Projet 1 de Biostatistiques

Alexandre BEWA & Adèle DESMAZIERES

2023-02-23

Introduction

Nous cherchons à analyser les variations de distances au sein de la protéine PR2, en fonction de sa conformation de départ : soit semie-ouverte (1HSI), soit complexée au darunavir (3EBZ).

Partie 1 : Impact de la structure de départ des simulations sur la distance d50 et la forme de la PR2

1. Visualisation des données

Chargement des données brutes.

```
rawData <- read.csv("./data/d50_3ebz_1hsi.csv", header=TRUE, sep=' ', dec='.')
str(rawData)
```

```
'data.frame':   2008 obs. of  2 variables:
 $ structure: chr  "trjconv_1hsi_284.pdb" "trjconv_1hsi_221.pdb" "trjconv_1hsi_70
5.pdb" "trjconv_1hsi_532.pdb" ...
 $ d50      : num  4.77 4.35 4.62 4.91 4.4 ...
```

On regroupe les résultats d'expérience en fonction de la structure de la protéine de départ. On a 6 lignes manquantes dans les données traitées, car ce sont des résulats étiquetés différemment des autres. On a choisi de les ignorer.

```
hsiData <- rawData[rawData$structure %like% "trjconv_1hsi_", ]
str(hsiData)
```

```
'data.frame':   1001 obs. of  2 variables:
 $ structure: chr  "trjconv_1hsi_284.pdb" "trjconv_1hsi_221.pdb" "trjconv_1hsi_70
5.pdb" "trjconv_1hsi_532.pdb" ...
 $ d50      : num  4.77 4.35 4.62 4.91 4.4 ...
```

```
ebzData <- rawData[rawData$structure %like% "trjconv_3ebz_", ]
str(ebzData)
```

```
'data.frame':   1001 obs. of  2 variables:
 $ structure: chr  "trjconv_3ebz_376.pdb" "trjconv_3ebz_499.pdb" "trjconv_3ebz_27
9.pdb" "trjconv_3ebz_252.pdb" ...
 $ d50      : num  6.41 8.33 7.97 5.92 6.1 ...
```

Distribution des d50 par conformation initiale de protéine.

```
par(mfrow=c(1, 2))
hist(hsiData$d50, main="A", xlab="d50", ylab="fréquence", xlim=c(0, 11), ylim=c(0,
500), col='orange')
axis(side=1, at=seq(0, 11, 1), labels=seq(0, 11, 1))
axis(side=2, at=seq(0, 500, 100), labels=seq(0, 500, 100))

hist(ebzData$d50, main="B", xlab="d50", ylab="fréquence", xlim=c(0, 11), ylim=c(0,
500), col='orange')
axis(side=1, at=seq(0, 11, 1), labels=seq(0, 11, 1))
axis(side=2, at=seq(0, 500, 100), labels=seq(0, 500, 100))
```

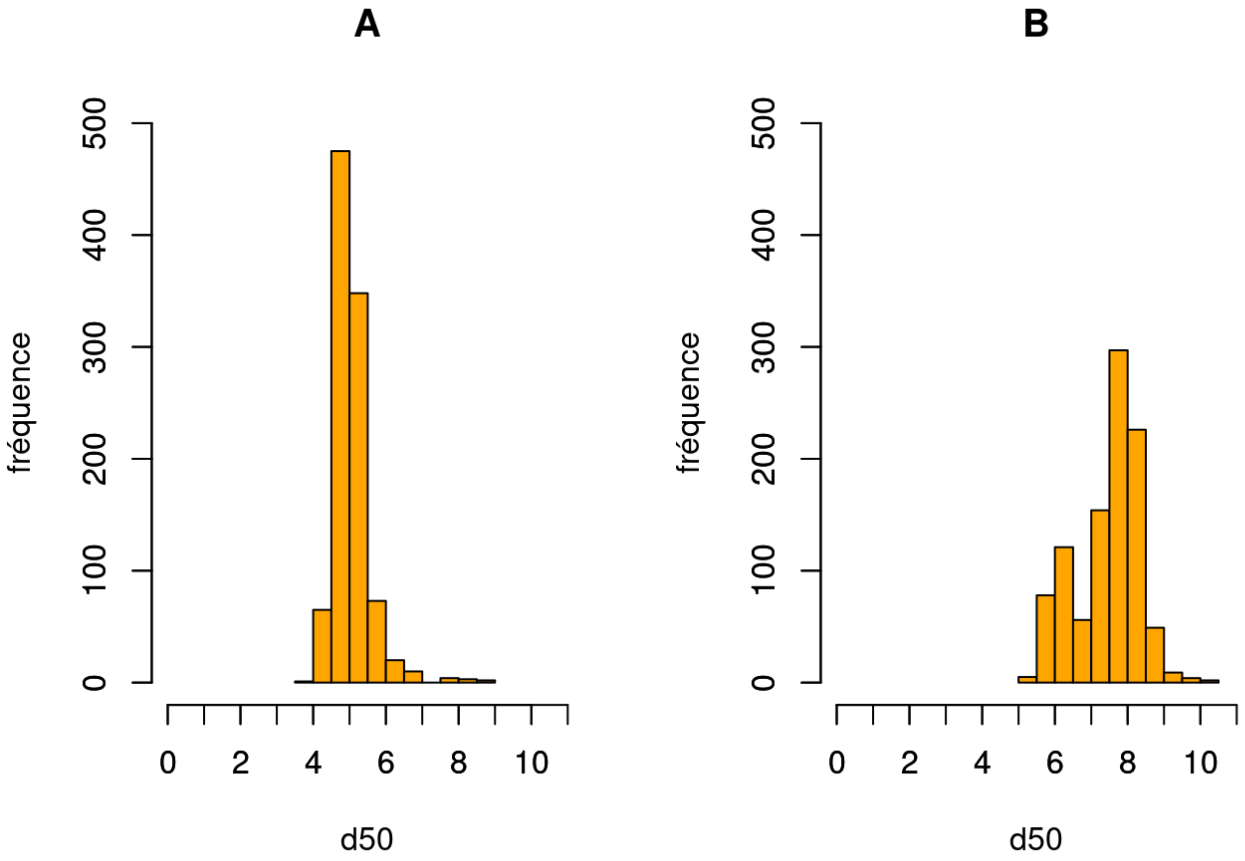
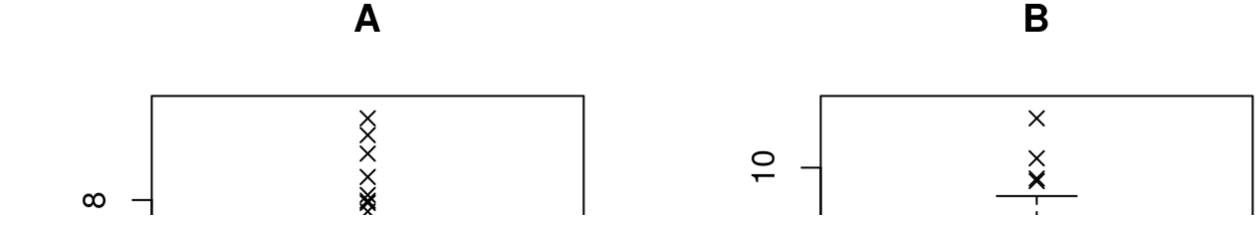
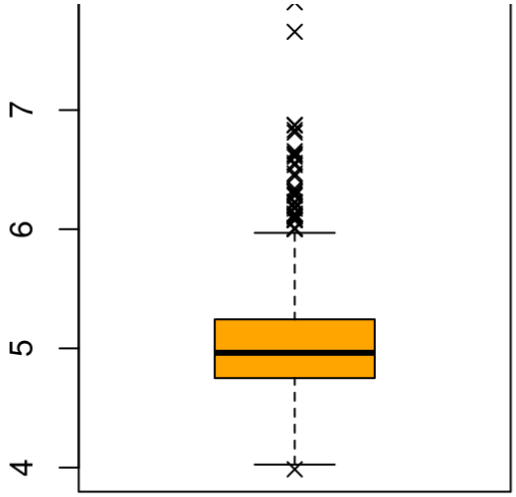


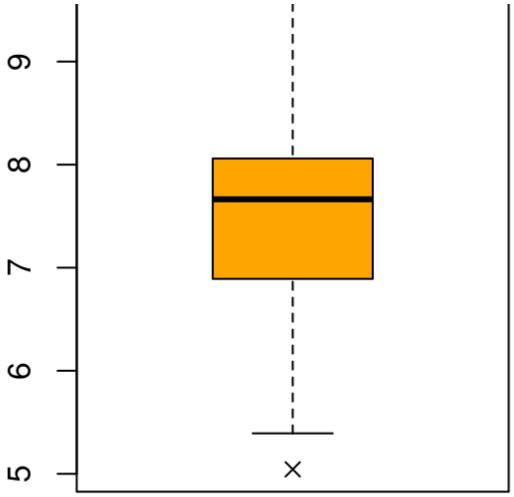
***Figure 1** : Répartition de la valeur de d50, pour les conformations initiales PDB 1HSI (A) et PDB 3EBZ (B).*

On remarque que la répartition de d50 diffère selon la structure de départ de la protéine. De plus les répartitions ne semblent pas suivre la loi normale.

```
par(mfrow = c(1, 2))
boxplot(hsiData$d50, main="A", col='orange', pch=4)
boxplot(ebzData$d50, main="B", col='orange', pch=4)
```







2. Comparaison des moyennes des distances

La distance d50 dépend-elle de la structure de départ de la protéine ?

On étudie la moyenne de d50. Cette variable aléatoire est quantitave continue de plus elle ne semble pas suivre une loi normale. Nous avons deux échantillons indépendants de taille > 30. Nous devons donc réaliser un test non paramétrique de comparaison de moyennes sur grands échantillons : c'est le test de Wilcoxon.

Échantillon 1 :

- structure de départ PDB 1HSI
- n1 = 1001
- m1 = 5.05
- v1 = 0.26

Échantillon 2 :

- structure de départ PDB 3EBZ
- n2 = 1001
- m2 = 7.45
- v2 = 0.76

Hypothèses :

H0 : la structure de départ n'influence pas la distance d50, $\mu_1 = \mu_2$

H1 : la structure de départ influence la distance d50, $\mu_1 < \mu_2$

```
wilcox.test(x=hsiData$d50, ebzData$d50)
```

```
    Wilcoxon rank sum test with continuity correction

data:  hsiData$d50 and ebzData$d50
W = 14062, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

Conclusion :

La p-value est inférieure à 2.2e-16. Pour alpha=5% on a p-value < alpha, donc le test de Wilcoxon est significatif au risque de première espèce 5%. On rejette H0, donc la moyenne de d50 de 1HSI est significativement inférieure à la moyenne de d50 de 3EBZ.

Evolution de la distance au cours du temps

Récupération des temps où les protéines ont été extraites.

```
# renvoie le temps d'extraction à partir d'une string du nom de l'expérience
protToTime <- function(protName) {
  s <- gsub(pattern="^.*_", replacement="", x=protName) # enlève le début du nom d
e la protéine
  s <- gsub(".pdb", "", s) # enlève la fin du nom
  s <- strtoi(s) # string to int
  return(s)
}

vecthsi <- mapply(FUN=protToTime, hsiData$structure) # crée le vecteur des temps d
'extraction
hsiData["time"] <- vecthsi # l'ajoute à la matrice
hsiData <- hsiData[order(hsiData[, 3], decreasing=FALSE), ] # ordonne par temps

vectebz <- mapply(FUN=protToTime, ebzData$structure)
ebzData["time"] <- vectebz
ebzData <- ebzData[order(ebzData[, 3], decreasing=FALSE), ]
```

```
par(mfrow=c(1, 2))
plot(x=hsiData$time, y=hsiData$d50, pch=4, col=rgb(0.9,0.4,0,0.6), ylim=c(4, 11),
main="A", xlab="temps d'extraction (en µs)", ylab="d50 (en Å)")
plot(x=ebzData$time, y=ebzData$d50, pch=4, col=rgb(0.9,0.4,0,0.6), ylim=c(4, 11),
main="B", xlab="temps d'extraction (en µs)", ylab="d50 (en Å)")
```

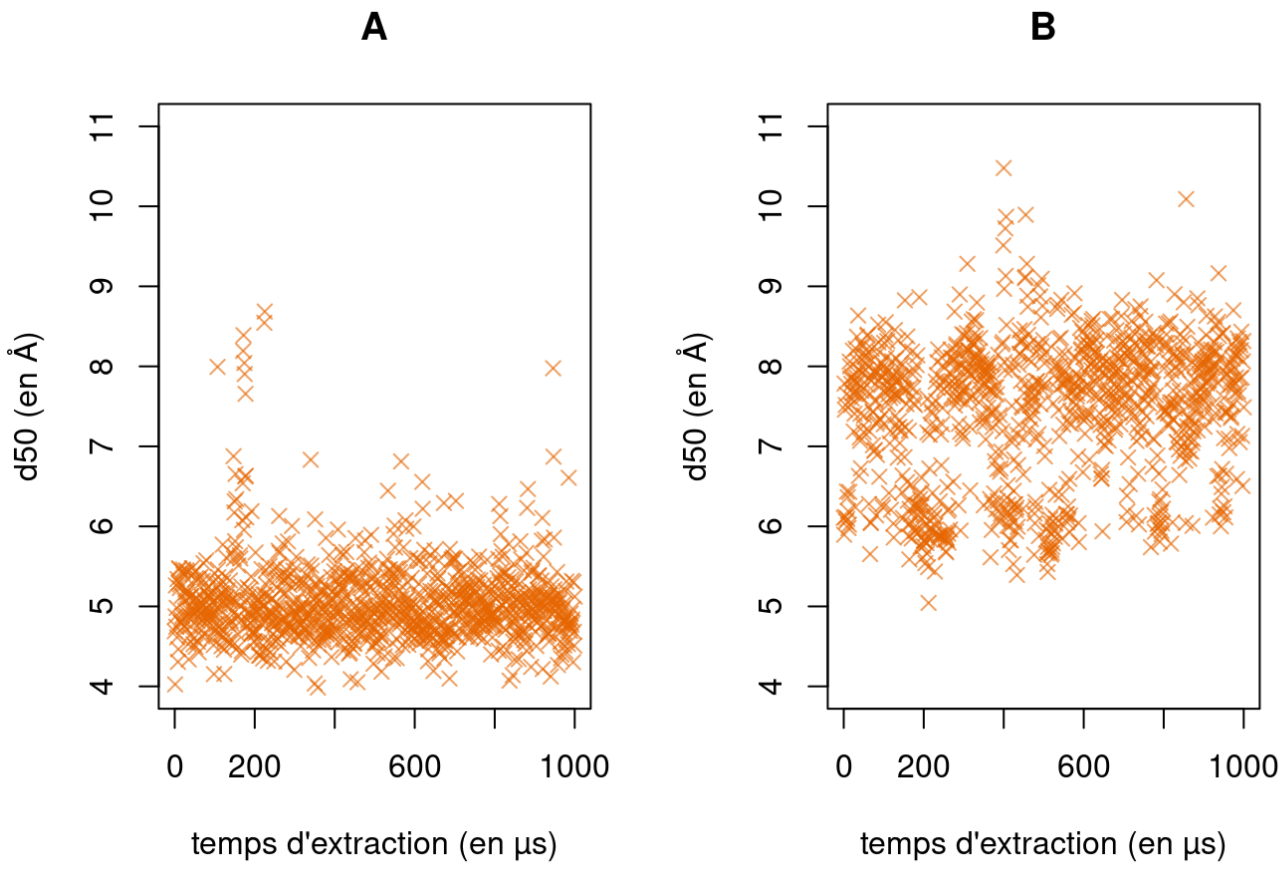


***Figure 2** : Évolution de la valeur de d50 selon le moment d'extraction de la protéine, pour les conformations initiales PDB 1HSI (A) et PDB 3EBZ (B).*

On observe pour 1HSI que la répartition de d50 semble stable au cours du temps d'extraction. Les d50 sont répartis autour de 5 Å. Par contre les valeurs de d50 pour 3EBZ semblent osciller entre 6 Å et 8 Å au cours du temps d'extraction. Cela forme une réparartition bimodale de d50.

Partie 2 : Impact de la conformation de départ sur la structure de la PR2

```
hsi2raw <- read.csv("./data/dist_SLCalpha_superlig_1HSI.csv", header=TRUE, sep='
', dec='.')
ebz2raw <- read.csv("./data/dist_SLCalpha_superlig_3EBZ.csv", header=TRUE, sep='
', dec='.')
```

4. Traitement des données de distance des carbones alpha au barycentre du ligand

On supprime les distances dSL−Cα qui ont une variance nulle. En effet si la variance est nulle cela signifie que peu importe la structure, la distance au barycentre reste la même, or on cherche à comparer les changements de distance.

```
ebz2var <- setDT(ebz2raw)[, list( structure,dSLCA,GroupVariance=var(dSLCA)) , by = CA]
hsi2var <- setDT(hsi2raw)[, list( structure,dSLCA,GroupVariance=var(dSLCA)) , by = CA]
#le data set avec les variances calculées pour chaque dSLCA par CA

#ici on crée la matrice des dSLCA pour chaque CA selon la structure
#c'est à dire on échange les lignes et colones des dtf, tout en maintenant l'ordre original des carbones asymétriques (colonne CA)
ebzDtf <- dcast(ebz2raw, structure ~ factor(CA, levels = unique(CA)), value.var = "dSLCA")
hsiDtf <- dcast(hsi2raw, structure ~ factor(CA, levels = unique(CA)), value.var = "dSLCA")

variance <- apply(ebzDtf[,2:199], MARGIN=2, FUN = var) #definitivement aucune variance egale à 0
print("Variances égales à 0 : ")
```

```
[1] "Variances égales à 0 : "
```

```
(variance[variance = 0])
```

```
named numeric(0)
```

Comme aucune variance n'est égale à 0, aucune distance ne doit être supprimée.

5. ACP des distances des carbones alpha pour les 2002 structures

```
ebzacp <- dudi.pca(scale(ebzDtf[,2:199]),scannf = FALSE , nf = 2)
hsiacp <- dudi.pca(scale(hsiDtf[,2:199]),scannf = FALSE , nf = 2)
```

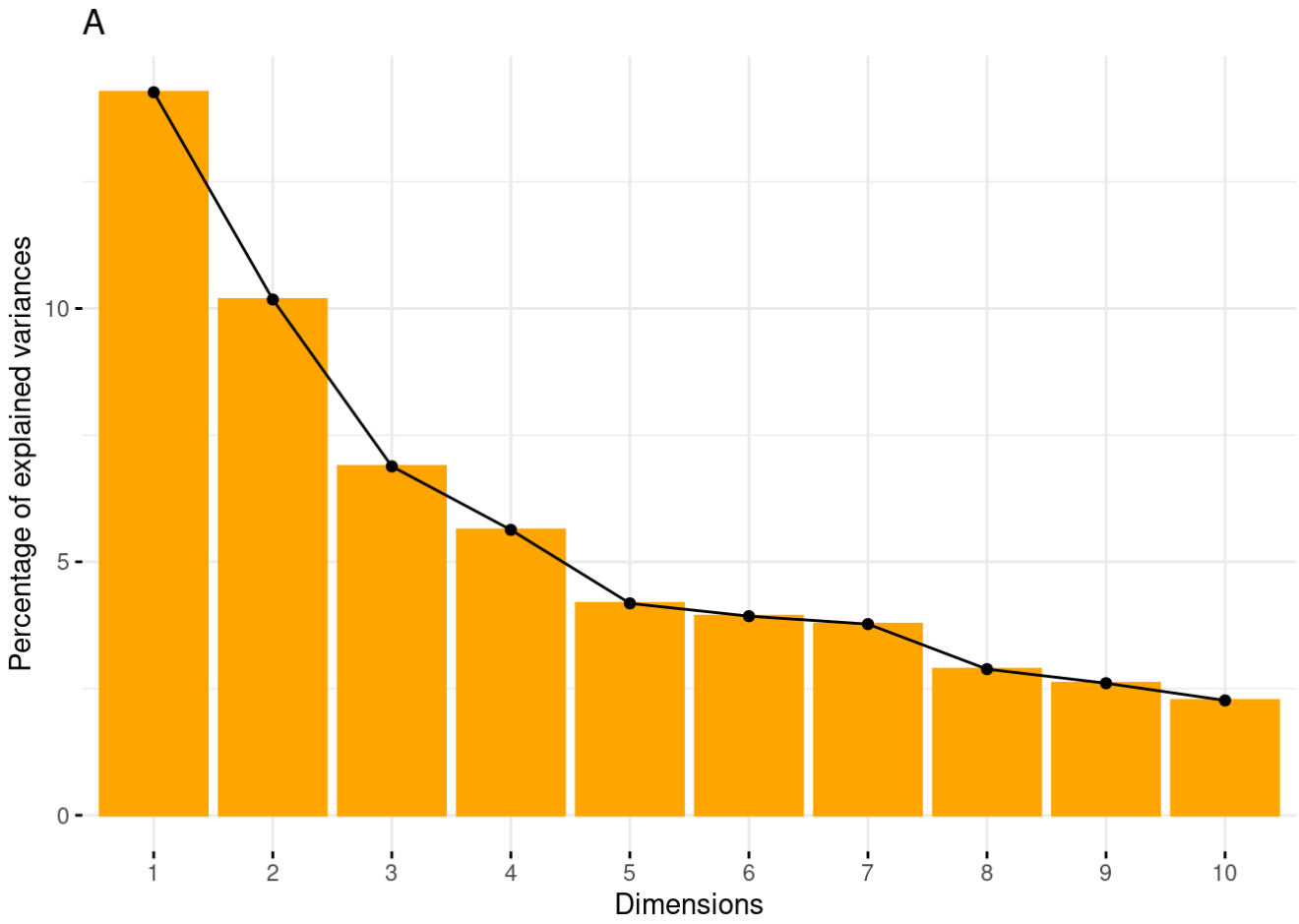


Figure 3A *: Prévalence des dimensions de l'ACP pour 1HSI (en A).*

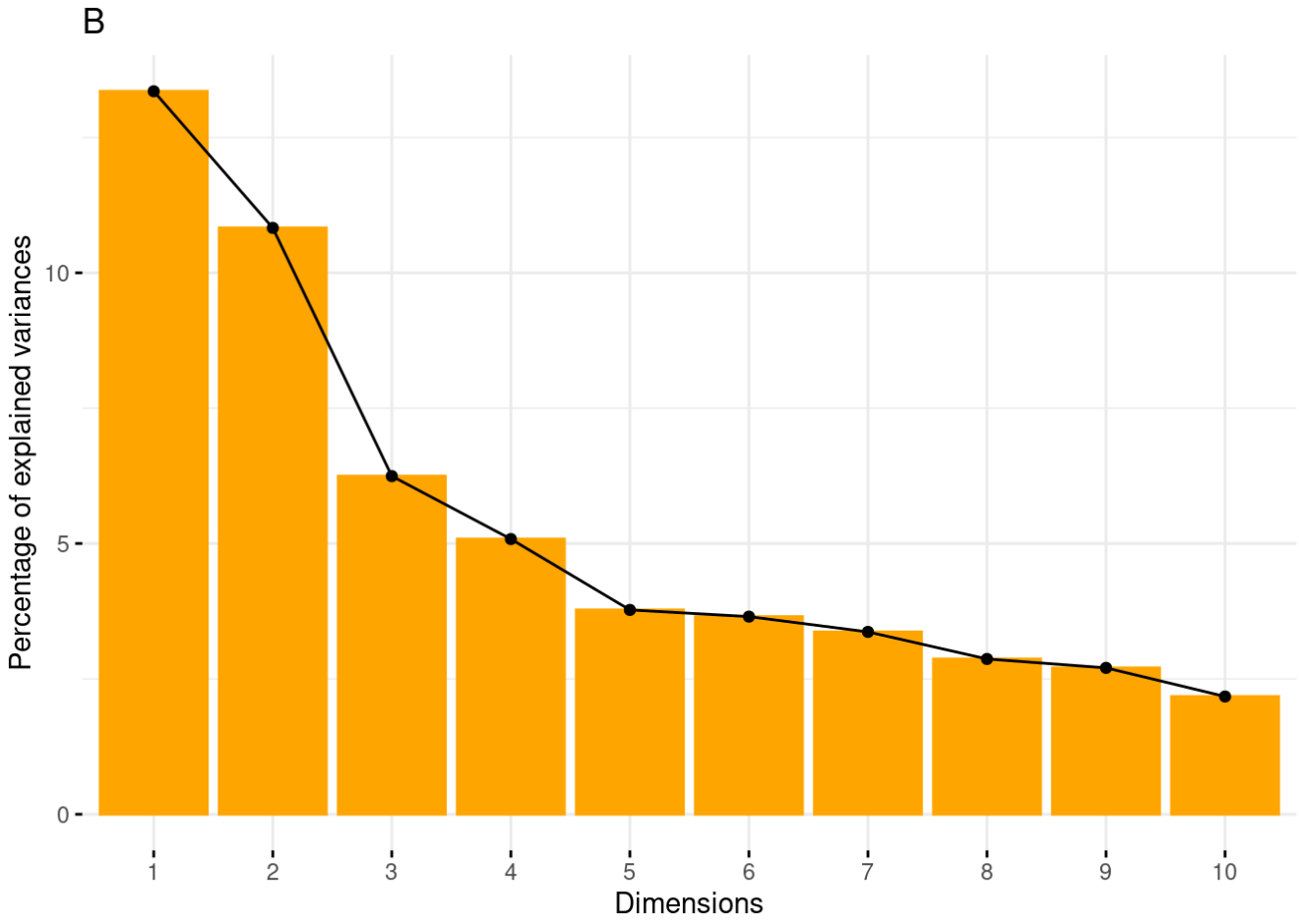


Figure 3B *: Prévalence des dimensions de l'ACP 3EBZ (en B).*

L'ACP sera peu représentatif car seulement 15% (dimension 1) et 10% (dimension 2) des variances seront expliquées pour chaque structure.

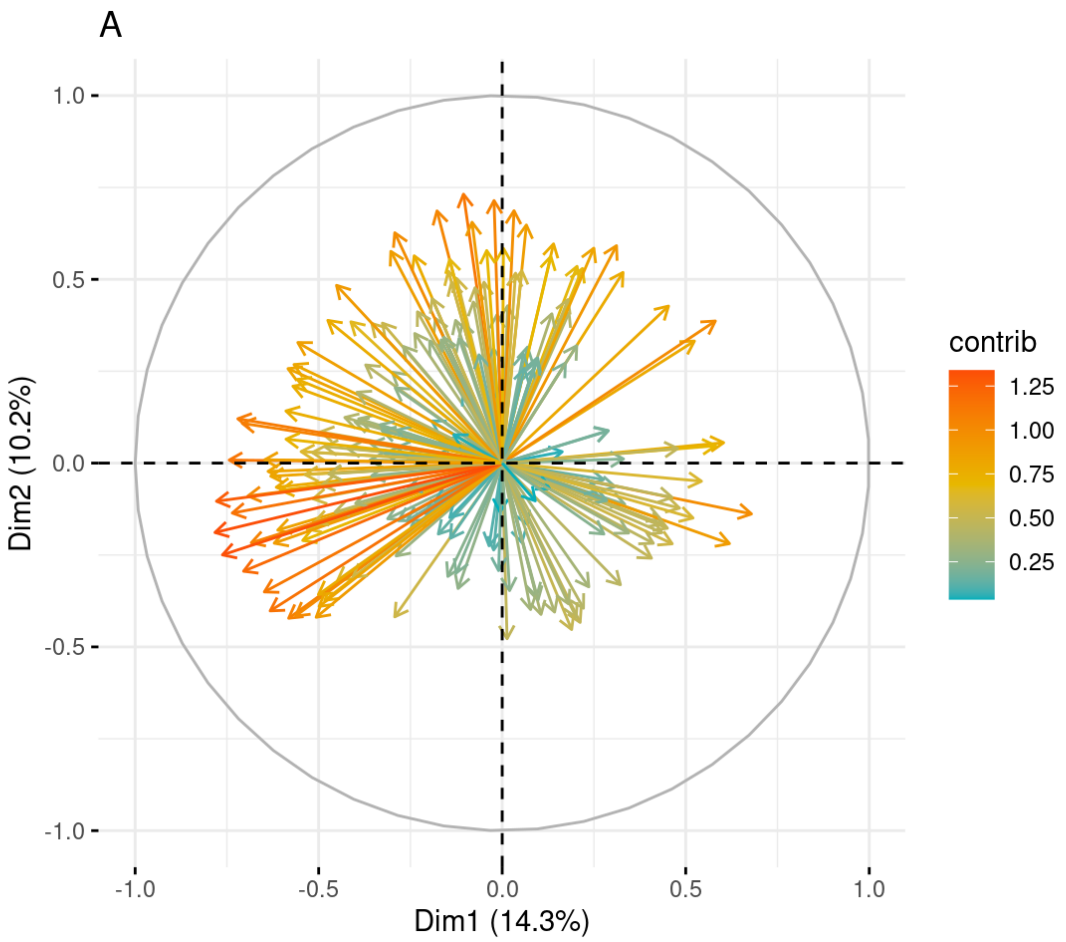


***Figure 4A** : Graphique d'ACP pour 1HSI (en A).*

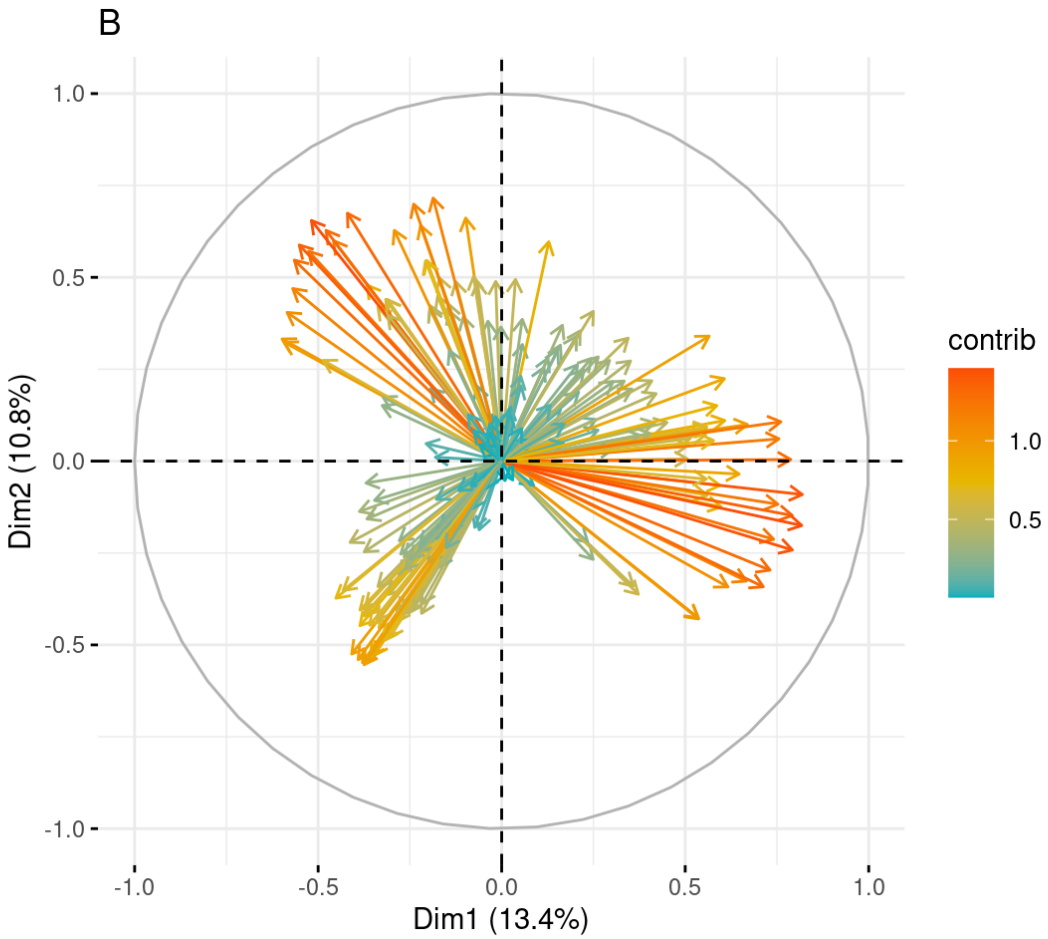


***Figure 4B** : Graphique d'ACP pour3EBZ (en B).*

On observe, pour la simulation 1HSI , une répartition homogène des distances des carbones alpha au barycentre. On remarque aussi une symétrie axiale cohérente avec la forme de la protéine. Quant à la simulation 3EBZ, on note que 3 tendances se dégagent. Tout comme la précedente, il y a aussi de la symétrie axiale.

6. Matrice de correlation

```
par(mfrow=c(1, 2))
Mhsi <- cor(hsiDtf[,2:20])
corrplot(Mhsi , title = "A")
Mebz<-cor(ebzDtf[,2:20])
corrplot(Mebz ,title = "B")
```

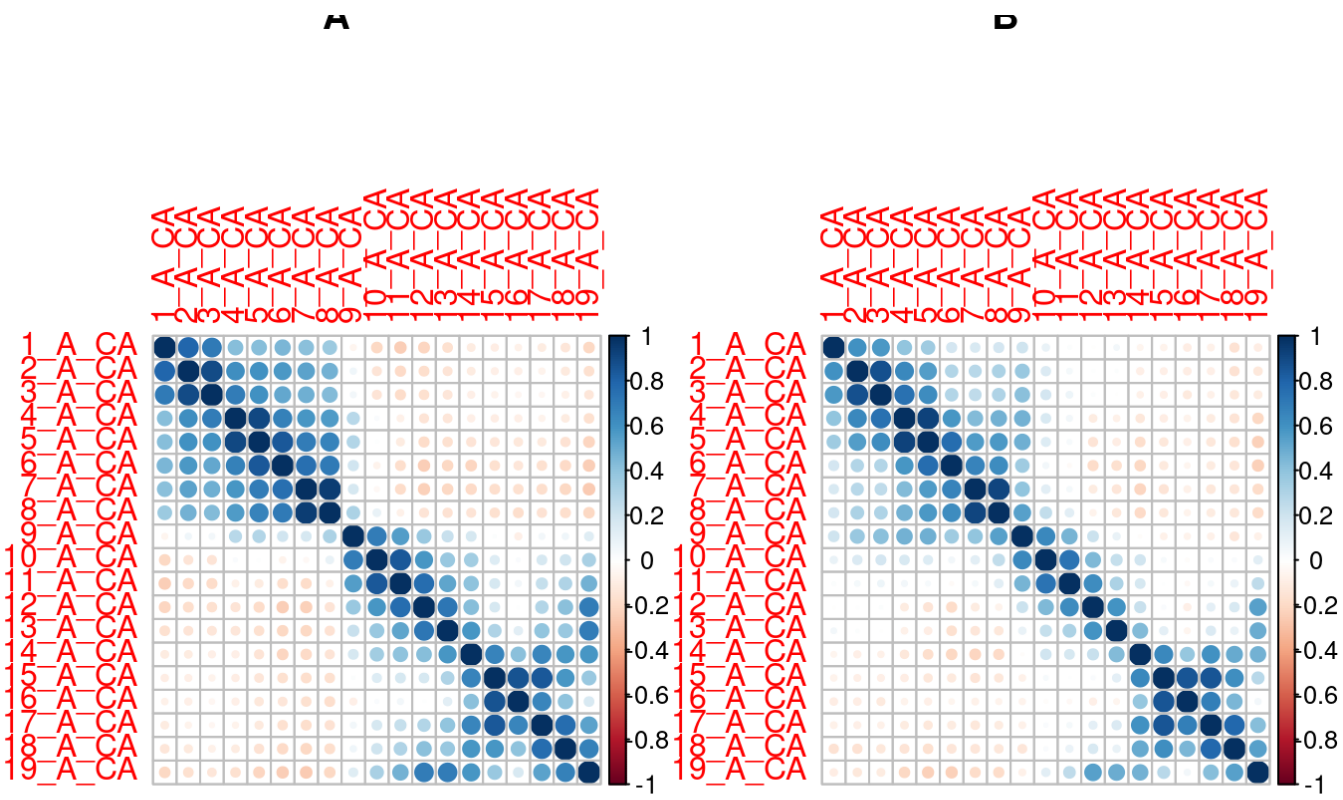


Figure 5 *: Matrice de coorélation pour 20 distances pour 1HSI (en A) et 3EBZ (en B).*

```
par(mfrow = c(1,2))
Mhsi <- cor(hsiDtf[,2:199])
corrplot(Mhsi, addgrid.col = NA , tl.pos ='n' , title = "A")
Mebz<-cor(ebzDtf[,2:199])
corrplot(Mebz, addgrid.col = NA ,tl.pos ='n' , title = "B")
```

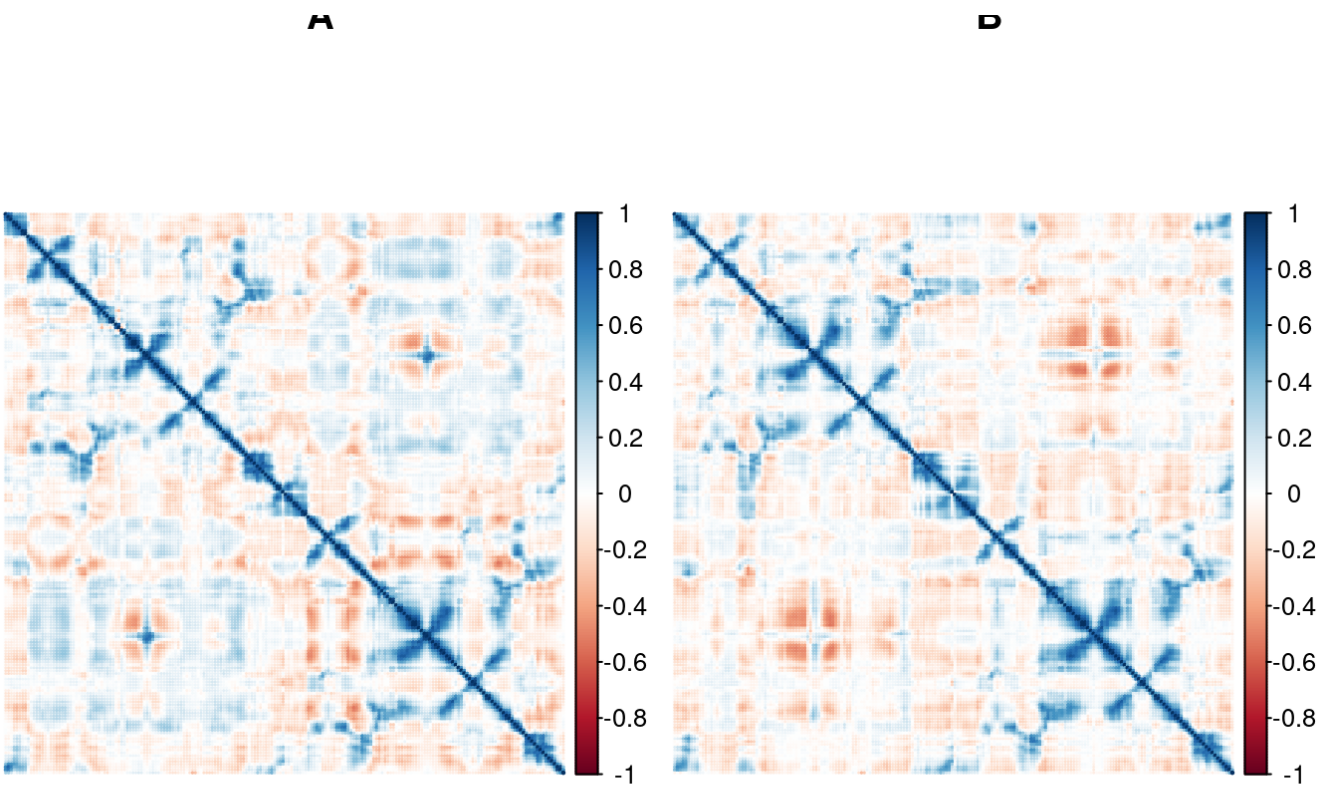


***Figure 6** : Matrice de coorélation pour 1HSI (en A) et 3EBZ (en B).*

On observe plusieurs groupes de carbones corrélés au niveau des diagonales pour les deux simulations. On peut identifier deux groupes similaires entre les simulations. Cependant, il y a généralement plus de carbones avec des distances corrélés pour la simulation 1HSI que pour la 3EBZ.

7. Suppression des distances avec des coeficient de corrélation supérieur à 0.85

Les distances avec une correlation de plus de 0.85 sont des distances qui sont liées directement entre elles. On souhaite comparer des distances similaires mais pas les mêmes, on supprime donc ces distances.

```
remhsi <- findCorrelation(Mhsi,cutoff = 0.85) #vecteur de taille 55 donc 55 distan
ces à supprimer
remebz <- findCorrelation(Mebz,cutoff = 0.85) #vecteur de taille 48 donc 48 distan
ces à supprimer

#permet de retirer le nom des carbones à retirer des dataframes
remebz2 <- colnames(Mebz[,remebz])
remhsi2 <- colnames(Mhsi[,remhsi])
```

```
#on retire de la table de coorélation les distances sélectionnées.
Mhsi <- Mhsi[,-remhsi]
Mebz <- Mebz[,-remebz]

#On supprime de nos dataframes les dslca à supprimer précedement
copyEbz <-ebzDtf %>% select(-all_of(remebz2))
copyHSI <-hsiDtf %>% select(-all_of(remhsi2))
```

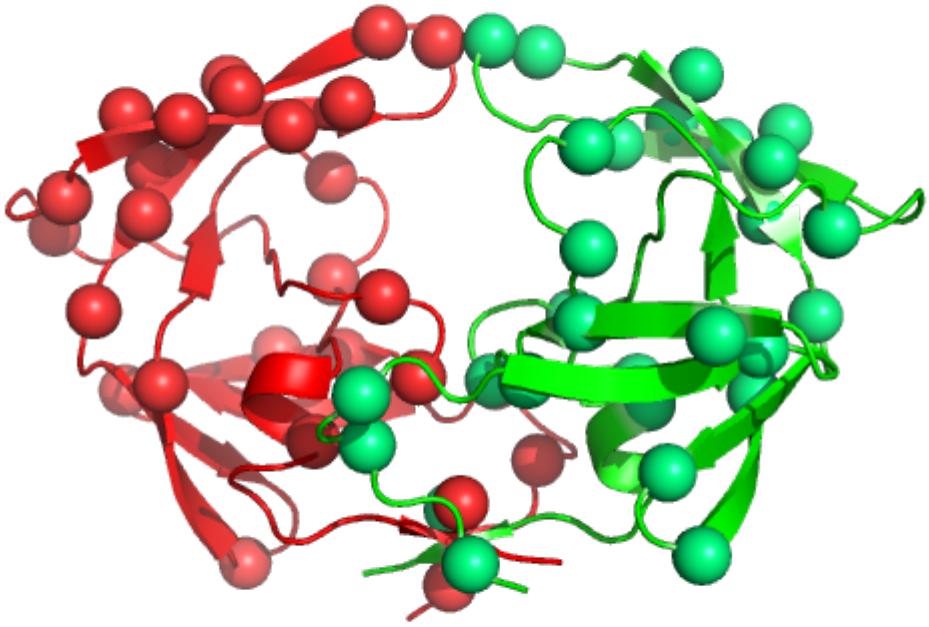
Dans le cas de la simulation 1HSI on a supprimé 55 distances.

Pour la simulation 3EBZ, 48 ont été supprimées

8. Visualisation sur PyMOL

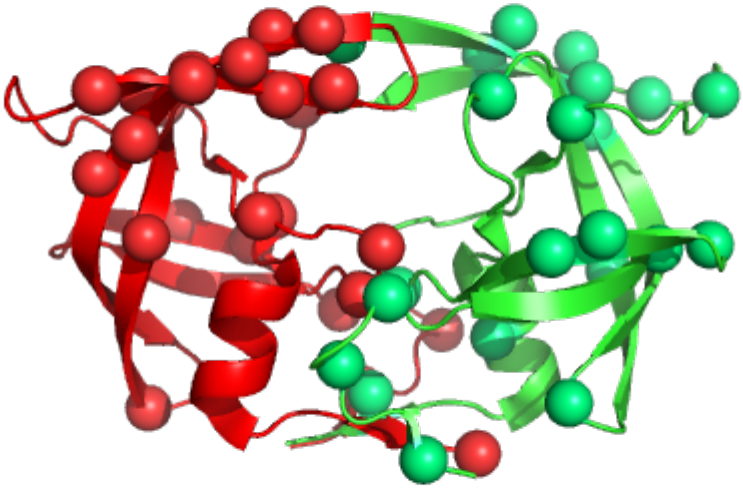
Représentation graphique des deux molécules 1HSI et 3EBZ, en particulier leurs carbones alpha sélectionnés.

1HSI



On peut voir la fente de 1HSI en haut de la molécule, comme attendu. Ces deux chaînes sont en vert et en rouge. Les carbones alphas sélectionnés sont en sphères. On voit qu'ils sont un répartis sur l'ensemble de la protéine, et qu'ils sont symétriques sur les deux chaînes.

3EBZ



La fente de 3EBZ est bien refermée dans cette représentation. On observe encore une fois une répartition des carbones alpha sur l'ensemble de la protéine, mais cette fois la symétrie par chaîne est moins évidente.

9. ACP sans les distance ayant une corrélation supérieure à 0.85

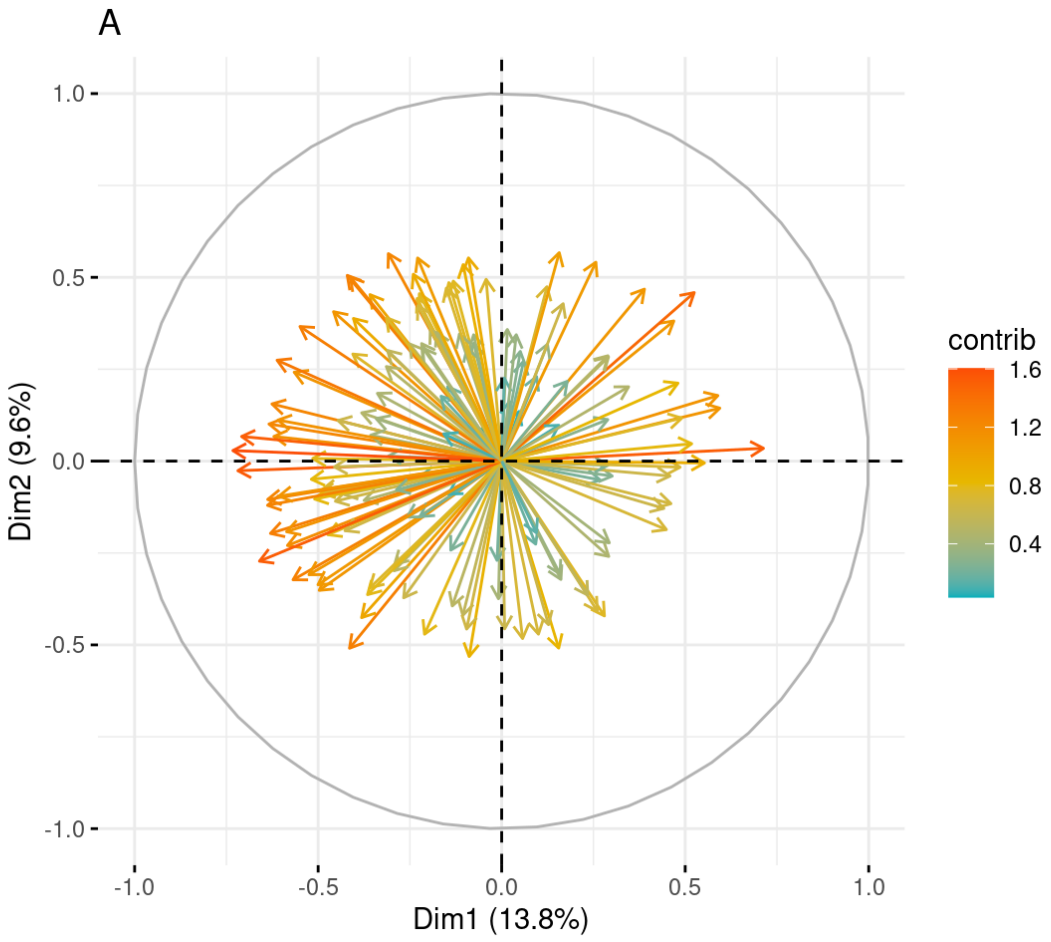


***Figure 7A** : Graphique d'ACP pour 1HSI après suppression des distances corrélées.*

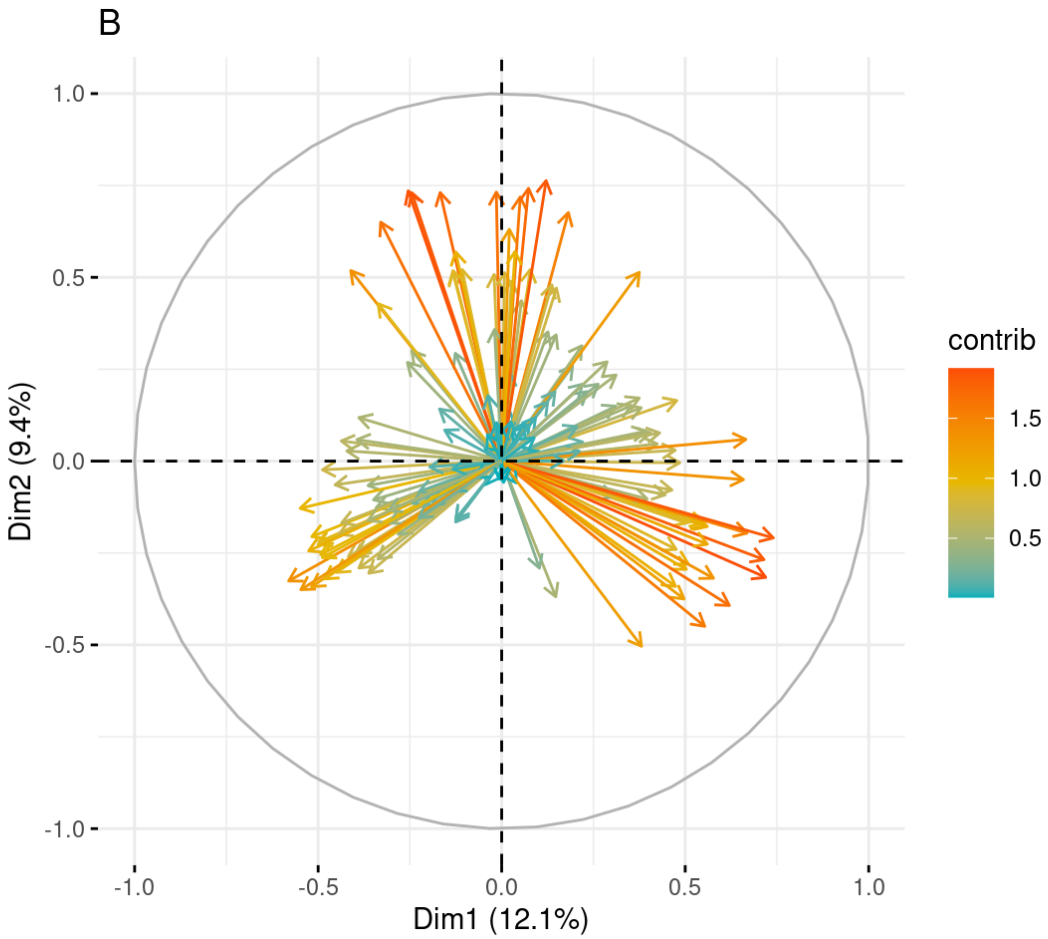


***Figure 7B** : Graphique d'ACP 3EBZ après suppression des distances corrélées.*

Les ACP donnent des résultats similaires aux précédents. Il y a quelques vecteurs en moins, grâce à l'élimination des distances très corrélées.

10. Classification hiérarchique des distances

On établie les distances entre chaque structure des deux simulations, pour faire une classification hiérarchique.

```
mhsi <- dist(copyHSI[,])
mebz <- dist(copyEbz[,])

Ahs <- hclust(mhsi)
Aeb <- hclust(mebz)
```

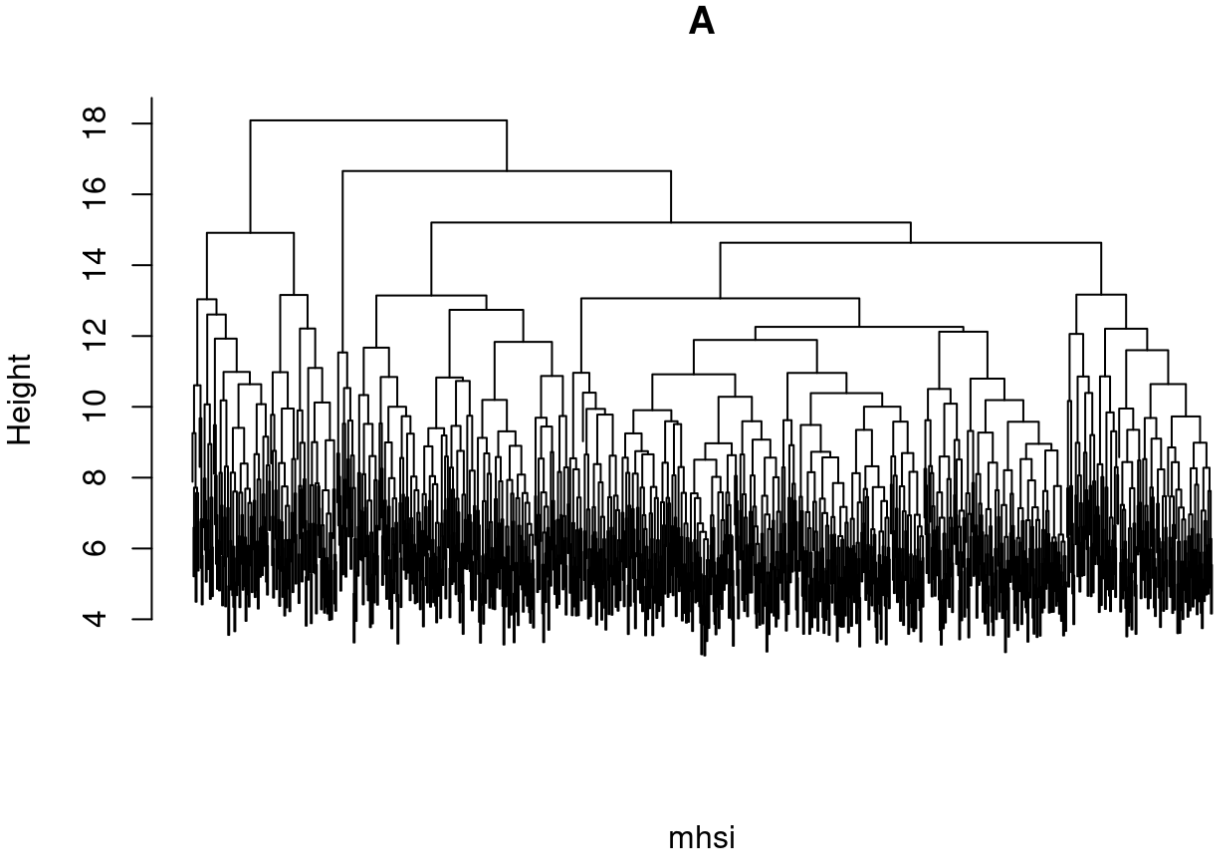


***Figure 8A** : Dendrogramme des 1001 structures de 1HSI*

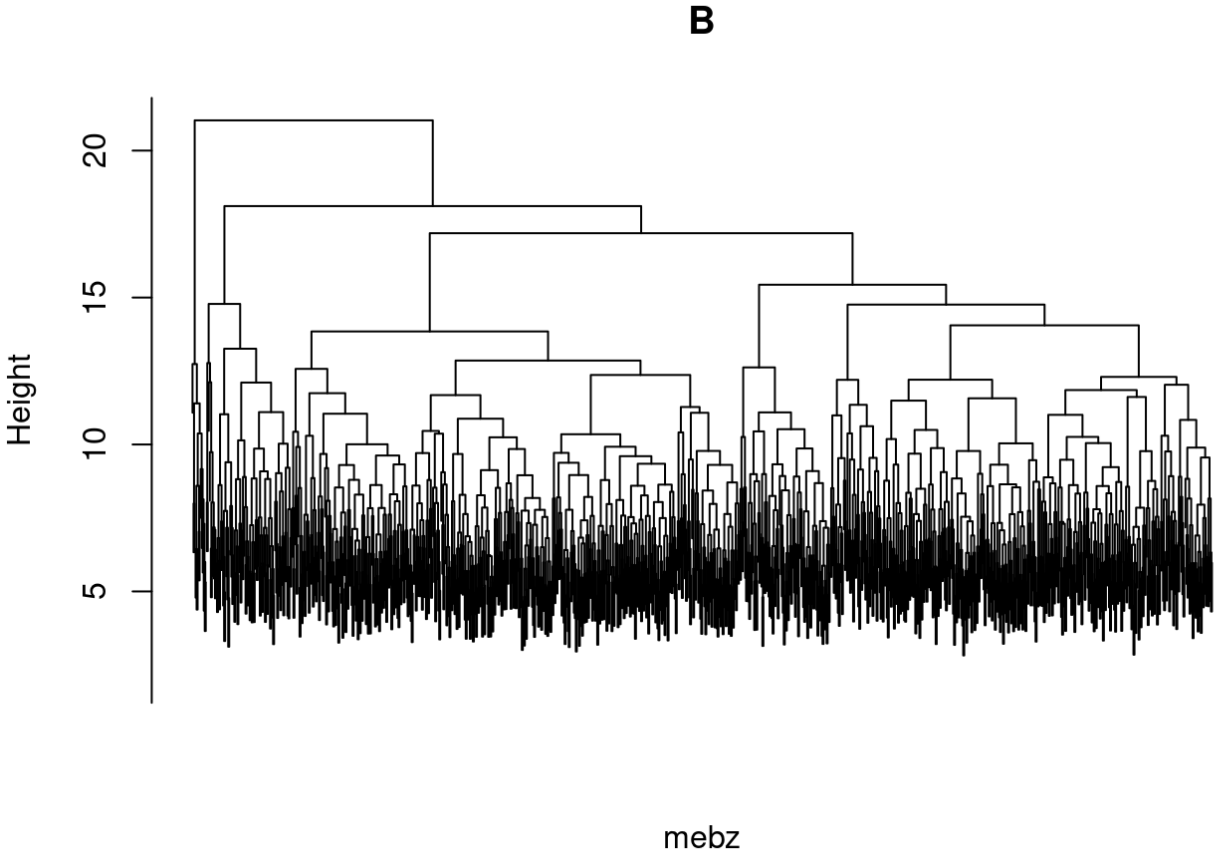


***Figure 8B** : Dendrogramme des 1001 structures de 3EBZ*

La clusterisation hiérarchique montre quelques petits clusters et un ou deux clusters principaux pour les deux arbres. Il existe donc certaines structures plus représentées que d'autres.

La grande quantités de données rend les graphes difficiles à analyser.

11. Mise en lumière de 6 clusters

```
cAhs <- cutree(Ahs, k=6)
cAeb <- cutree(Aeb, k=6)
par(mfrow=c(1, 2))
plot(cAhs, pch=4, col=rgb(0.9,0.4,0,0.3), main="A", ylab="cluster")
plot(cAeb, pch=4, col=rgb(0.9,0.4,0,0.3), main="B", ylab="cluster")
```

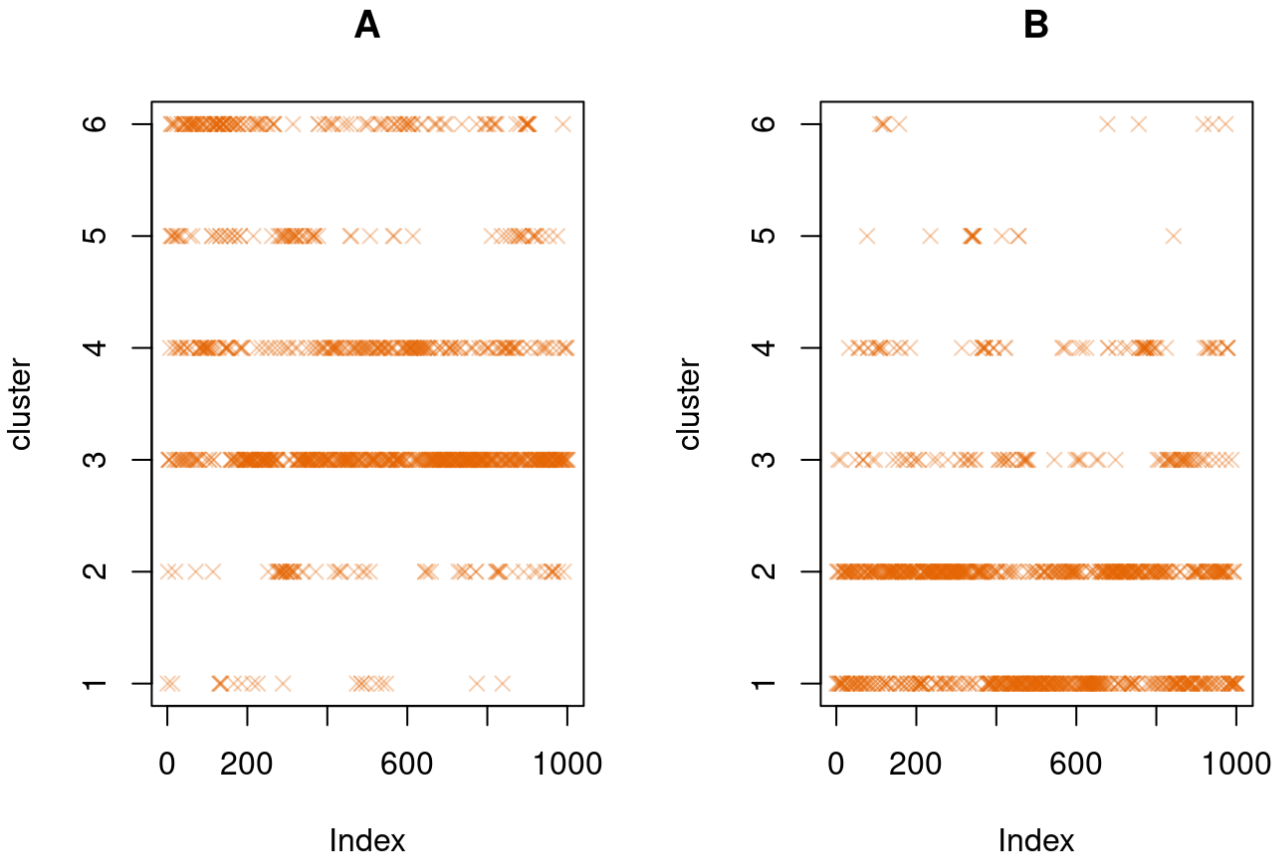


***Figure 9** : Représentation de 6 clusters hierarchiques, pour 1HSI (en A) et 3EBZ (en B)*

Pour 1HSI, on observe un des cluster est plus peuplé que les autres. Donc la majorité de structures sélectionnés est similaire. En revanche deux clusters de structures de 3EBZ semblent plus peuplés que les autres. Donc il existe deux groupes de clusters importants chez 3EBZ. Ces résultats sont cohérents avec la Figure 2, donc 1HSI présente une répartition de structures autour d'une valeur, alors que 3EBZ présente une répartition bimodale de ses structures.

On peut en déduire que la forme 3EBZ alterne entre deux conformations, expliquant la répartition bimodale de ses structures. Par contre 1HSI a une conformation qui varie autour d'une position d'équilibre.

Conclusion

La forme de départ de la protéine PR2 a une forte influence sur sa conformation durant l'expérience. Quand elle est semi-ouverte, la conformation de PR2 varie dans l'espace autour d'une position d'équilibre. En revanche, quand elle est liée à un ligand, PR2 semble osciller entre deux positions d'équilibre.