Prédiction d'espèces à partir de séquences ADN

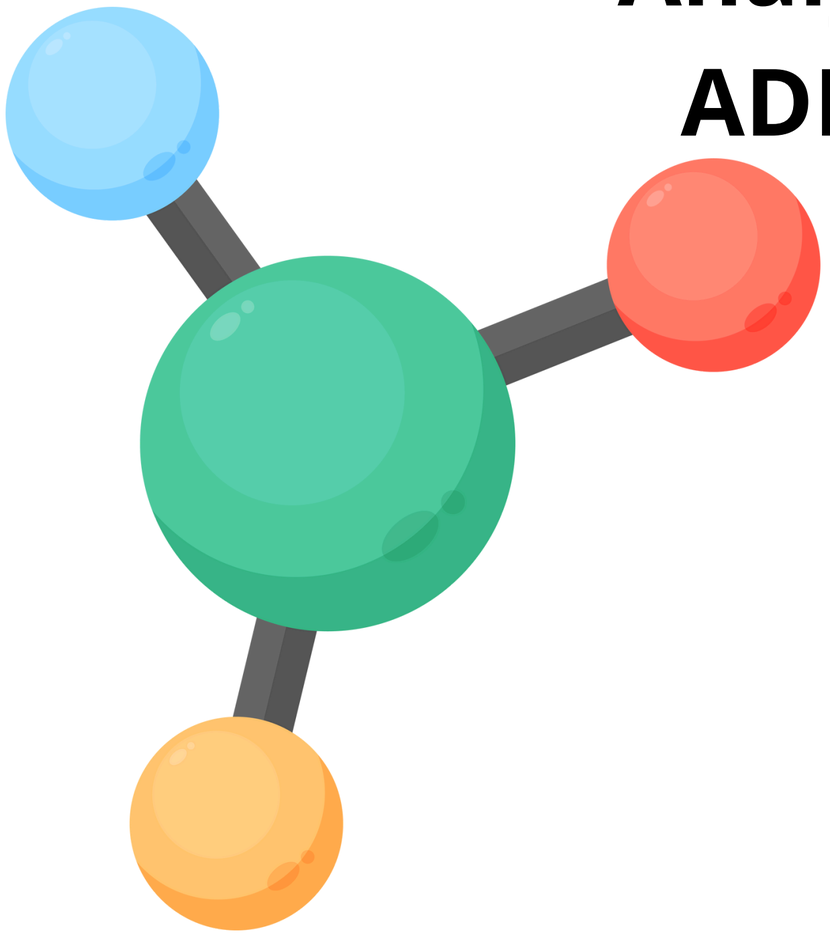
DNA Barcoding

ASRI WASSILA

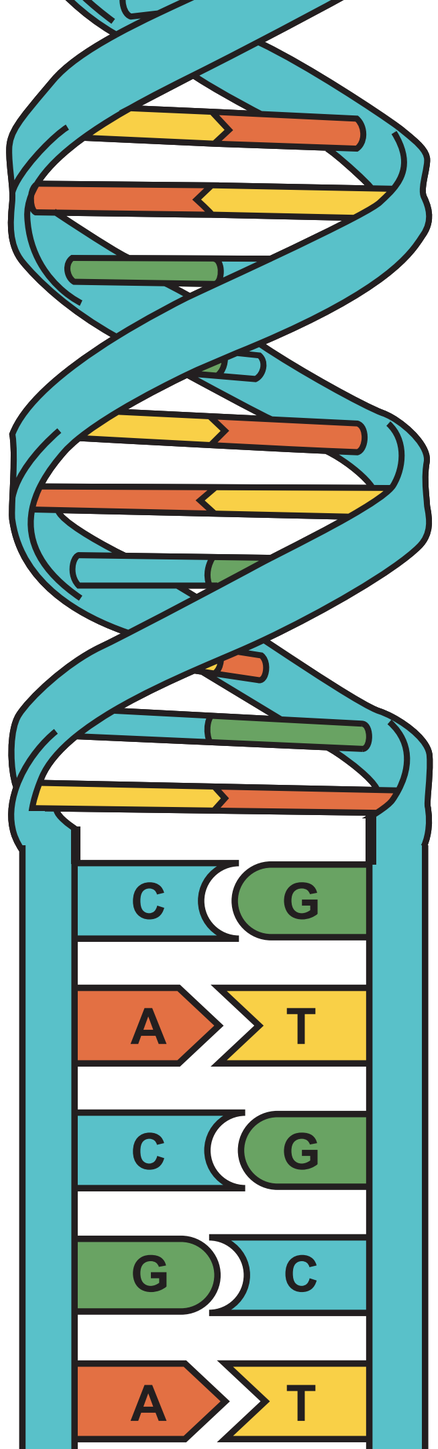
UTILISATION DE MODÈLES MACHINE LEARNING POUR LA CLASSIFICATION DES PLANTES

Présentation du projet

Analyse de séquences ADN pour identifier les espèces



Utilisation de méthodes de ML et DL



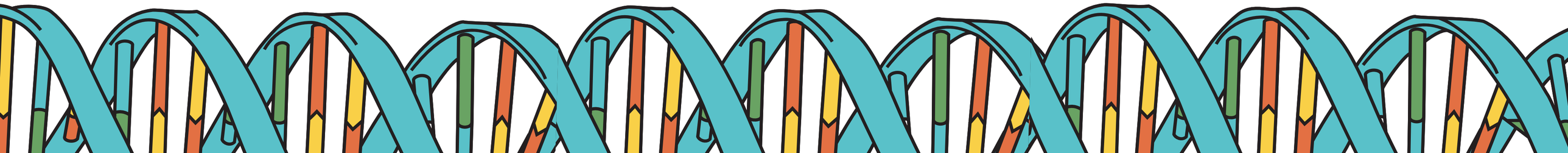
WHAT IS DNA BARCODING?

CONTEXTE DU PROBLÈME

- LE DNA BARCODING EST UNE MÉTHODE D'IDENTIFICATION DES ESPÈCES BASÉE SUR L'ADN
- UTILISE DES RÉGIONS GÉNOMIQUES COURTES ET STANDARDISÉES
- CHAQUE ESPÈCE POSSÈDE UNE SIGNATURE GÉNÉTIQUE UNIQUE
- PERMET L'IDENTIFICATION RAPIDE DES PLANTES ET LA CONSERVATION DE LA BIODIVERSITÉ

Objectif du Projet

- Prédire automatiquement l'espèce végétale à partir d'une séquence ADN
- Utiliser plusieurs modèles :
- Random Forest
- SVM
- XGBoost
- Réseaux de Neurones
- Améliorer les performances avec des techniques avancées de prétraitement



DONNÉES UTILISÉES

	Gene_Region	Class	Order	Family	Genus	Species	Sequence
0	rbcLa	Magnoliopsida	Gentianales	Apocynaceae	Schizoglossum	Schizoglossum atropurpureum	AGTGTTGGATTCAAAGCCGGTGTTAAAGAGTACAAATTGACTTATT...
1	matK	Magnoliopsida	Gentianales	Apocynaceae	Schizoglossum	Schizoglossum atropurpureum	GATATACTAATACCCTACCCTGTTCATCTGGAAATCTTGGTTCAAA...
3	rbcLa	Magnoliopsida	Gentianales	Apocynaceae	Schizoglossum	Schizoglossum atropurpureum	AGTGTTGGATTCAAAGCCGGTGTTAAAGAGTACAAATTGACTTATT...
4	matK	Magnoliopsida	Gentianales	Apocynaceae	Schizoglossum	Schizoglossum atropurpureum	GATATACTAATACCCTACCCTGTTCATCTGGAAATCTTGGTTCAAA...
11	matK	Magnoliopsida	Gentianales	Apocynaceae	Schizoglossum	Schizoglossum bidens	TCTGGAAATCTTGGTTCAAACCCTTCGCTATTGGGTAAAGGATGCC...

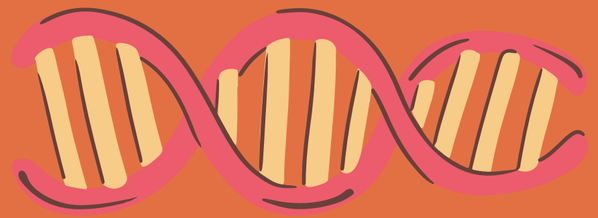
Colonnes principales :

- Gene Region
- Species
- Sequence

SOURCE : **BOLD SYSTEMS**

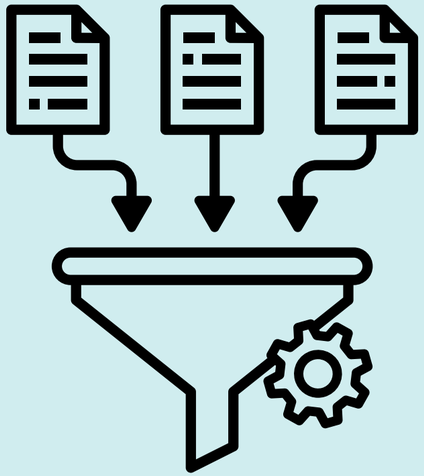
Volume des données après nettoyage :

- 3322 séquences
- 149 espèces

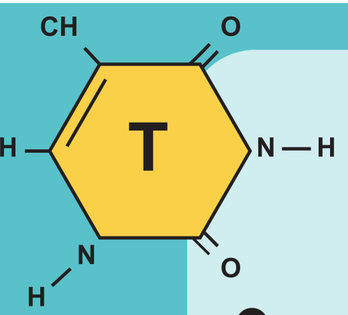


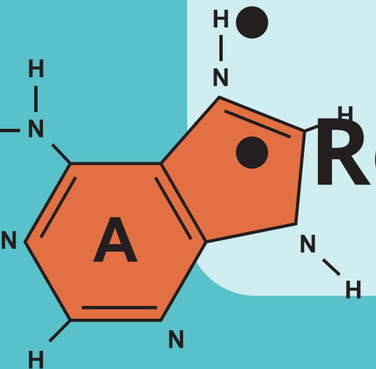
PRÉTRAITEMENT DES DONNÉES

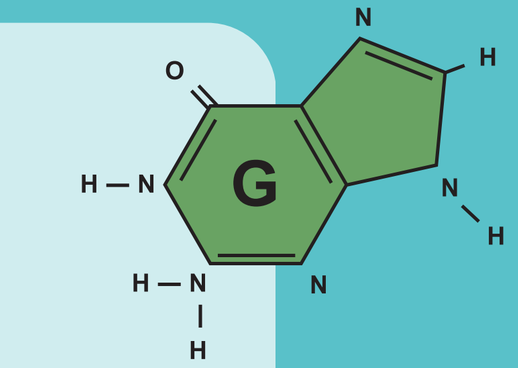
- SUPPRESSION DES COLONNES INUTILES
- ENCODAGE DES SÉQUENCES EN K-MERS (K=3 À K=8)
- VECTORISATION AVEC COUNTVECTORIZER ET TF-IDF
- INTÉGRATION DE LA RÉGION GÉNIQUE COMME VARIABLE EXPLICITE

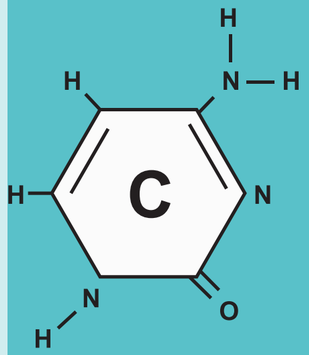


MODÉLISATION DES MODÈLES









Modèles utilisés :

- **Random Forest**
- **Naive Bayes**
- **SVM (SGDClassifier)**
- **XGBoost**
- **Réseaux de Neurones (CNN, MLP, RNN)**

Évaluation des Modèles

Modèle	Accuracy	F1 Macro	Précision Macro	Recall Macro
Random Forest	0.87	0.84	0.86	0.85
Naive Bayes	0.58	0.53	0.58	0.55
SVM (SGDClassifier)	0.84	0.80	0.82	0.83
XGBoost	0.83	0.80	0.82	0.81
Réseau de Neurones (MLP)	0.80	0.77	0.78	0.79
CNN	0.34	–	–	–
RNN (LSTM)	0.05	–	–	–

Le Random Forest reste le meilleur choix pour ce projet, notamment avec une vectorisation de type k-mer (k=6) et un encodage explicite de la région génique. Les approches de deep learning n'ont pas surpassé les méthodes classiques ici, ce qui souligne l'importance de l'adéquation entre modèle, volume de données, et qualité de la représentation des séquences.

AMÉLIORATIONS APPORTÉES

- OPTIMISATION DES HYPERPARAMÈTRES AVEC GRIDSEARCH
- TEST DE DIFFÉRENTES TAILLES DE K-MERS (3 À 8)
- UTILISATION DE STRATIFIED K-FOLD POUR ÉQUILIBRER LES CLASSES

PRÉDICTION RÉELLE

ON VA PRÉDICTER L'ESPÈCE DE LA SÉQUENCE SUIVANTE :

SEQUENCE="GGTGTTGGATTTCAAGCTGGTGTTAAAGATTATAAATTGACTTACTACACCCCAGAGTATGAAACTAAGGATACTGATATC
TTGGCAGCATTCCGAGTAAGTCCTCAGCCTGGGGTTCCGCCCGAAGAAGCAGGGGCTGCAGTAGCTGCCGAATCTTCTACTGGTACATGGA
CAACTGTTTGGACTGATGGACTTACCAGTCTTGATCGTTACAAAGGACGATGCTATCACATCGAGCCTGTTGCTGGGGAAGACAACCAATG
GATCTGTTATGTAGCTTATCCATTAGACCTATTTGAGGAGGGTTCCGTTACTAACATGTTTACTTCCATTGTGGGTAACGTATTTGGGTTCA
AAGCCCTACGTGCTCCCCCCTACTTATTCAAAAACTTTCCAAGGCCCGCCTCATGGTATCCAAGTTGAAAGAGATAAGTTGAACAAGTAT
GGTCGTCCTTTATTGGGATGTACTATTAAACCAAAATTGGGATTATCCGC AAAAAATTATGGTAGAGCGTGTTATGAGTGTCTA"

GENE_REGION="RBCLA"

Résultats :

- **RANDOM FOREST** :Espèce prédite : Hordeum jubatum
- **SVM** :Espèce prédite : Hordeum jubatum

les deux modèles ont bien prédictés l'espèce.

MERCI

Le code complet, les données utilisées et les notebooks de ce projet sont disponibles sur GitHub :

👉 Accéder au Repository GitHub : https://github.com/Wassila00/DNA-Barcoding.git