnée qui classe les espèces en psychrophiles, mesophiles, thermophiles et hyperthermophiles.

```
opar <- par(no.readonly = TRUE)
par(mar = par("mar") + c(0, 5, 0, 0))
boxplot(afcinin$topt$topt ~ afcinin$topt$typephile, las = 1, varwidth = TRUE,
    horizontal = TRUE, main = "Les températures optimales\n de croissance",
    col = grey(0.8), xlab = "Topt (C)")
par(opar)
```



Quelle est la classe de thermophilie la mieux documentée ? La moins bien documentée ? Quelles valeurs seuil de température ont-elles été utilisée pour définir les classes thermophilie ? Quelle est l'espèce ayant la haute température de croissance optimum ?

```
afcinin$topt[which.max(afcinin$topt$topt), ]
```

```
                  topt        typephile
PYROCOCCUS FURIOSUS 98.5 hyperthermophile
```

Quelle est l'espèce ayant la plus faible température de croissance optimum ?

```
afcinin$topt[which.min(afcinin$topt$topt), ]
```

```
                               topt    typephile
DESULFOTALEA PSYCHROPHILA LSV54    7 psychrophile
```

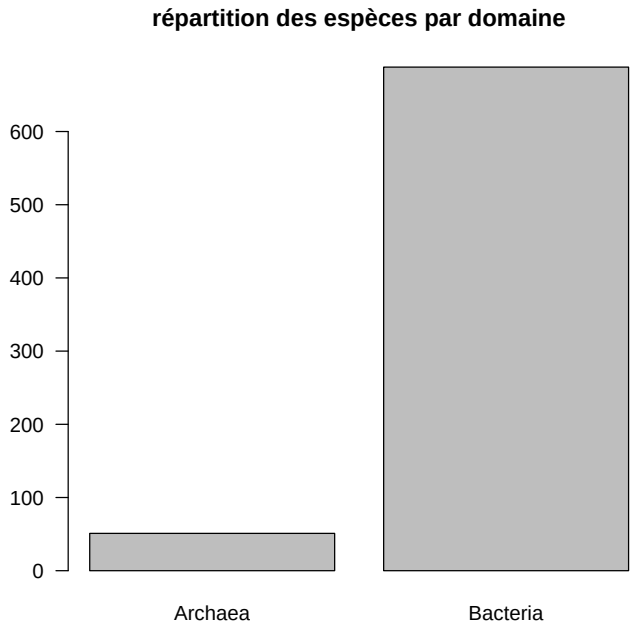
1.3 domain

```
class(afcinin$domain)
```

```
[1] "factor"
 length(afcinin$domain)
[1] 739
 head(afcinin$domain)
[1] Bacteria Bacteria Archaea  Bacteria Bacteria Bacteria
Levels: Archaea Bacteria
```

C'est une variable qualitative non ordonnée à deux modalité : `Bacteria` pour les bactéries au sens strict et `Archaea` pour les archées. Représenter graphiquement les effectifs de chaque modalité :

```
barplot(table(afcinin$domain), main = "répartition des espèces par domaine",
    las = 1)
```

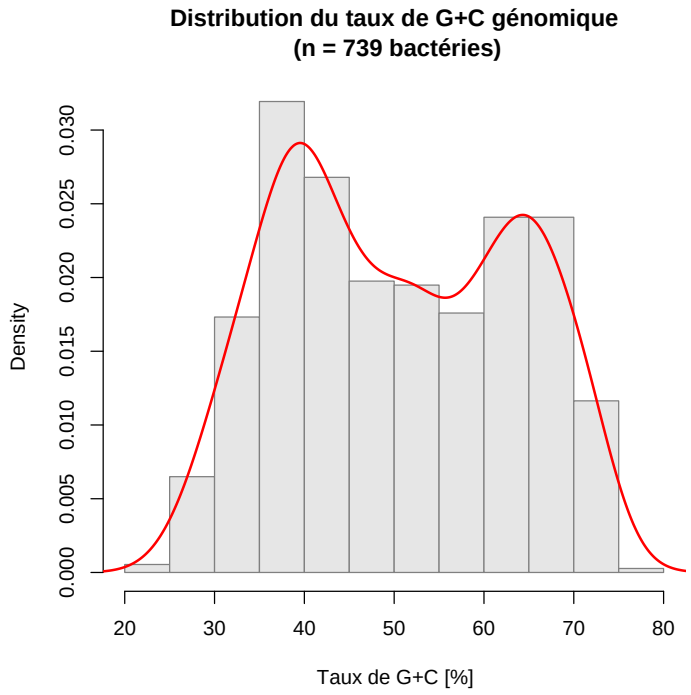


1.4 gc

```
 class(afcinin$gc)
[1] "numeric"
 length(afcinin$gc)
[1] 739
 head(afcinin$gc)
[1] 61.88166 63.37462 37.59371 58.72497 43.92444 42.96045
```

C'est le taux de G+C exprimé en % des 739 génomes bactériens. Représenter graphiquement la distribution du taux de G+C dans le jeu de données :

```
dst <- density(afcinin$gc)
hist(afcinin$gc, col = grey(0.9), border = grey(0.5), proba = TRUE,
    main = "Distribution du taux de G+C génomique\n (n = 739 bactéries)",
    xlab = "Taux de G+C [%]")
lines(dst, lwd = 2, col = "red")
```



1.5 aa

```
class(afcinin$aa)
```

```
[1] "factor"
```

```
length(afcinin$aa)
```

```
[1] 61
```

```
afcinin$aa
```

```
 [1] Lys Asn Lys Asn Thr Thr Thr Thr Arg Ser Arg Ser Ile Ile Met Ile Gln His Gln His
[21] Pro Pro Pro Pro Arg Arg Arg Arg Leu Leu Leu Leu Glu Asp Glu Asp Ala Ala Ala Ala
[41] Gly Gly Gly Gly Val Val Val Val Tyr Tyr Ser Ser Ser Ser Cys Trp Cys Leu Phe Leu
[61] Phe
20 Levels: Ala Arg Asn Asp Cys Gln Glu Gly His Ile Leu Lys Met Phe Pro Ser ... Val
```

C'est une variable qualitative à 20 modalités donnant les acides aminés codés par les 61 codons. Ils sont rangés dans le même ordre que les colonnes de `afcinin$count`. Donner l'acide aminé codé par le codon acg :

```
afcinin$aa[match("acg", names(afcinin$count))]
```

```
[1] Thr
20 Levels: Ala Arg Asn Asp Cys Gln Glu Gly His Ile Leu Lys Met Phe Pro Ser ... Val
```

2 Analyse factorielle des correspondances

2.1 Calculs

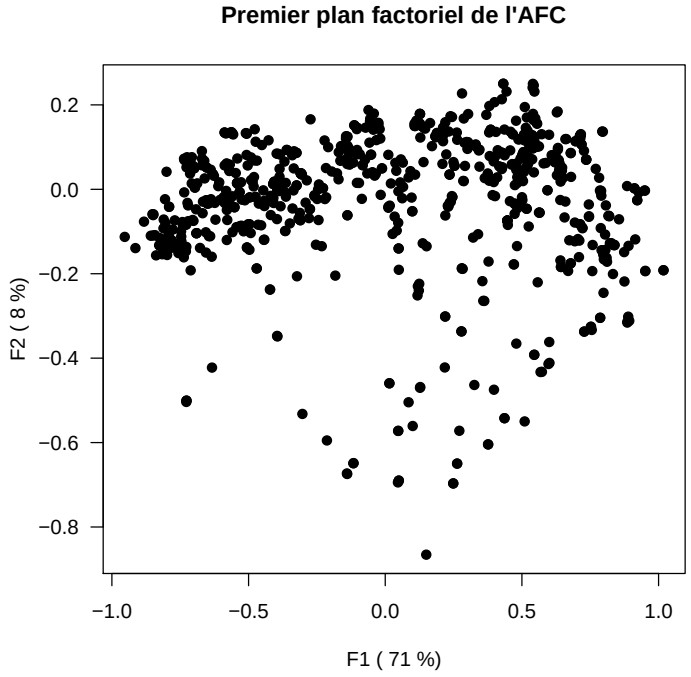
Faire l'AFC du tableau :

```
library(ade4)
afc <- dudi.coa(afcinin$count, scann = FALSE, nf = 2)
```

2.2 Plans factoriels

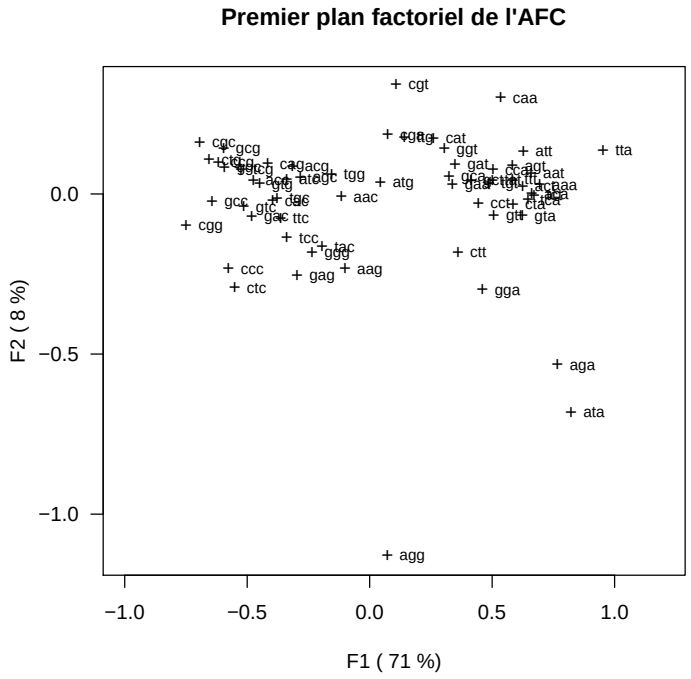
Représenter les individus sur le premier plan factoriel :

```
plot(afc$li[, 1], afc$li[, 2], pch = 19, main = "Premier plan factoriel de l'AFC",
    xlab = paste("F1 (", round(100 * afc$eig[1]/sum(afc$eig)), "%)"),
    ylab = paste("F2 (", round(100 * afc$eig[2]/sum(afc$eig)), "%)"),
    las = 1)
```



Représenter les variables sur le premier plan factoriel :

```
plot(afc$co[, 1], afc$co[, 2], pch = "+", xlim = c(-1, 1.2), main = "Premier plan factoriel de l'AFC",
    xlab = paste("F1 (", round(100 * afc$eig[1]/sum(afc$eig)), "%)"),
    ylab = paste("F2 (", round(100 * afc$eig[2]/sum(afc$eig)), "%)"),
    las = 1)
text(afc$co[, 1], afc$co[, 2], names(afcinin$count), pos = 4, cex = 0.75)
```



2.3 Interprétation

Voici une fonction qui permet de tester si un codon finit par une base C ou G :

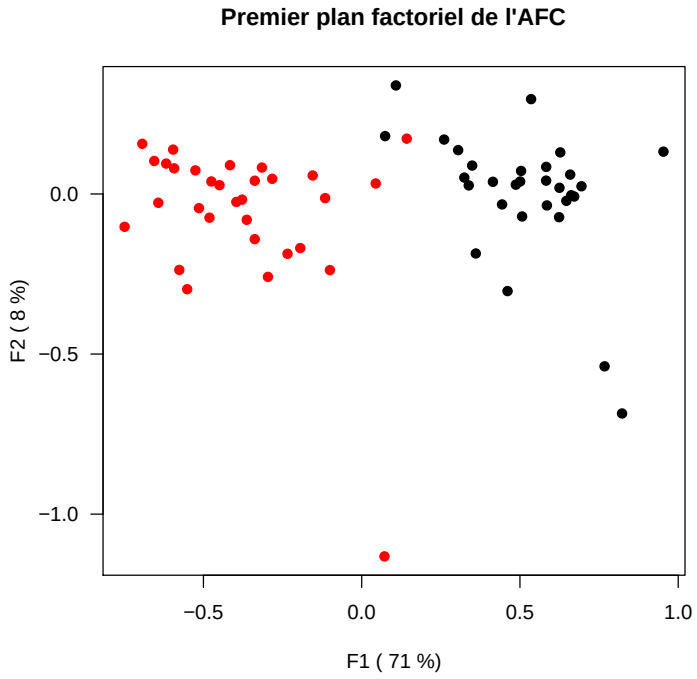
```
isgc3 <- function(codon) {
    codon <- tolower(codon)
    base3 <- unlist(strsplit(codon, split = ""))[3]
    return(base3 == "c" | base3 == "g")
}
isgc3("aaa")
```

```
[1] FALSE
```

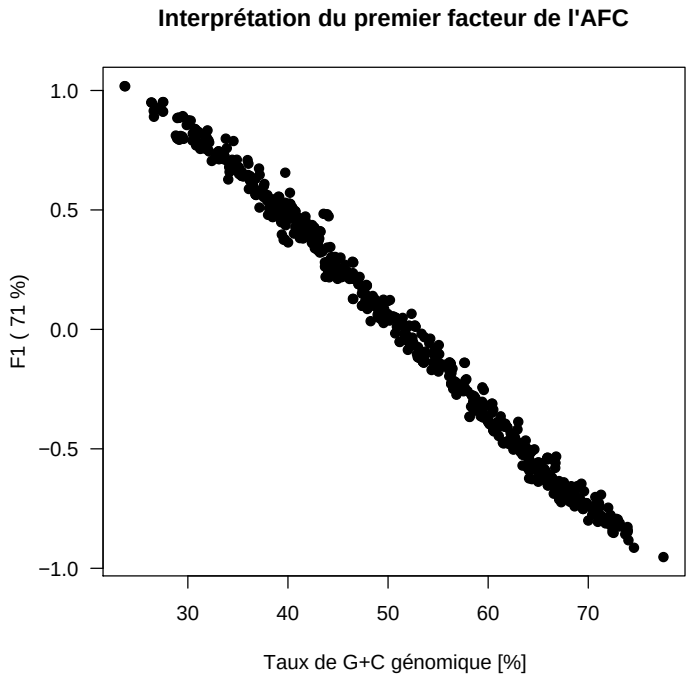
```
isgc3("aac")
```

```
[1] TRUE
```

Utiliser cette fonction pour colorier en rouge les codons finissant par C ou G sur le premier plan factoriel :



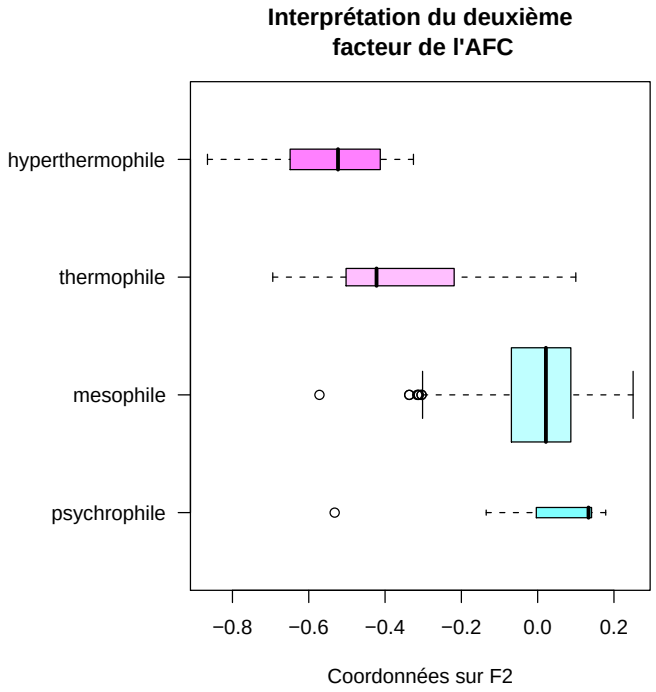
Représenter les coordonnées factorielles des individus sur le premier facteur en fonction du taux de G+C génomique (`afcinin$gc`).



Quelle interprétation donnez vous au premier facteur de l'AFC ? Les résultats sont ils cohérents entre le plan factoriel des individus et des variables ?

Représentez les coordonnées des individus sur le deuxième facteur de l'AFC en fonction des classes de thermophilie. Proposer une interprétation du deuxième facteur.

```
opar <- par(no.readonly = TRUE)
par(mar = par("mar") + c(0, 5, 0, 0))
boxplot(afc$li[, 2] ~ afcinin$topt$typephile, horizontal = TRUE,
    las = 1, main = "Interprétation du deuxième\n facteur de l'AFC",
    col = cm.colors(4), varwidth = TRUE, xlab = "Coordonnées sur F2")
par(opar)
```



3 Analyse factorielle des correspondances inter-groupes

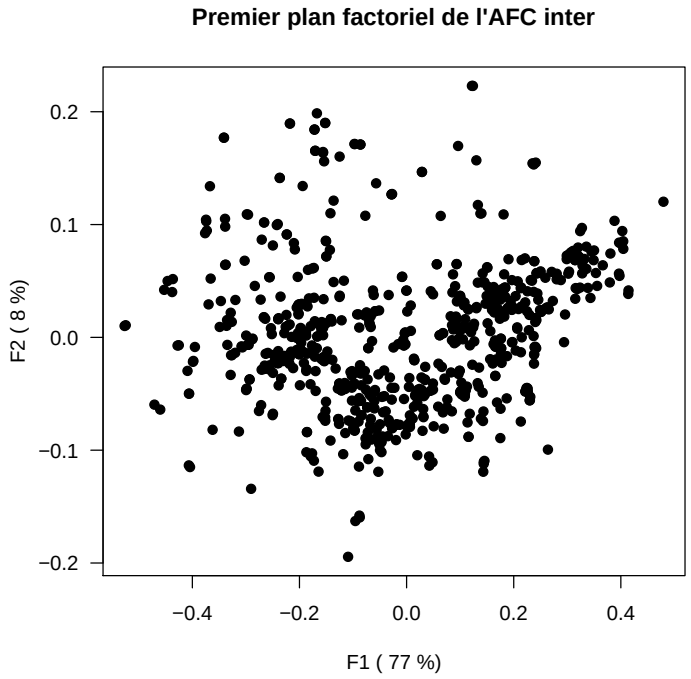
3.1 Calculs

Les acides aminés définissent des groupes de codons. L'analyse inter-groupe correspond donc ici à l'analyse de l'usage des acides aminés, ou encore à la partie non synonyme de la variabilité. La fonction de `ade4` qui permet cette analyse s'appelle `between()`, elle suppose que les groupes sont définis en ligne, d'où l'utilisation ici de la fonction de transposition `t()`.

```
inter <- t(between(t(afc), scann = FALSE, nf = 2, fac = afcinin$aa))
```

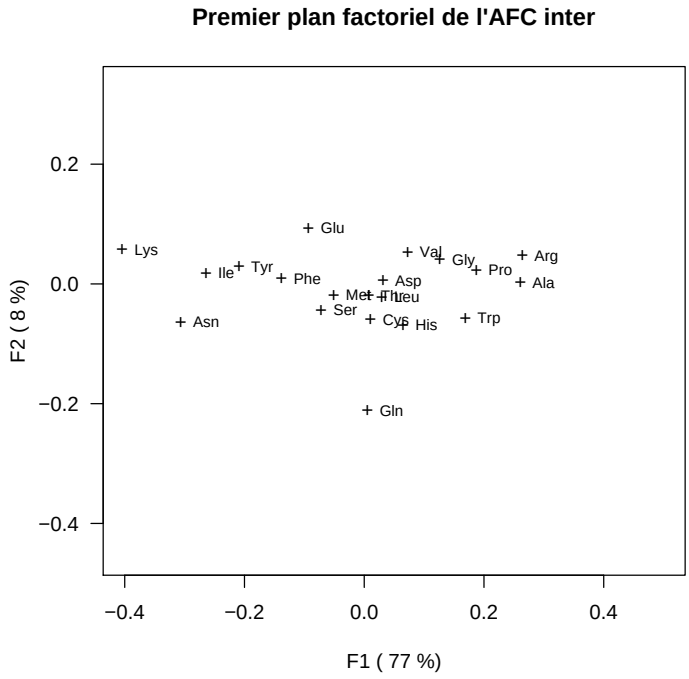
3.2 Plans factoriels

Représenter les individus sur le premier plan factoriel :



Représenter les variables sur le premier plan factoriel :

```
plot(inter$co[, 1], inter$co[, 2], pch = "+", xlim = c(-0.4, 0.5),
    asp = 1, main = "Premier plan factoriel de l'AFC inter", xlab = paste("F1 (",
        round(100 * inter$eig[1]/sum(inter$eig)), "%)"), ylab = paste("F2 (",
        round(100 * inter$eig[2]/sum(inter$eig)), "%)"), las = 1)
text(inter$co[, 1], inter$co[, 2], names(inter$tab), pos = 4, cex = 0.75)
```



3.3 Interprétation

Essayer de donner une interprétation biologique au premier et deuxième facteur de l'AFC inter.

4 Analyse factorielle des correspondances intra-groupes

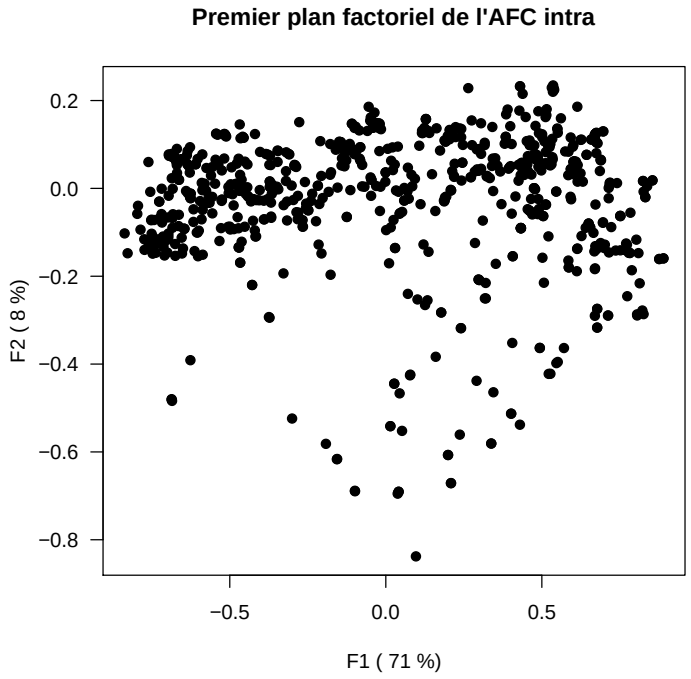
4.1 Calculs

L'analyse intra-groupe correspond ici à l'analyse de l'usage du code synonyme, c'est à dire à la part de la variabilité qui n'est pas visible au niveau des acides aminés. C'est la fonction `within()` de `ade4` qui permet de faire cette analyse :

```
intra <- t(within(t(afc), scann = FALSE, nf = 2, fac = afcinin$aa))
```

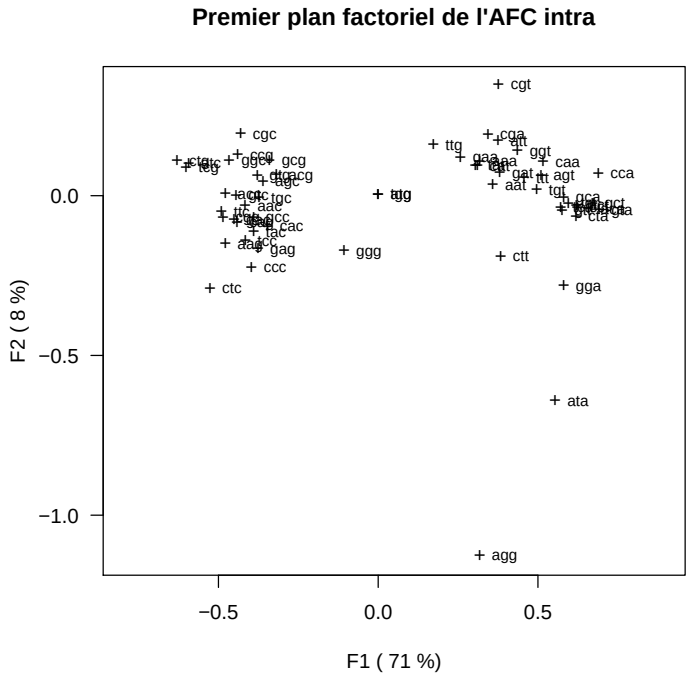
4.2 Plans factoriels

Représenter les individus sur le premier plan factoriel :



Représenter les variables sur le premier plan factoriel :

```
plot(intra$co[, 1], intra$co[, 2], pch = "+", xlim = c(-0.4, 0.5),
    asp = 1, main = "Premier plan factoriel de l'AFC intra", xlab = paste("F1 (",
        round(100 * intra$eig[1]/sum(intra$eig)), "%)"), ylab = paste("F2 (",
        round(100 * intra$eig[2]/sum(intra$eig)), "%)"), las = 1)
text(intra$co[, 1], intra$co[, 2], names(afcinin$count), pos = 4,
    cex = 0.75)
```



4.3 Interprétation

Essayer de donner une interprétation biologique au premier et deuxième facteur de l'AFC intra.

5 Décomposition de la variabilité totale

5.1 Pourquoi tout ça ?

Nous avons analysé la variabilité d'un tableau d'usage du code de trois points de vue différents :

1. Avec l'AFC classique, nous avons analysé la variabilité de l'usage du code sans nous soucier du fait que certains codons codent pour certains acides acide aminés. C'est une analyse globale dans laquelle nous n'avons pas tenu compte de la structuration en groupe des codons.
2. Avec l'AFC inter nous avons oublié les détails de la variabilité intra-groupe, nous avons analysés les groupes en tant que tels. Nous avons en fait analysé la variabilité de l'usage des acides aminés.
3. Avec l'AFC intra nous oublié les différences d'usage des acides aminés pour nous concentrer sur la part de la variabilité qui est synonyme.

Nous avons décomposé la variabilité "proprement", dans le sens où la variance totale est égale à la somme des variance des analyses partielles :

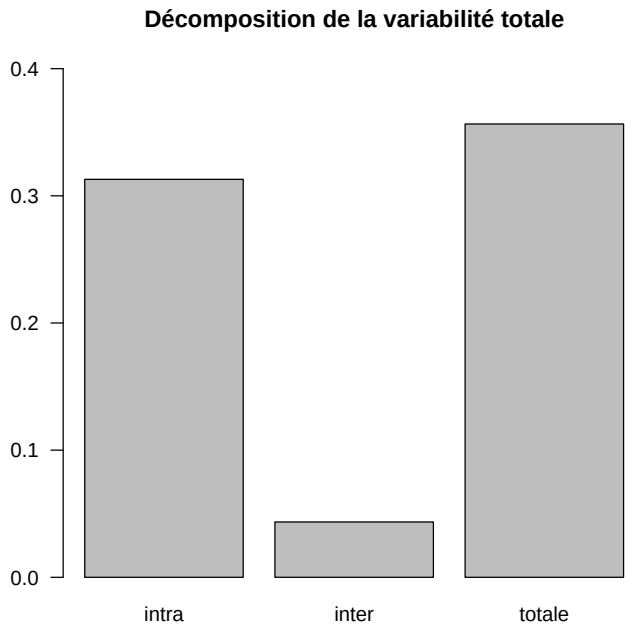
```
sum(afc$eig)
```

```
[1] 0.356484
```

```
 sum(inter$eig)
[1] 0.0435149
 sum(intra$eig)
[1] 0.3129691
 sum(inter$eig) + sum(intra$eig)
[1] 0.356484
 all.equal(sum(inter$eig) + sum(intra$eig), sum(afc$eig))
[1] TRUE
```

L'idée générale est simplement de décomposer un problème en sous problèmes pour mieux cerner les facteurs en jeu, pour faciliter l'interprétation des résultats. Il est intéressant de représenter graphiquement comment la variabilité totale se décompose en variabilité inter et intra, par exemple :

```
barplot(c(sum(intra$eig), sum(inter$eig), sum(afc$eig)), main = "Décomposition de la variabilité totale",
    ylim = c(0, 0.4), las = 1, names.arg = c("intra", "inter", "totale"))
```



Il saute aux yeux que la grosse part de la variabilité est intra. Sachant que la variabilité intra correspond à la variabilité synonyme et que la variabilité inter correspond à la variabilité non-synonyme, expliquer pourquoi il est raisonnable qu'il en soit ainsi d'un point de vue évolutif (vous avez besoin de quelques notions en évolution au niveau moléculaire pour pouvoir répondre à cette question).

5.2 Pour aller plus loin dans l'interprétation

Ne vous contentez pas d'une simple décomposition globale de la variabilité mais comparez dans le détail les structures trouvées au niveau inter et intra (*i.e.* facteur par facteur). C'est relativement "simple" ici parce que les deux premiers facteurs ont la propriété d'être corrélés entre l'analyse inter et intra, ceci est dû à la nature des données, ce n'est pas forcément vrai dans le cas général.

Références

[1] J.R. Lobry and A. Necşulea. Synonymous codon usage and its potential link with optimal growth temperature in prokaryotes. *Gene*, 385 :128–136, 2006.