

✓ Présentation du Data Challenge : identification de souches de *Escherichia coli* par spectrométrie de masse (MALDI-TOF)

Ce TP est basé sur les travaux de Christner *et al.* (2014) publiés dans *PLOS ONE*, où la spectrométrie de masse de type MALDI-TOF a été utilisée pour identifier rapidement des souches d'*Escherichia coli* responsables d'une épidémie.

Objectif de l'étude

Utiliser les spectres MALDI-TOF pour **typer les souches bactériennes** en identifiant celles associées à l'épidémie (*E. coli* O104:H4).

Nous allons explorer ces données, les visualiser, appliquer une **réduction de dimension** (PCA) puis construire un **modèle de classification** à l'aide de `scikit-learn`.

Introduction à la spectrométrie de masse (MALDI-TOF)

La **spectrométrie de masse MALDI-TOF** (Matrix-Assisted Laser Desorption Ionization Time-Of-Flight) est une méthode utilisée pour analyser des biomolécules comme les protéines bactériennes.

Que mesure-t-on ?

Chaque échantillon génère un **spectre** : une courbe représentant l'intensité en fonction du **rapport masse/charge (m/z)**.

Les **pics du spectre** correspondent à différentes protéines ou fragments présents dans l'échantillon.

Comment obtient-on la matrice d'intensité ?

Après acquisition et traitement du signal :

- Chaque spectre (chaque souche bactérienne) devient une **ligne** dans une matrice
- Chaque **colonne** correspond à un **pic m/z détecté**
- La valeur représente l'**intensité du signal** à ce m/z

Description des données

1. RawIntensityMatrixChristner.tsv

- **Format** : fichier `.tsv`
- **Contenu** : valeurs d'intensité du spectre MALDI-TOF après traitement des signaux
- Chaque ligne = un spectre (une souche)
- Chaque colonne = un pic m/z

2. MetadataShigatoxChristner.csv

- **Format** : fichier `.csv`
- **Dimensions** : 891×4 colonnes
- **Colonnes** :
 - `Toxigenic_status` : statut du clone :
 - `norec` : non lié à l'épidémie (190 souches)
 - `orec` : lié à l'épidémie (104 souches)
 - `ref` : souches de référence (3 réplicats biologiques)
 - `id_number` : identifiant numérique unique de la souche
 - `Strain_number` : nom de la souche (nommage dans l'étude)
 - `spot` : position de dépôt sur la plaque MALDI
 - `type_of_extraction` : toujours "fae" (formic acid extraction)

Ce Data Challenge vous guide dans l'importation, l'exploration et la modélisation de ces données pour identifier les souches responsables de l'épidémie.

Objectifs du Data Challenge

- Évaluer la capacité de la **spectrométrie de masse** à détecter la présence d'un clone pathogène d'*E. coli* à l'aide d'**outils de Machine Learning**.
 - Vous devez créer un modèle de Machine Learning performant pour détecter efficacement les clones responsables nommés `orec` dans le fichier `MetadataShigatoxChristner.csv`
-

1. Import des bibliothèques

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split, cross_val_score
from sklearn.ensemble import RandomForestClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import classification_report, confusion_matrix

import warnings
warnings.simplefilter("ignore")
```

2. Chargement des données


```
# Charger la matrice d'intensité (.tsv)
url_matrix = "https://raw.githubusercontent.com/agodmer/MSData/main/ShigaToxigenicEscherichia/RawIntensityMatrixChristnerSN3.tsv"
intensity_df = pd.read_csv(url_matrix, sep="\t")

# Charger les métadonnées (.csv)
url_meta = "https://raw.githubusercontent.com/agodmer/MSData/main/ShigaToxigenicEscherichia/MetadataShigatoxChristner.csv"
meta_df = pd.read_csv(url_meta)

# Aperçu des données
intensity_df.head()
```

3. Dimensions des jeux de données

```
print("Matrice d'intensité :", intensity_df.shape)
print("Matrice des métadonnées :", meta_df.shape)
```



```
Matrice d'intensité : (891, 956)
Matrice des métadonnées : (891, 5)
```

4. A vous de jouer !

