Analyse en composantes principales

Exercice 2.1. Les crabes

Le jeu de données considéré est constitué de 200 crabes décrits par huit variables (3 qualitatives et 5 quantitatives). Charger le jeu de données et sélectionner les variables quantitatives en utilisant le code R suivant :

```
> library(MASS)
> data(crabs)
> crabsquant<-crabs[,4:8]
```

Cette étude vise à utiliser l'ACP pour trouver une représentation des crabes qui permettent des distinguer visuellement différents groupes, liés à l'espèce et au sexe.

FIGURE 2.1 – l'individu en question

Quelques programmes R utiles :
— `princomp` implémente l'ACP,
— `biplot` qui permet de représenter variables et données dans un plan principal.

1. Testez une ACP sur `crabsquant` sans traitement préalable. Que constatez vous ?
2. Trouvez une solution pour améliorer la qualité de votre représentation en terme de visualisation des différents groupes.
3. Que dire de la qualité de représentation de cette nouvelle ACP ? Combien d'axes retenez-vous ? Pourquoi ?
4. Comment interprétez-vous les axes retenus à partir du cercle des corrélations ?
5. Que pouvez-vous en déduire sur la caractérisation des mâles/femelles, crabes oranges/bleus ?

Solution de l'exercice 2.1. **ACP des crabes**

Analyse descriptive

```
> library(MASS)
> data(crabs)
> n=dim(crabs)[1]
> crabsquant<-crabs[,4:8]
```

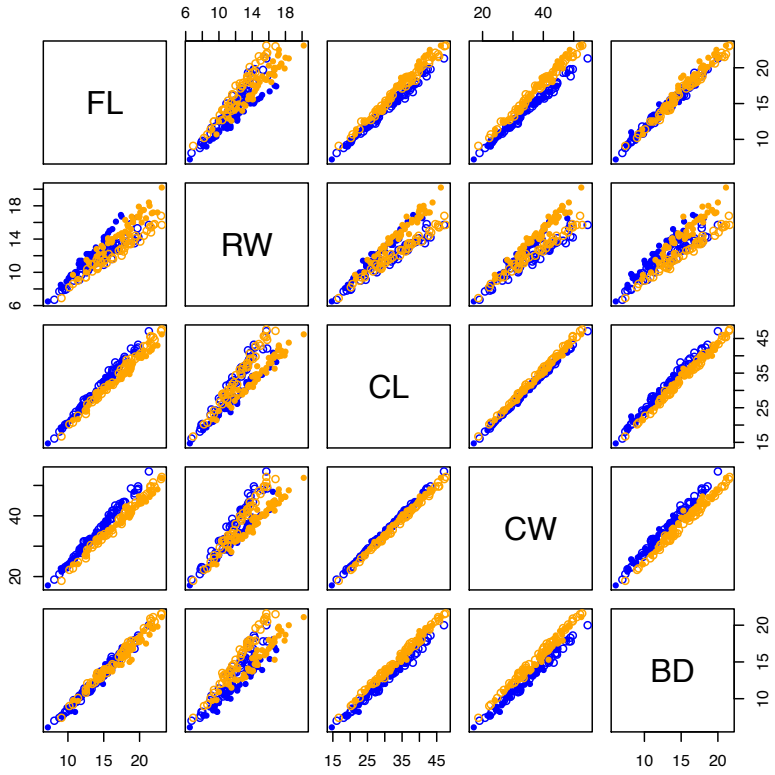
Visualisation des corrélations

```
> cor(crabsquant)
```

```
          FL        RW        CL        CW        BD
FL 1.0000000 0.9069876 0.9788418 0.9649558 0.9876272
RW 0.9069876 1.0000000 0.8927430 0.9004021 0.8892054
CL 0.9788418 0.8927430 1.0000000 0.9950225 0.9832038
CW 0.9649558 0.9004021 0.9950225 1.0000000 0.9678117
BD 0.9876272 0.8892054 0.9832038 0.9678117 1.0000000
```

Les variables sont extrêmement correlées (effet taille à prévoir).

```
> pairs(crabsquant,col=c("blue","orange")[crabs$sp],pch=c(20,21)[crabs$sex])
```



ACP brute

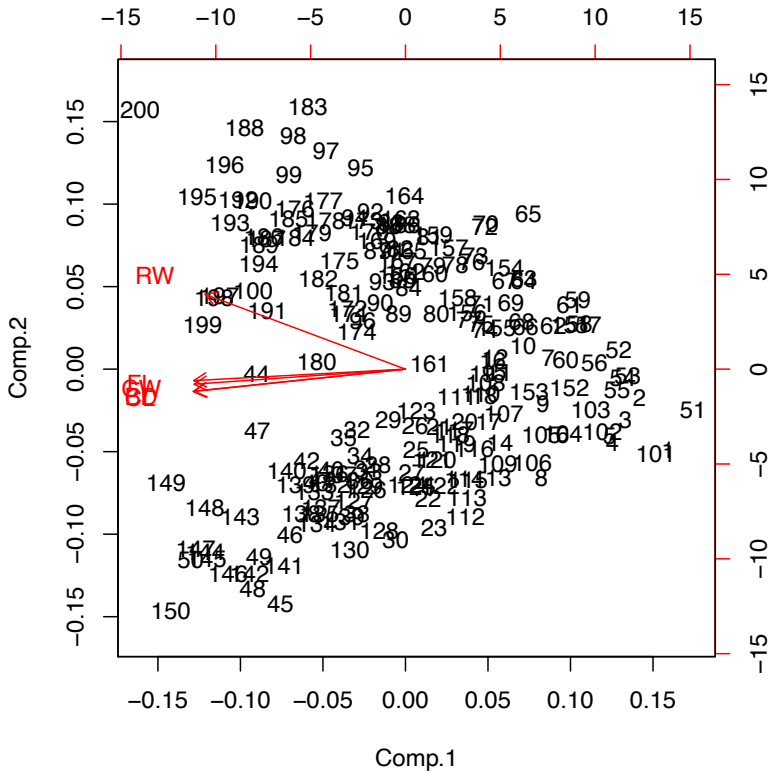
```
> res<-princomp(scale(crabsquant))
```

Le premier axe capture toute la variation du nuage, et les autres n'ont plus rien à dire. On le voit bien si on regarde les valeurs propres : la première composante capture 95% de l'inertie du nuage !

```
> summary(res)
```

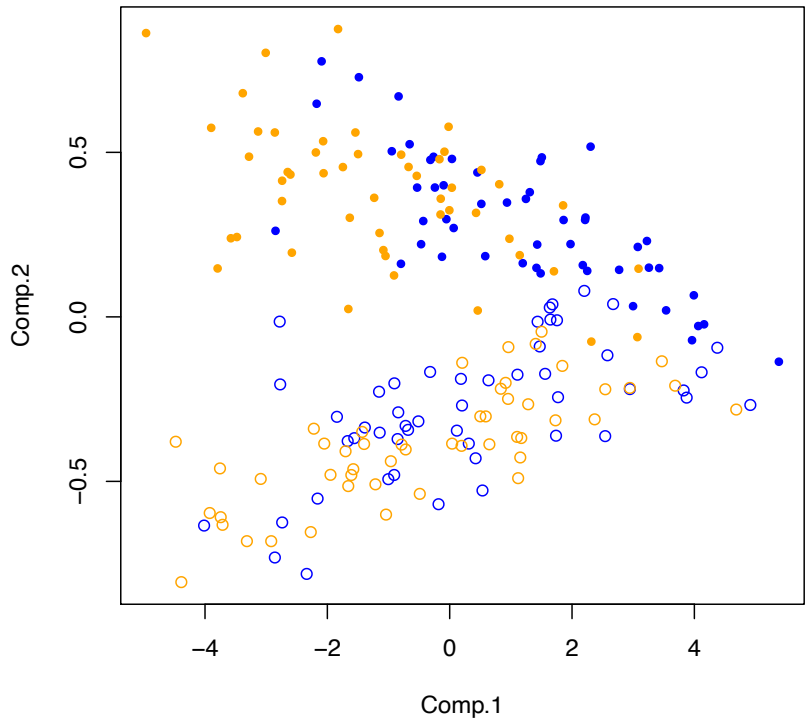
```
Importance of components:
                          Comp.1     Comp.2      Comp.3      Comp.4       Comp.5
Standard deviation     2.182863 0.38849296 0.215406149 0.105260061 0.0412688656
Proportion of Variance 0.957767 0.03033704 0.009326595 0.002227071 0.0003423355
Cumulative Proportion  0.957767 0.98810400 0.997430593 0.999657664 1.0000000000
```

```
> biplot(res)
```



L'effet taille est ici très visible. Toutes les variables sont très corrélées, et l'on ne départage pas bien du tout les crabes, ni du point de vue de l'espèce, ni du point de vue du sexe.

```
> plot(res$scores[,1:2],col=c("blue","orange")[crabs$sp],pch=c(20,21)[crabs$sex])
```



ACP après transformation Faisons une simple petite transformation pour enlever l'effet taille, toutes les variables sont renormalisées par la 3ème (variable la plus corrélée avec les autres).

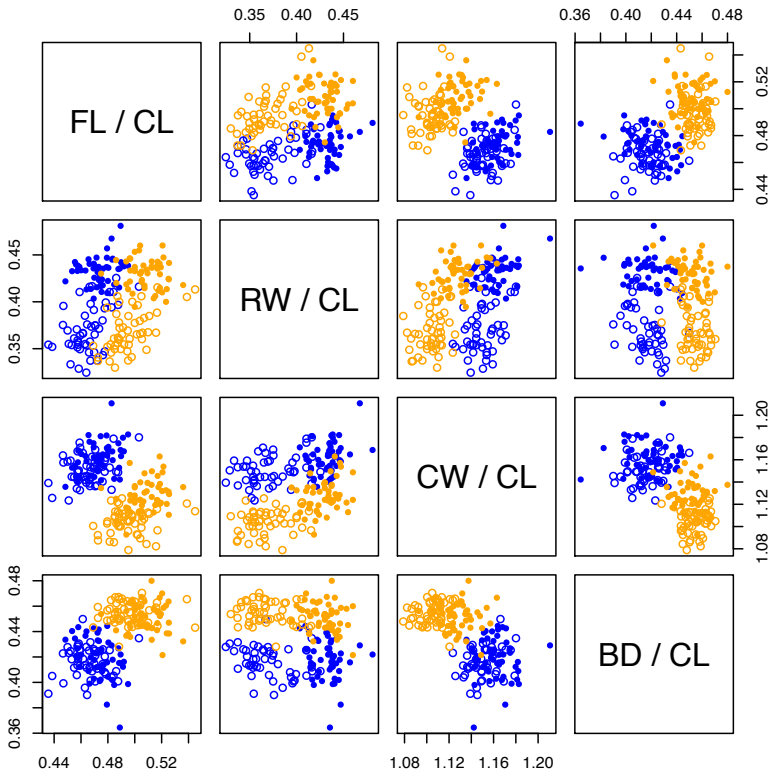
```
> crabsquant2<-(crabsquant/crabsquant[,3])[,-3]
```

On met à jour les noms des variables :

```
> j=0
> for(i in c(1,2,4,5))
+ {
+  j=j+1
+  names(crabsquant2)[j]<-c(paste(names(crabsquant)[i],"/",names(crabsquant[3])))
+ }
```

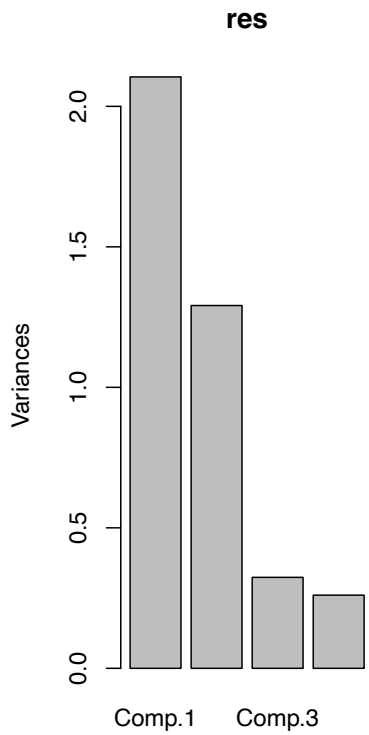
Visualisons à nouveau les corrélations entre variables :

```
> pairs(crabsquant2,col=c("blue","orange")[crabs$sp],pch=c(20,21)[crabs$sex])
```



Effectuons une nouvelle fois l'ACP :

```
> res<-princomp(scale(crabsquant2))
> par(mfrow=c(1,2))
> plot(res)
```



Analyse de la sortie

> str(res)

```
List of 7
 $ sdev    : Named num [1:4] 1.451 1.136 0.569 0.51
  ..- attr(*, "names")= chr [1:4] "Comp.1" "Comp.2" "Comp.3" "Comp.4"
 $ loadings: loadings [1:4, 1:4] -0.51 0.134 0.589 -0.613 -0.485 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:4] "FL / CL" "RW / CL" "CW / CL" "BD / CL"
  .. ..$ : chr [1:4] "Comp.1" "Comp.2" "Comp.3" "Comp.4"
 $ center  : Named num [1:4] -9.48e-16 -5.07e-16 -8.39e-17 1.13e-15
  ..- attr(*, "names")= chr [1:4] "FL / CL" "RW / CL" "CW / CL" "BD / CL"
 $ scale   : Named num [1:4] 1 1 1 1
  ..- attr(*, "names")= chr [1:4] "FL / CL" "RW / CL" "CW / CL" "BD / CL"
 $ n.obs   : int 200
 $ scores  : num [1:200, 1:4] 0.665 1.129 1.907 1.242 0.864 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:200] "1" "2" "3" "4" ...
  .. ..$ : chr [1:4] "Comp.1" "Comp.2" "Comp.3" "Comp.4"
 $ call    : language princomp(x = scale(crabsquant2))
 - attr(*, "class")= chr "princomp"
```

On trouve les éléments suivants :

sdev Ecarts types des composantes principales, soit la racine carrée des valeurs propres.
loadings Matrice des vecteurs propres : axes principaux
center Moyennes utilisées pour le centrage des données.
scale Ecarts-types utilisés pour la réduction des données.
n.obs Nombre d'observations
scores Composantes principales
call Rappel de l'appel fait à `princomp`.

Choix des axes

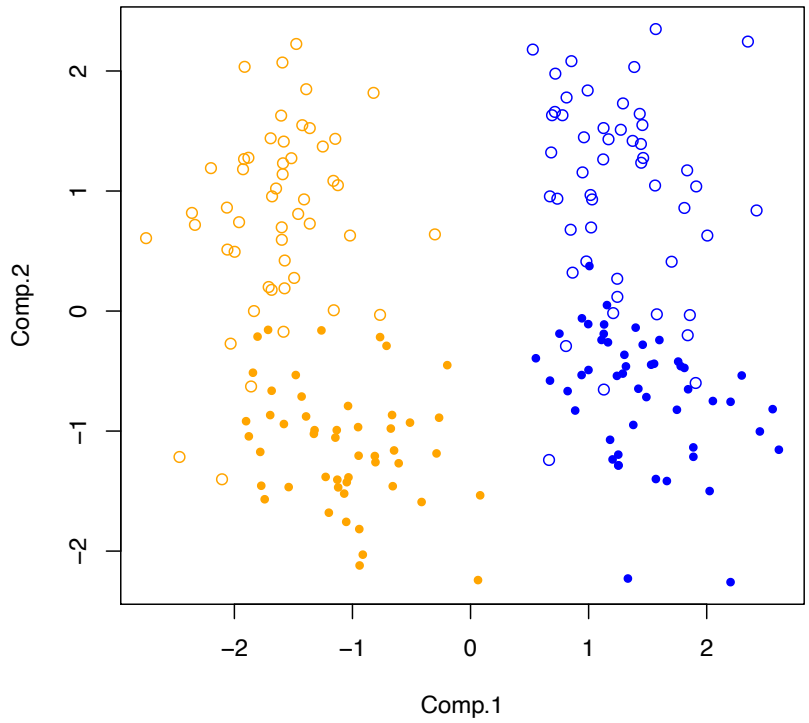
```
> summary(res)

Importance of components:
                          Comp.1    Comp.2     Comp.3     Comp.4
Standard deviation     1.4507641 1.1362749 0.56899999 0.51029581
Proportion of Variance 0.5288233 0.3244022 0.08134698 0.06542759
Cumulative Proportion  0.5288233 0.8532254 0.93457241 1.00000000
```

On conserve ici les deux premiers axes, qui portent tous les deux une variance supérieure à la moyenne (1). Au total, le plan des deux premiers facteurs représentera 85% de la variance totale.

Cette fois, ce plan parvient à très bien distinguer les espèces et les sexes. L'axe 1 sépare les 2 espèces, tandis que l'axe 2 sépare les femelles des mâles.

```
> plot(res$scores[,1:2],col=c("blue","orange")[crabs$sp],pch=c(20,21)[crabs$sex])
```

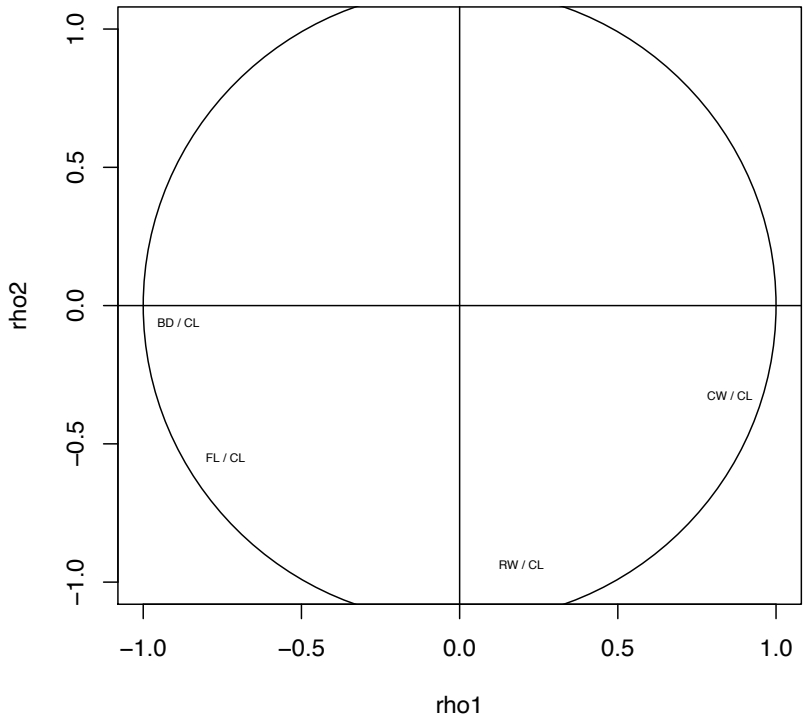


Analyse des variables : construction du cercle des corrélations

```
> # Corrélations variables-facteurs principaux
> rho1 <- res$loadings[,1] * res$sdev[1]
> rho2 <- res$loadings[,2] * res$sdev[2]
> corr <- cbind(rho1,rho2)
> print(corr,digits=2)

         rho1   rho2
FL / CL -0.74 -0.551
RW / CL  0.19 -0.937
CW / CL  0.85 -0.326
BD / CL -0.89 -0.061

> # Affichage du cercle des corrélations
> plot(rho1,rho2, xlim=c(-1,1), ylim=c(-1,1), type = "n")
> abline(h=0,v=0)
> text(rho1,rho2, labels=names(crabsquant2), cex=0.5)
> symbols(0,0,circles=1,inches= F, add=T)
```



Le premier axe est positivement corrélé aux variables BD/CL et FL/CL, négativement corrélé à CW/CL. Le second axe est principalement (négativement) corrélé à RW/CL.

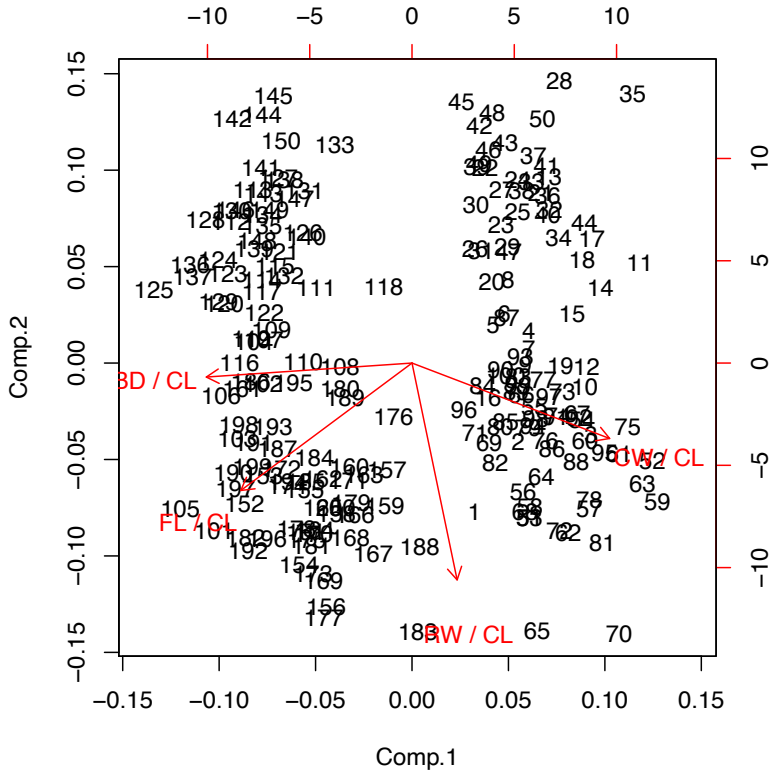
```
> # Carrés des corrélations (cosinus carrés)
> print(corr^2,digits=2)

         rho1   rho2
FL / CL 0.547 0.3031
RW / CL 0.038 0.8781
CW / CL 0.729 0.1063
BD / CL 0.790 0.0037

> # Cumul des carrés des corrélations
> print(t(apply(corr^2,1,cumsum)),digits=2)

         rho1 rho2
FL / CL 0.547 0.85
RW / CL 0.038 0.92
CW / CL 0.729 0.84
BD / CL 0.790 0.79

> biplot(res)
```



Analyse des individus

```
> ctrb <- NULL
> for (k in 1:2){
+   ctrb <- cbind( ctrb,res$scores[,k]^2/res$sdev[k]^2/nrow(crabs))
+ }
> o1 <-order(ctrb[,1],decreasing=T)
> o2 <-order(ctrb[,2],decreasing=T)
> best1 <- cbind(ctrb[o1,1],res$scores[o1,1],crabs$sp[o1],crabs$sex[o1])
> best2 <- cbind(ctrb[o2,2],res$scores[o2,2],crabs$sp[o2],crabs$sex[o2])
> print(best1[1:10,])

          [,1]      [,2] [,3] [,4]
125 0.01793696 -2.747807    2    2
59  0.01619717  2.611148    1    1
52  0.01555194  2.558610    1    1
105 0.01443306 -2.464853    2    2
63  0.01426134  2.450146    1    1
11  0.01393991  2.422377    1    2
136 0.01322529 -2.359470    2    2
35  0.01310833  2.349013    1    2
137 0.01293975 -2.333860    2    2
75  0.01252811  2.296437    1    1
```

```
> print(best2[1:10,])

          [,1]      [,2] [,3] [,4]
28  0.02137378  2.349303    1    2
70  0.01975185 -2.258407    1    1
35  0.01952067  2.245152    1    2
183 0.01945851 -2.241575    2    1
65  0.01922960 -2.228351    1    1
145 0.01917601  2.225244    2    2
45  0.01838197  2.178685    1    2
177 0.01740040 -2.119718    2    1
48  0.01678720  2.082033    1    2
144 0.01660556  2.070738    2    2
```

Exercice 2.2. Single Nucleotide Polymorphism

Le jeu de données considéré est constitué de 5500 SNP concernant 728 individus issus de 6 populations différentes : africaines (YRI, MKK et LWK), indienne (GIH), caucasiennes (CEU, TSI). Charger le jeu de données DonneesSNPnormalisées.RData.

Quelques programmes R utiles :

— `princomp` implémente l'ACP en acceptant qu'il y ait plus de variables que d'observations,

— `biplot` qui permet de représenter variables et données dans un plan principal.

1. Observez ce que contient le jeu de données. Que trouve-t-on dans l'objet `data` que vous venez de charger ? Que représentent ses différents éléments ?
2. Réaliser l'ACP sur la matrice des génotypes.
3. Que constatez-vous sur le nombre de composantes principales ? Comment expliquez-vous cela ?
4. Combien d'axes conservez-vous ? Que pensez-vous de la qualité de représentation ?
5. Représentez les individus dans les plans principaux en distinguant les différentes populations d'origine par des couleurs. Interprétez les axes.

Solution de l'exercice 2.2. A faire

Exercice 2.3. Phylogénie des globines

On se propose d'effectuer une analyse factorielle des dissimilarités de séquence protéique de plusieurs globines issues de différentes espèces et de comparer les résultats obtenus à l'arbre phylogénétique de la Figure 2.2.

1. Télécharger le fichier neighbor_globin.txt et importer les données dans `R` dans un `data.frame` d. Elles contiennent les scores d'alignement deux à deux de diverses globines chez différentes espèces tel que décrit dans le fichier Globines_liste.txt.
2. Vérifier que ces scores correspondent bien à des dissimilarités. Nommer les colonnes.
3. Calculer la matrice Δ des carrés des dissimilarités.

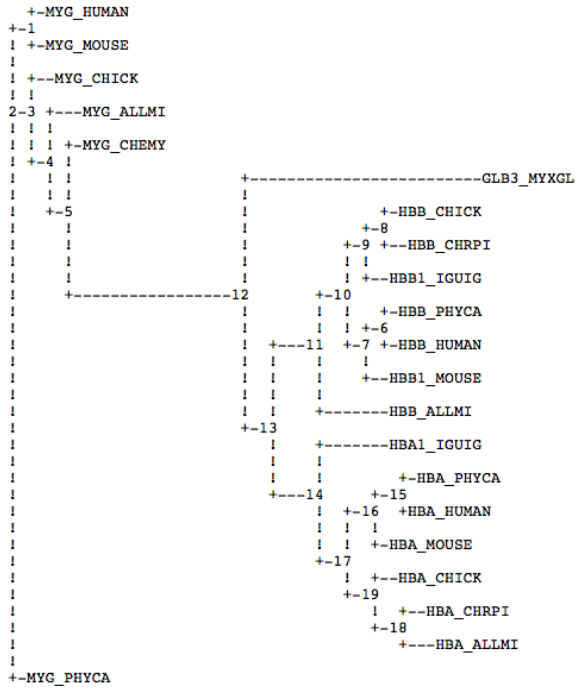


FIGURE 2.2 – Arbre phylogénétique des globines

4. Calculer la matrice de centrage J définie par :

$$J = I - \frac{1}{n} 1_{(n,n)}$$

5. Calculer $B = -\frac{1}{2} J \Delta J$. Comment peut-on interpréter B ?
6. Effectuer la décomposition spectrale de B :

$$B = U \Lambda U^{\intercal}$$

7. Dans cette décomposition, quels sont les facteurs principaux ? Combien en conserverez-vous pour la suite de l'analyse ? Que concluez-vous aussi de l'observation des valeurs propres ?
8. Calculer les composantes principales associées aux axes principaux retenus et représenter les plans principaux correspondants en différenciant les types de globines par type de point (1=myoglobine, 2=hémoglobine β, 3 = hémoglobine α, 4 = globine-3) et les espèces par couleur (on mettra la même couleur pour les deux espèces de tortues).
9. Que remarquez-vous en comparant ces différents plans à l'arbre phylogénétique ?
10. Effectuer la même chose sur les sous-ensembles d'hémoglobines α, β puis de myoglobines, toujours en comparant les résultats à l'arbre phylogénétique. Pour gagner du temps, vous pouvez vous aider de la fonction `cmdscale`.

Solution en R :

```
> d <- read.table("TD-ACP/neighbor_globin.txt",header = FALSE,row.names=1)
> colnames(d) <- rownames(d)
```