Analyse de données génomiques concernant la résistance aux antibiotiques : le cas de la tuberculose

Tom AGONNOUDE, Laurence LACMAGO & Horeb SEIDOU
Tuteur: Stéphane CHRETIEN
Coach : Marion ROBERT

Contexte & Problématique

Plusieurs méthodes statistiques ont été utilisées dans le cadre de l'étude de la multirésistance de M. tuberculosis. Mais la généralisation et l'adoption clinique de ces méthodes ont été limitées par un manque d'interprétabilité et de vérifiabilité. Notre étude s'inscrit dans ce contexte et se décline en deux grandes parties:

1. étudier le modèle de réseau de neurones convolutif présenté par Anna G. Green et al. ;
2. mettre en oeuvre une approche de **prédiction conformelle** qui permettra de garantir la fiabilité des prédictions faites.

Données de l'étude

L'étude a impliqué des données sur 23 049 isolats de M.tuberculosis et a concerné **13 antibiotiques** dont 4 dits de première ligne et 9 dits de deuxième ligne.

Pour chaque isolat, nous disposons de :

- la séquence génomique et
- le statut (résistant ou susceptible) à l'un au moins des 13 antibiotiques auxquels on veut prédire la résistance.

Pour chacun des treize médicaments, on dispose des phénotypes pour au moins 250 isolats.

Données en entrée : Seuls les loci des séquences de gènes suspectés de causer la résistance ont été donnés en entrée aux modèles (18 loci maximum).

Utilisation du jeu de données : Le jeu de données a été divisé en données d'entraînement (10 201 isolats) et de test (le reste des isolats).

Méthodologie de l'étude

Au total, quatre types modèles ont été mis en œuvre :

- un modèle multi-médicaments de réseau de neurones large et profond (MD-WDNN, état de l'art);
- 13 modèles de réseau de neurones convolutif à médicament unique (SD-CNNs);
- un modèle multi-médicaments de réseau de neurones convolutif (MD-CNN);
- 13 modèles de régression logistique avec pénalité de régularisation L2 à médicament unique (modèle de benchmarck).

Performances des modèles CNN

1 Comparaison des modèles avec l'état de l'art

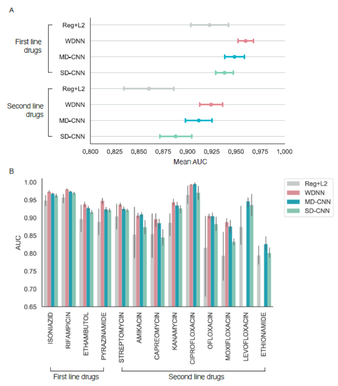


Figure 1: Comparaison des modèles CNN

Les modèles CNN ont de bien meilleures performances que la régression L2. De plus, le modèle MD-CNN est le meilleur, avec des performances semblables à celles du MD-WDNN.

2 Généralisation et applicabilité des modèles CNN dans le monde réel

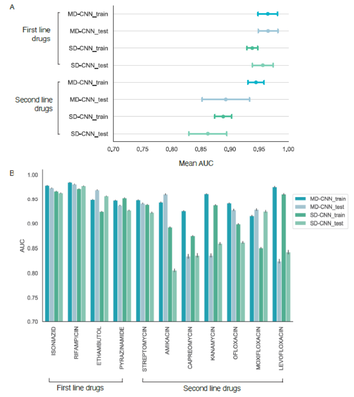


Figure 2: Généralisation des modèles CNN

Le modèle MD-CNN a de bonnes performances pour les médicaments de 1^{ere} ligne mais moins pour ceux de 2^{eme} ligne, tandis que le modèle SD-CNN a de bonnes performances pour les deux types de médicaments.

3 Interprétabilité des modèles CNN

La méthode **Deep LIFT** a permis de trouver les sites nucléotides dans les séquences de M. tuberculosis qui lui confèrent le phénotype de résistance. Les résultats montrent que le modèle MD-CNN atteint des performances élevées en s'appuyant sur des corrélations entre les résistances aux médicaments.

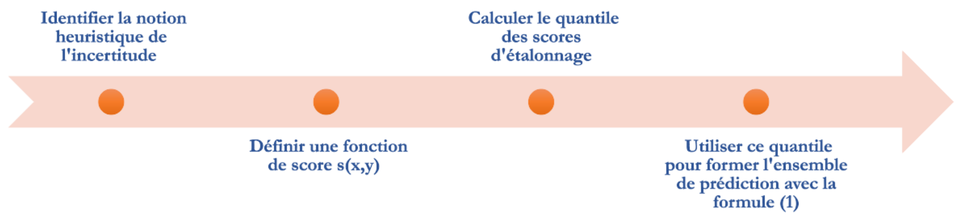
Prédiction conformelle

1 De quoi s'agit t-il?

La **prédiction conformelle** permet de transformer une notion heuristique en une notion rigoureuse de l'incertitude, en "prédisant" un ensemble de prédictions contenant la vraie valeur du paramètre avec une grande probabilité, tout en reflétant fidèlement l'incertitude du modèle. Cet ensemble doit :

- avoir une **bonne couverture**,
- être de **taille minimale**,
- être **adaptatifs** (refléter la difficulté liée à la prédiction).

Son principe est basé sur les quantiles empiriques, utilise les prédictions faites et la mise en oeuvre se fait en 4 étapes:



$$\mathcal{C}(X_{test}) = \{y : s(X_{test}, y) \leq \hat{q}\}. \quad (1)$$

2 Application aux modèles SD-CNN

Nous l'avons appliqué sur deux modèles SD-CNN : ceux de la rifampicine et de l'amikacine. Ces modèles étant des modèles de classification binaire (résistant ou susceptible) nous avons appliqué la prédiction conformelle selon la méthodologie **outlier detection** présenté par Angelopoulos et al. Les résultats sont les suivants:

Modèle SD-CNN	Erreur de type 1	Erreur de type 2
Rifampicine	9,5%	4,46%
Amikacine	9,8%	75,27%

Table 1: Tableau récapitulatif des résultats de la prédiction conformelle

Au vu de ces résultats, le modèle SD-CNN de l'antibiotique rifampicine est plus performant et prédit avec moins d'incertitude que celui de l'amikacine.

Limite

La principale limite de notre étude réside dans le fait que nous n'ayons pas eu les ressources nécessaires pour tourner le modèle principal (MD-CNN) dont les résultats pourraient être bien plus intéressants que ceux du SD-CNN.