Le TP de la séance 4, Clustering

Anne Badel & Frederic Guyon

Choisir son environnement de travail

1. vous pouvez choisir de travail sur le cluster de l'IFB, ici
   - les données sont dans le répertoire :
2. ou en local, sur nos machines
   - les données sont dans le répertoire :
   - vous devez avoir la commande suivante dans votre `.bashrc` : `. /opt/sdv/anaconda/etc/profile.d/conda.sh`
   - puis vous devez lancer l'environnement `conda` adéquat : `conda activate`
   - et enfin lancer `rstudio` : `rstudio`

Lire les données

Les données sont issues de la base Recount2 (https://jhubiostatistics.shinyapps.io/recount/). Nous nous intéressons à la base TCGA, regroupant des données RNA-seq pour différents types de cancer. Nous nous intéressons ici uniquement aux données BIC concernant le cancer du sein.

Les données ont été préparées pour vous, filtrage des données, normalisation, extraction des gènes différentiellement exprimés (à l'aide d'une analyse de la variance (anova) et du calcul des *p.value* ajustées par une méthode de FDR (fonction `aov` et `p.adjust(, method="fdr")`).

1. A l'IFB : sur le serveur de l'IFB, les données sont dans le répertoire `../../projects/du_bii_2019/data/module3/seanc`
   Vous pouvez donc les lire à l'aide de la commande : `mes.expr <- read.table("../../projects/du_bii_2019/data/mo`
   `h=T)`
2. En local : en local, les données sont dans le répertoire `../../dubii/data-m3/s4`. Vous pouvez donc les lire à l'aide de la commande : `mes.expr <- read.table("../../dubii/data-m3/s4/BIC_diff_exp.tsv",`
   `h=T)`

- prenez le temps d'identifier
- la taille du jeu de données
- les individus
- les variables

rq : Classiquement, en analyse de données, les individus sont les lignes du tableau de données, les colonnes sont les variables. Pour des raisons histroiques, ce n'est pas le cas en analyse transcriptomique, où :

- 1 ligne = 1 gène
- 1 échantillon biologique = 1 colonne

Ce qui implique de faire attention, et éventuellement de travailler sur la matrice transposée (fonction `t` en R) pour utiliser correctement les fonctions classiques.

Calcul de la matrice de distance

Nous allons utiliser comme métrique le coefficient de corrélation de Spearman, plus adapté à ce type de données que la distance euclidienne utilisée en cours.

1. calcul de la matrice de corrélation de Spearman à l'aide de la fonction `cor` avec l'option `method="spearman"`
2. transformation du corrélation de Spearman en une distance à l'aide de la trnsformation : $d = 1 - r^2$

hclust

1. faire un premier clustering hiérarchique, avec le critère d'aggrégation par défaut
2. faire un deuxième clustering hiérarchique, avec le critère d'aggrégation de Ward
3. Comparer les classifications obtenues
    - en particulier sur les nombres de cluster
    - utiliser les commandes `rect.hclust` et `cutree` pour visualiser les clusters sur le dendrogramme, puis récupérer les clusters.

kmeans

1. faire un premier kmeans, par exemple, en prenant le nombre de groupe trouvé sur le `hclust`
2. faire une boucle pour trouver le nombre optimal de cluster, en calculant l'inertie intra totale en fonction du nombre de groupe `kmeans()$totss` [faire une boucle pour i allant de 1 à 10 `for (i in 1:10) {}`]
3. refaire le kmeans avec ce nombre optimal
4. visualiser ces groupes par exemple sur une projection des données dans le plan par PCA, à l'aide de la fonction 'plot(PCA(mon.data.frame, choix="ind", col.ind=mon.kmeans$cluster)).

Comparaisons

kmeans versus hclust

Nous allons maintenant comparer les résultats de ces deux méthodes de clustering.

1. à l'aide de la fonction `table`, calculer la matrice de confusion de vos deux clustering. Commentez.
2. à l'aide de la fonction `adjustedRand(clues)` calculez le RI et le ARI de vos clustering. Commentez.

clustering versus Her2 status

Nous connaissons les types de cancer des différentes tumeurs, définie en combinant trois marqueurs immunologiques :

- HER2,
- ER (récepteur d'œstrogène)
- Pr (récepteur de progestérone)

et nous obtenons les classes suivantes :

- Basal.like
- HER2pos
- Luminal.A
- Luminal.B

qqs tumeurs sont non classées

Vous pouvez lire les données concernant le type de cancer grâce à la fonction `read.table`, la ligne de commande est : `mes.classes <- read.table("../../xxxx/BIC_sample-classes.tsv", h=T)`. A l'aide de la fonction `summary`, déterminez le nombre de tumeurs pour chaque type de cancer

1. comparez vos résultats de clustering avec la réalité
    - par des visualisations
    - le calcul de la matrice de confusion
    - le calcul des rand index et adjusted rand index
2. commentez