

Data Challenge

Apprentissage Statistique II – Data Challenge

Objectif : prédire la résistance à la streptomycine d'une souche de *Mycobacterium 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

OPEN

Accurate and rapid prediction of tuberculosis drug resistance from genome sequence data using traditional machine learning algorithms and CNN

Xingyan Kuang^{1,2}, Fan Wang¹, Kyle M. Hernandez^{1,2}, Zhenyu Zhang¹ & Robert L. Grossman^{1,2,3}

⇒ Scientific Reports, 14 février 2022.



bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

bioRxiv posts many COVID19-related papers. A reminder: they have not been formally peer-reviewed and should not guide health-related behavior or be reported in the press as conclusive.

New Results

[Follow this preprint](#)

A convolutional neural network highlights mutations relevant to antimicrobial resistance in *Mycobacterium tuberculosis*

Anna G. Green, Chang H. Yoon, Michael L. Chen, Luca Freschi, Matthias I. Gröschel, Isaac Kohane, Andrew Beam, Maha Farhat

doi: <https://doi.org/10.1101/2021.12.06.471431>

⇒ bioRxiv, 7 décembre 2021.

Outline

UE App.Stat II

Objectif

Jeu de données

Objectifs

Objectif #1 : prédire la résistance à la streptomycine d'une souche de *Mycobacterium 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

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

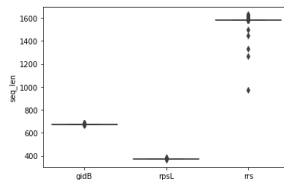
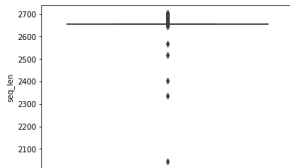
Critères d'évaluation :

- ▶ 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 ~ 2654 bp :



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

- Les jeux de train / test :

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

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

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

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

► 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`