Résistance aux antibiotiques

Prédiction par apprentissage statistique

Vadim BERTRAND, Cheikh-Darou BEYE

9 janvier 2023

Sommaire

1 L'antibiorésistance

Les antibiotiques sont développés pour contrer les infections dues aux bactéries. Certaines bactéries peuvent acquérir une résistance à des antibiotiques, via l'obtention de nouveaux gènes ou par la mutation de gènes existants. L'antibiorésistance représente un grand risque pour la santé publique, il est donc important de la limiter. Cela passe notamment par une meilleure compréhension des mécanismes de résistances comme l'identification de gènes résistants ou des bactéries résistantes. Cette résistance de certaines bactéries à des antibiotiques peut être traitée comme une tâche de classification en apprentissage statistique.

Dans cette étude nous aurons à notre disposition un jeu de données constitué de 3 matrices de régresseurs et une matrice réponse pour 414 bactéries :

- X_gpa, codant la présence ou l'absence de 16005 gènes,
- X_nsps, codant la présence ou l'absence de 72236 mutations génétiques,
- X_genexp, représentant l'expression génétique de 6026 gènes ;
- Y, codant la résistance ou la sensibilité à 5 antibiotiques : la Ceftazidime, la Ciprofloxacine, la Colistine, le Méropenème et la Tobramycine.

Notre objectif est de prédire la résistance des bactéries aux antibiotiques à partir des régresseurs et d'identifier quelles matrices de régresseurs sont les plus intéressantes pour cette tâche, selon l'antibiotique considéré.

Dans un premier temps, nous procéderons à une courte exploration des données. Puis, nous détaillerons notre démarche : pré-traitements utilisés sur les données, proposition d'approches de réduction de dimension, classifieurs considérés et mise en œuvre via la librairie *scikit-learn*. Enfin, nous présenterons les résultats obtenus et nous proposerons quelques pistes d'amélioration.

2 Exploration des données

Avant de nous lancer dans la prédiction de la résistance aux antibiotiques, nous avons souhaité nous pencher sur les données que nous manipulons.

Naturellement nous avons commencé par observer les types de données que nous manipulons et l'éventuelle présence de données manquantes. Sur les 4 matrices à notre disposition, 3 contiennent des données binaires (Y, X_gpa, X_nsps) tandis que X_genexp contient des données quantitatives.

Comme le montre la table 1, les matrices de régresseurs ne contiennent pas de données manquantes, mais certaines informations de résistance aux antibiotiques sont manquantes, notamment pour la Ceftazidim avec 20% de données absentes. Etant donné que la taille du jeu de données est réduite, que nous procèderons par la suite à une validation croisée et que nous ne disposerons donc pas d'un jeu de test, nous avons choisi de ne pas imputer les données manquantes afin d'éviter de fausser la généralisation des résultats. Par conséquent, les bactéries dont la résistance à un antibiotique est manquante ne seront pas utilisées lors de l'évaluation des classifieurs sur l'antibiotique correspondant.

Nous pouvons également observer sur la table 1 que les variables réponses ne sont pas toujours équilibrées : 2 fois plus de bactéries résistantes à la Méropenem, et à l'inverse 2 à 3 fois plus de bactéries susceptibles à la Tobramycin et la Colistin. De même, les gènes ou les mutations sont bien plus souvent absentes que présentes.

Table 1: Résumé des variables réponses et des régresseurs.

	Résistance					Présence		
	Tobramycin	Ceftazidim	Ciprofloxacin	Meropenem	Colistin	gpa	snps	genexp
# **NA**	8	80	56	60	0	0	0	0
# **VRAI**	130	165	199	244	85	3581	8218	NaN
# **FAUX**	276	169	159	110	329	12424	64018	NaN

Pour aller un peu plus loin, nous avons représenté nos données regroupées par clustering hiérarchique avec des cartes de chaleur afin de faire apparaître des structures. La figure 1 correspondant à la carte de chaleur ainsi obtenue pour la matrice X_gpa permet par exemple de supposer que cette matrice porte de l'information intéressante pour prédire la résistance à la Tobramycin et la Ciprofloxacine, mais probablement moins pour la Colistin. Pour celle-ci, nous avons observé par le même biais que la matrice X_genexp sera sûrement indispensable.

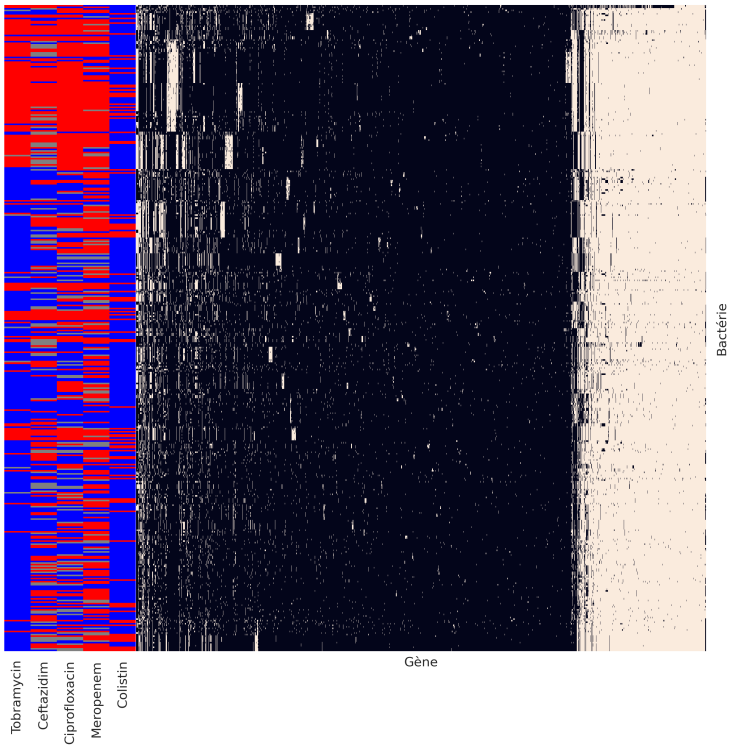


Figure 1: Carte de chaleur du clustering hiérarchique sur les lignes et les colonnes de la matrice X_gpa. Les couleurs à gauche des lignes permettent de déterminer si la bactérie est sensible (bleu) ou résistante (rouge) à l'antibiotique, le gris correspond aux données manquantes.

3 Approche pour la prédiction

3.1 Pré-traitements

Comme expliqué dans la section §2, nous avons fait le choix de supprimer les données manquantes. Cette suppression est faite de manière "intelligente" en ce sens où les bactéries dont la résistance est

absente sont éliminées uniquement pour les antibiotiques concernés et demeurent disponible pour les autres antibiotiques.

Nous nous sommes ensuite contentés de centrer/réduire les expressions génétiques de la matrice X_genexp grâce au transformateur ***StandardScaler*** de *scikit-learn*.

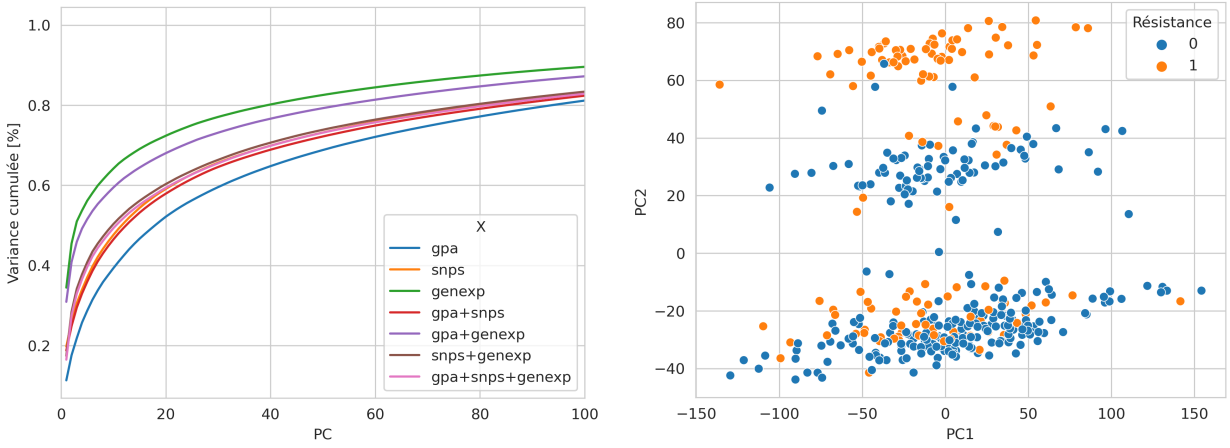
S'est ensuite posée la question de la mutualisation des informations contenues dans les 3 matrices de classifieurs. En traçant les cartes de chaleur de ces matrices, nous avons aperçu que toutes ne sont pas nécessairement pertinentes pour tous les antibiotiques. C'est pourquoi nous avons fait le choix de ne pas systématiquement agréger les matrices, mais de considérer les 7 arrangements possibles : 1 seule matrice (3), 2 matrices (3) et les 3 matrices (1).

3.2 Réduction de la dimensionnalité

Nous avons vu que nos matrices de régresseurs contiennent beaucoup de covariables, jusqu'à 94267 lorsque nous les concaténons entres-elles, relativement aux nombres d'observations dont nous disposons. Il est donc impératif de réduire cette dimensionnalité. Cela peut se faire en amont de la tâche de classification, ou alors de manière intégrée en incluant une pénalisation sur les poids du modèle associés aux regresseurs. Nous avons considéré les deux approches et nous détaillerons dans cette partie les 3 méthodes que nous avons employées en amont.

3.2.1 ACP (à Noyau)

L'Analyse en Composantes Principales (ACP) est une approche bien connue permettant de représenter les observations dans un sous-espace vectoriel dont les composantes sont décorrélées.



(a) Variance cumulée exprimée en pourcentage (b) Nuage de points selon les 2 premières dimensions

Figure 2: Représentations de l'ACP avec un noyau linéaire