Data Challenge

Apprentissage Statistique II – Data Challenge

Data Challenge

Objectif : prédire la résistance à la streptomycine d'une souche de *Mycobaterium tuberculosis* à partir de ses gènes *rrs*, *rpsL* et *gidB*.

Donnée d'entrée pour chaque souche :

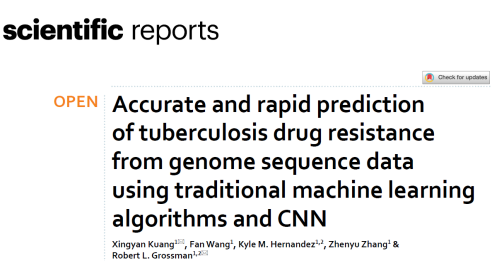
- 1 séquence nucléique contenant la concaténation des 3 gènes
- 1 phénotype binaire R/S

Jeu de données :

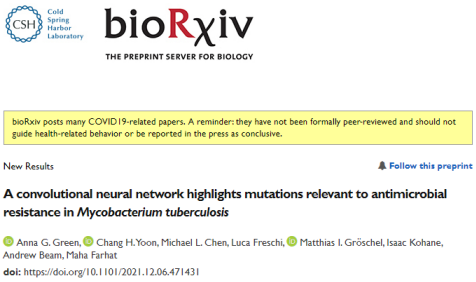
- training set : 2991 génomes (918 R et 2073 S)
- test set : 1497 génomes

Critère de performance : "balanced-accuracy" (moyenne de sensi. et spéci.)

A hot topic !



⇒ Scientific Reports, 14 février 2022.



⇒ bioRxiv, 7 décembre 2021.

Objectifs

Objectif

Objectif #1 : prédire la résistance à la streptomycine d'une souche de *Mycobaterium tuberculosis* à partir de ses gènes *rrs*, *rpsL* et *gidB*.

- critère de performance : "balanced-accuracy" (moyenne de sensi. et spéci.)

Objectif #2 : évaluer l'intérêt des approches de "deep learning" dans ce contexte

- construire une "baseline" au moyen d'algorithmes "standards"
- essayer de faire mieux avec différentes architectures et manières de représenter les séquences

Evaluation

A rendre pour le 09/01, 19h (via moodle) :

- prédictions sur le jeu de test
  - fichier texte **predictions-group-XXX.txt** de 1497 lignes
  - chaque ligne = R ou S (pas de 0/1, +1/-1 ou autre)
- rapport d'analyse
  - 8 pages maxi
  - expliciter les différentes architectures mises en oeuvre et leur intérêt

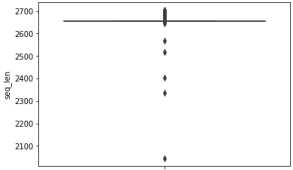
Critères d'évaluation :

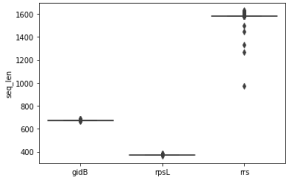
- performance (balanced accuracy)
- clarté du rapport
- créativité et exhaustivité (surprenez moi !)

Session pratique - prise en main du jeu de données

Jeu de données

- Longueur de la séquence $\sim$ 2654 bp :





$\Rightarrow$ longueur (concaténée) minimum / maximum = 2043 / 2706

- Les jeux de train / test :

```
df_train.head(10)
```

idx	label	seq
0	S	ATGTCTCCGATCGAGCCCGCGGCGTCTGCGATCTTCGGACCGCGGC...
1	R	ATGTCTCCGATCGAGCCCGCGGCGTCTGCGATCTTCGGACCGCGGC...
2	S	ATGTCTCCGATCGAGCCCGCGGCGTCTGCGATCTTCGGACCGCGGC...
3	S	ATGTCTCCGATCGAGCCCGCGGCGTCTGCGATCTTCGGACCGCGGC...
4	S	ATGTCTCCGATCGAGCCCGCGGCGTCTGCGATCTTCGGACCGCGGC...
5	S	ATGTCTCCGATCGAGCCCGCGGCGTCTGCGATCTTCGGACCGCGGC...
6	S	ATGTCTCCGATCGAGCCCGCGGCGTCTGCGATCTTCGGACCGCGGC...
7	S	ATGTCTCCGATCGAGCCCGCGGCGTCTGCGATCTTCGGACCGCGGC...
8	S	ATGTCTCCGATCGAGCCCGCGGCGTCTGCGATCTTCGGACCGCGGC...
9	S	ATGTCTCCGATCGAGCCCGCGGCGTCTGCGATCTTCGGACCGCGGC...

```
df_test.head(10)
```

idx	seq
0	ATGTCTCCGATCGAGCCCGCGGCGTCTGCGATCTTCGGACCGCGGC...
1	ATGTCTCCGATCGAGCCCGCGGCGTCTGCGATCTTCGGACCGCGGC...
2	ATGTCTCCGATCGAGCCCGCGGCGTCTGCGATCTTCGGACCGCGGC...
3	ATGTCTCCGATCGAGCCCGCGGCGTCTGCGATCTTCGGACCGCGGC...
4	ATGTCTCCGATCGAGCCCGCGGCGTCTGCGATCTTCGGACCGCGGC...
5	ATGTCTCCGATCGAGCCCGCGGCGTCTGCGATCTTCGGACCGCGGC...
6	ATGTCTCCGATCGAGCCCGCGGCGTCTGCGATCTTCGGACCGCGGC...
7	ATGTCTCCGATCGAGCCCGCGGCGTCTGCGATCTTCGGACCGCGGC...
8	ATGTCTCCGATCGAGCCCGCGGCGTCTGCGATCTTCGGACCGCGGC...
9	ATGTCTCCGATCGAGCCCGCGGCGTCTGCGATCTTCGGACCGCGGC...

Session pratique

- Pré-traitement des séquences
  - One-hot encoding, kmers tokens, kmers profiles...
  - Ouvrir le notebook `DataChallenge_sequence-preprocessing.ipynb`

- Baseline Random Forest
  - validation croisée / GridSearchCV par RF sur profils de kmers
  - Ouvrir le notebook `DataChallenge_RF-baseline.ipynb`