Projet : étude du cancer du sein ER+

M2 MMA - Université de Paris

2020-2021

Instructions pour le rendu du projet

Le projet est à rendre pour le Vendredi 18 Décembre 23h59 au plus tard via Moodle, sous le format .pdf, qui pourra avoir été généré à l'aide du logiciel suivant (au choix) :

- Word,
- LaTeX(mieux), qui permet d'écrire des mathématiques proprement,
- `Rmarkdown` (encore mieux), qui permet de faire des rendus automatisés de code `R` et d'inclure du LaTeX.

Les codes `R`* devront également être envoyés (ou visibles si vous travaillez sur `Rmarkdown`). Ce projet est **personnel** et devra donc être rédigé comme tel. Les codes `R` seuls ne suffiront pas, ils doivent être accompagnés de commentaires, graphiques, bibliographie et analyses personnelles. Les questions proposées ne sont là que pour vous guider.

* *Un guide d'utilisation de* `R` *est accessible via Moodle. Vous trouverez aussi de nombreux éléments sur le site suivant* http://www.sthda.com/.

Introduction

Les informations suivantes sont extraites du site de l'Institut National du Cancer.

Un cancer du sein résulte d'un dérèglement de certaines cellules qui se multiplient et forment le plus souvent une masse appelée tumeur. Il en existe différents types qui n'évoluent pas de la même manière. Certains sont "agressifs" et évoluent très rapidement, d'autres plus lentement. Les cellules cancéreuses peuvent rester dans le sein. Elles peuvent aussi se propager dans d'autres organes, ce qui est une situation encore plus menaçante. On parle alors de métastases. Dans la majorité des cas, le développement d'un cancer du sein prend plusieurs mois, voire plusieurs années. Le cancer du sein est le cancer le plus fréquent chez la femme. Il représente plus du tiers de l'ensemble des nouveaux cas de cancer chez la femme.

Certaines tumeurs du sein ont pour caractéristique d'être hormonosensibles, ce qui signifie que les hormones féminines (œstrogènes, progestérone), naturellement produites par l'organisme, stimulent leur croissance. Les cellules cancéreuses hormonosensibles possèdent en fait des récepteurs hormonaux, qui sont des protéines situées à la surface de la cellule cancéreuse. Ils détectent les œstrogènes ou la progestérone qui passent dans le sang et les captent. La liaison entre les hormones et leurs récepteurs sur les cellules déclenche la stimulation de la croissance de ces cellules cancéreuses.

On distingue deux types de cancer du sein positif de récepteur hormonal, selon le type d'hormone avec lequel le récepteur est associé :

- Cancer du sein récepteur-positif d'œstrogène (ER+) dans le cas de récepteurs d'œstrogène,
- Cancer du sein récepteur-positif de progestérone (PR+) dans le cas de récepteurs de progestérone.

Le but de ce projet est d'identifier les gènes responsables du développement d'un cancer du sein ER+.

I. Chargement et nettoyage des données

Les données utilisées dans ce projet ont été téléchargées depuis le site http://gdac.broadinstitute.org/, qui regroupe des données concernant différents types de cancer, collectées dans le cadre du projet américain TCGA. Ce projet à grande échelle TCGA (*The Cancer Genome Atlas* en anglais) a été lancé en 2005 pour cataloguer les mutations génétiques responsables du cancer en utilisant le séquençage génomique et la bio-informatique, l'objectif étant d'appliquer des techniques d'analyse génomique à haut débit pour améliorer la capacité à diagnostiquer, traiter et prévenir le cancer grâce à une meilleure compréhension de la base génétique du cancer.*

* *extrait de Wikipédia*

Une version pré-traitée des données est disponible sur Moodle. Vous pouvez les charger à l'aide des commandes suivantes :

```
load("ProcessedDataBRCA.Rdata") # données d'expression des gènes
# gènes en colonnes, patients en lignes
# le jeu de données est accessible via MA_TCGA
load("clinicalData.Rdata") # données cliniques associées
```

Dans cette première partie, on se propose de nettoyer les données avant de les utiliser. Pour cela, vous procéderez notamment de la manière suivante :

1. Centrer les données. *On ne réduira pas les données afin de garder l'effet sur/sous-expression des gènes*
2. Supprimer du jeu de données d'expression des gènes les gènes dont la variance est la plus petite. *Le seuil est à définir, on pourra par exemple ne garder que* 75% *des gènes les plus variants.*

II. Tests multiples pour réduire la dimension du jeu de données

Dans cette 2ième partie, on se propose de réduire la dimension du jeu de données d'expression des gènes pour ne garder que les gènes qui sont différentiellement exprimés entre les conditions "cancer du sein ER+" et "cancer du sein non ER+".

La variable clinique indiquant si la patiente est "ER+" ou non est accessible via la commande suivante :

```
ER <- clinicalData$patient.breast_carcinoma_estrogen_receptor_status
```

Pour cette étude, vous pouvez vous appuyer sur les questions suivantes :

1. Enlever, s'il y a lieu, les valeurs manquantes puis transformer la variable `ER` en variable binaire (2 modalités uniquement).
2. Effectuer une analyse différentielle entre les deux conditions "cancer du sein ER+" et "cancer du sein non ER+" à l'aide d'un test de Student. *La correction pour tests multiples appliquée ainsi que le niveau de risque associé est à expliciter.*

III. ACP pour différencier les deux groupes

Dans cette 3ième partie, on se propose de mener une analyse plus détaillée des gènes qui permettent de différencier la caractéristique ER+ à l'aide d'une ACP*. Pour cela, vous pouvez vous appuyer sur les questions suivantes :

1. A l'aide du jeu de données réduit obtenu dans la partie II, effectuer une ACP.
2. Tracer le diagramme représentant la contribution des individus puis afficher l'appartenance des individus au groupe ER+.
3. Identifier les gènes qui contribuent le plus aux composantes principales construites.

* *Ici, l'ACP est utilisée dans un cadre exploratoire pour discriminer deux groupes et non dans un cadre de réduction de dimension.*

IV. Apprentissage par méthodes régularisées pour valider les résultats

Dans la dernière partie, on se propose de mettre en place un modèle de régression linéaire pour étudier l'effet de l'expression des gènes sur la caractéristique ER+ et comparer les résultats obtenus avec ceux de la partie III. Pour cela, vous pouvez vous appuyer sur les questions suivantes :

1. Mettre en place un modèle de régression logistique pénalisée pour expliquer l'appartenance au groupe ER+ en fonction des données d'expression des gènes. *On utilisera le jeu de données réduit obtenu dans la partie II.*
2. Identifier les gènes responsables de la caractéristique ER+ et les comparer avec ceux obtenus dans la partie III.
3. Dans le jeu de données TCGA initial, le statut ER+ de certaines patientes n'a pas pu être déterminé. Ces données sont accessibles à l'aide de la commande suivante :

```
load("NewData_BRCA.Rdata")
# gènes en colonnes, patients en ligne
# les données sont accessibles via MA_TCGA_test
```

A l'aide du modèle logistique mis en place, prédire le statut de ces patients à partir des données d'expression des gènes.