ENSAI
École nationale de la statistique et de l'analyse de l'information

Éline Chartier Granouilhac
Eulalie Delaune
Ignacio Espinosa
Léo Gruchociak

CANCER DE LA VESSIE & PROFIL GÉNÉTIQUE

tuteur :
Stéphane Chrétien

Université Lumière Lyon 2

Contexte et objectifs

Le cancer de la vessie figure parmi les plus fréquents dans le monde. Une prise en charge précoce et adaptée garantit un taux de survie relativement élevé par rapport à d'autres types de cancers. Pouvoir évaluer la progression du cancer et assigner les traitements adéquats constitue ainsi un enjeu majeur de santé publique .
Nos travaux visent à évaluer l'évolution d'une tumeur de la vessie cancéreuse à partir du profil génétique.

Notre étude porte sur **100 individus** atteints d'un cancer de la vessie. Pour chaque individu, nous disposons de l'expression de **34 gènes** ainsi que d'informations sur l'avancée du cancer.

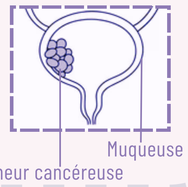
Nos objectifs sont les suivants :
- déterminer quels gènes sont significatifs dans la progression du cancer
- construire un modèle de prédiction de l'avancée du cancer en fonction de l'expression génétique

1. Comment analyser la progression du cancer de la vessie ?

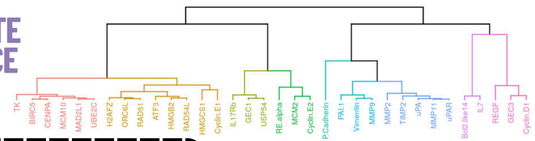
Le cancer de la vessie est caractérisé par son stade, son grade, l'examen anatomopathologique du patient et l'état de la tumeur (invasive ou superficielle).

La variable résumant le mieux ces informations est l'**anatomopathologie** :
- niveau 1 : tumeur superficielle et confinée à la muqueuse
- niveau 2 : tumeur superficielle et plus ancrée
- niveau 3 : tumeur invasive et cancer plus avancé

Exemple de cancer de la vessie de niveau anatomopathologique 3 :

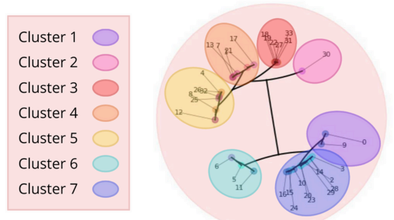


2. CLASSIFICATION ASCENDANTE HIÉRARCHIQUE DANS UN ESPACE EUCLYDIEN

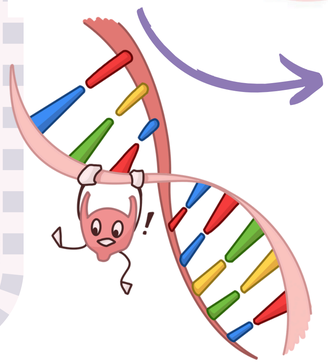


Face au fléau de la dimension, nous effectuons un regroupement de variables.
Objectif : rassembler en clusters les gènes aux trajectoires similaires

3. CLASSIFICATION ASCENDANTE HIÉRARCHIQUE DANS UN ESPACE HYPERBOLIQUE



La projection des données dans un espace hyperbolique permet d'encoder naturellement les relations hiérarchiques, et ainsi d'obtenir une meilleure classification. Nous utilisons la classification précédente comme base.
Résultats : 7 clusters de gènes

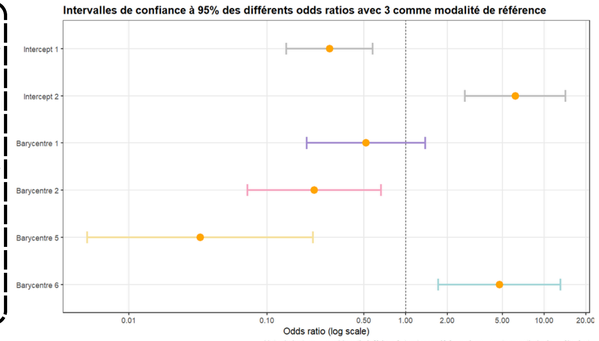


4. SÉLECTION DE VARIABLES PAR RÉGRESSION LASSO, KNOCKOFF FILTERS ET SUSIE

Nous mettons en place trois méthodes complémentaires pour déterminer quels clusters de gènes sont significatifs dans la progression du cancer de la vessie. Nous regroupons alors les gènes d'un même cluster en calculant le barycentre associé. Dans notre cas, elles s'accordent à sélectionner les mêmes barycentres.
Résultats : 4 barycentres significatifs avec un taux de mauvaise découverte de 0,25

5. CONSTRUCTION D'UN MODÈLE DE RÉGRESSION

Nous construisons un modèle de régression ordinale à odds proportionnels sur les 4 barycentres significatifs. Cela permet ensuite de déterminer la probabilité qu'une tumeur soit de niveau anatomopathologique de 1, 2 et 3 selon l'expression génétique. Une première approche consiste alors à renvoyer la modalité avec la plus grande probabilité associée.
Résultats : Le modèle semble concorder avec les études actuelles en cancérologie, en soulignant l'impact positif de gènes tels que Vimentin (Barycentre 5) ou TK (Barycentre 2) sur le cancer de la vessie.



6. Méthodes de prédiction conforme

Le modèle de régression que nous avons construit permet de prédire le niveau anatomopathologique, mais n'apporte aucune garantie en terme de niveau de confiance. Nous souhaitons alors fixer le taux d'erreur α des prédictions réalisées. Pour cela, nous utilisons des méthodes de **prédiction conforme**. Ces méthodes reposent sur l'hypothèse d'échangeabilité des données et utilisent une fonction de score. La prédiction conforme garantit alors que $1-\alpha$ des futures observations seront couvertes par les ensembles prédits.

Dans notre étude, nous utilisons en particulier deux méthodes de prédiction conforme :
1. *Le Split Conformal :* Construction du modèle sur l'ensemble d'entraînement, calcul des scores et du quantile sur l'ensemble de calibration; puis prédiction sur de nouveaux individus.
2. *Le Full Conformal :* Pour chaque nouvel individu et modalité de la variable d'intérêt, construction d'un modèle sur l'ensemble d'entraînement auquel on rajoute ce nouveau couple; calcul des scores et du quantile; prédiction pour ce nouvel individu .

Full Conformal

Nouveau point :

Construction de 3 modèles :

Chaque modèle est entraîné sur les données avec en plus le nouveau point associé à un label.
- Modèle 1 : entraîné avec le label 1
- Modèle 2 : entraîné avec le label 2
- Modèle 3 : entraîné avec le label 3

Calcul des scores :

Pour chacun des points des trois modèles, on calcule le score à partir de la fonction suivante :

$$S(Y_i, X_i) = 1 - \hat{P}(Y_i = y_i \mid X_i = x), \quad \forall i \in \{1, 2, ..., n\}$$

Pour le nouveau point en particulier (avec k la modalité correspondant au modèle) : $s_{n+1,k} = S(k, X_{n+1}) = 1 - \hat{P}(Y_i = k \mid X_{n+1} = x_{n+1})$

Détermination du quantile empirique et de l'ensemble de prédiction final :

On détermine pour chaque modèle le quantile empirique des scores pour renvoyer l'ensemble de prédiction suivant :

$$\hat{C}(X_{n+1}) = \{\text{Tous les labels } k \in \{1, 2, 3\} : S(k, X_{n+1}) \leq \hat{q}_{1-\alpha,k}\}$$

Cet ensemble est associé à un niveau de confiance $1-\alpha$

Résultats et conclusion

Au cours de notre étude, nous avons identifié **24 gènes** caractéristiques de la progression du cancer de la vessie et **10 gènes** qui ne semblent pas y être liés. Ces gènes sont regroupés en clusters, à partir desquels nous avons construit un modèle de prédiction de l'anatomie pathologique. Il s'avère être relativement fiable quand il est associé à la prédiction *full conformal*. Nos résultats corroborent les recherches existantes en cancérologie. Nous espérons que ce modèle ou les résultats qui y ont conduits pourront être utilisés dans l'analyse de la progression du cancer de la vessie.

Niveau de confiance	Taux de couverture	Taille moyenne des ensembles	Temps de calcul
0.1	90%	1.65	23 secondes
0.05	95%	1.86	19.5 secondes

Bien que coûteux d'un point de vue computationnel, le *full conformal* offre une prédiction précise à partir d'un petit échantillon de données.

Une extension possible de nos travaux consisterait à intégrer une procédure de Benjamini-Hochberg, afin d'identifier les profils d'individus pour lesquels le modèle tend à se tromper.