RÉPUBLIQUE FRANÇAISE
Liberté Égalité Fraternité

INRAE

GenIALearn

PROJET EXPLORATOIRE

2021-2023

Application du machine learning et deep learning pour perfectionner la sélection génomique animale

Coordination
Eric Barrey
UMR GABI
eric.barrey@inrae.fr
Didier Boichard
UMR GABI
didier.boichard@inrae.fr

Mots clés
Génomique
Interactions géniques
Apprentissage statistique
Apprentissage automatique
Apprentissage profond

Unités INRAE impliquées
GABI
MIA Paris Saclay

Partenaires
UEVE Université Paris-Saclay

Contexte et enjeux

Le développement de la sélection génomique - et des autres analyses « omiques » telles que la métagénomique, transcriptomique, métabolomique et protéomique - permet aujourd'hui de caractériser les animaux grâce à des milliers de mesures. Ces données massives sont intégrées dans des modèles, afin de prédire des caractères de production avec la plus grande précision possible.

Les modèles les plus couramment utilisés en prédiction génomique (modèle génétique additif type GBLUP) sont très efficaces pour prédire la valeur génétique des animaux sur quelques caractères génétiquement corrélés. En revanche, ce type de modèle ne permet pas d'intégrer un très grand nombre de mesures hétérogènes, ni de prédire beaucoup de caractères en sortie sans connaitre leurs corrélations génétiques. De plus, ce modèle reste limité pour tenir compte des nombreuses interactions non-linéaires qui interviennent entre les régions du génome ou des facteurs environnementaux.

Afin de lever ces verrous, nous proposons d'utiliser les méthodes d'apprentissage statistique (Machine Learning) et d'apprentissage profond, issues de l'IA, pour à la fois traiter les informations génétiques additives mais également les informations génétiques non-linéaires présentes dans les données massives de génotypage.

©kjpargeter, CS - freepik

Métaprogramme
DIGIT-BIO

digitbio@inrae.fr
www.inrae.fr/digitbio/

Objectifs

Le projet GenIALearn propose d'évaluer les performances des méthodes d'apprentissage statistique et profond pour la prédiction conjointe de multiples caractères complexes, par l'intégration de données massives de génotypage. Deux grandes familles de méthodes seront comparées entre-elles et à la méthode de référence le GBLUP :

- d'une part, les méthodes d'apprentissage ensemblistes (random forests, gradient boosting), couplées à une étape d'apprentissage de représentation des données d'entrées, afin de proposer des niveaux de prédiction de référence ;
- d'autre part, les réseaux de neurones avec différentes architectures, couplés à une étape d'apprentissage profond sur des bases de données massives, permettront de concevoir et de comparer des modèles prédictifs pour la sélection génomique animale.

Partenaires

Départements INRAE	Unités INRAE	Expertises
GA	GABI	Phénotypage fin de caractères complexes, multi-omiques (génotypage, trans-criptomique, métagénomique, métabolomique), prédictions des valeurs génétiques des reproducteurs et prédiction multi-caractères complexes
MathNum	MIA Paris Saclay	Modélisation, apprentissage statistique, machine learning, données de grande dimension et hétérogènes, application aux sciences du vivants
Partenaires		**Expertises**
UEVE Université Paris-Saclay	UBISC	Méthodes de construction de réseaux de neurones et deep learning, applications à l'analyse d'images et du transcriptome