RÉPUBLIQUE FRANÇAISE
Liberté Égalité Fraternité

INRAE

GenIALearn

PROJET EXPLORATOIRE

2021-2023

Application du machine learning et du 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
UMR GABI
MIA – Paris Saclay

Partenaires
UEVE, Université Paris-Saclay
IBISC

© kjpargeter, CS - freepik

Contexte, enjeux et objectifs

Le développement de la sélection génomique – et des autres analyses « omiques » – 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 des animaux avec la plus grande précision possible. Les modèles les plus couramment utilisés en prédiction génomique (modèles génétiques additifs de type GBLUP[1]) 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 connaître 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.

Le projet GenIALearn propose d'évaluer la performance des méthodes d'apprentissage statistique et profond pour la prédiction conjointe de multiples caractères complexes chez le bovin laitier en intégrant des données massives de génotypage. Deux approches complémentaires ont été mises en œuvre - des méthodes ensemblistes (machine learning) et des réseaux de neurones (deep learning) - pour prédire 33 caractères phénotypiques (production, morphologie, boiterie, fertilité, etc.) à partir du seul génotype, et sont comparées au modèle de référence GBLUP.

Pour mener à bien le projet GenIALearn, les équipes ont utilisé un jeu de données massives de génotypes et phénotypes de bovins laitiers (113 599 femelles Holstein) produites et gérées par l'UMR GABI, porteuse du projet.

Résultats

Acquisition de moyens de calculs adaptés et évaluation des méthodes d'apprentissage

Le projet a permis de financer l'acquisition d'un serveur GPU[2] et d'un serveur d'archivage qui ont été intégrés à la plateforme ColabIA, entièrement dédiée aux calculs d'intelligence artificielle au sein du Data Center Toulouse. Les différents stages de Master 2 durant le projet ont testé plusieurs méthodes : certaines se sont révélées inadaptées

[1] Genomic Best Linear Unbiased Prediction
[2] Graphics processing unit (serveur équipé de plusieurs processeurs graphiques)

au contexte, tandis que d'autres montrent un réel potentiel (mais nécessiteront des améliorations pour être pleinement efficaces pour ce type d'applications).

Le projet a notamment permis d'observer que :

- Sur 33 caractères phénotypiques étudiés, environ une dizaine sont mieux prédits par les modèles d'IA explorés. Cependant, sur la majorité des caractères le modèle de référence GBLUP reste suffisant. Les méthodes explorées offrent cependant l'avantage de la rapidité, par rapport au GBLUP unicaractère répété individuellement 33 fois.
- Les modèles de deep learning (réseaux de neurones) semblent globalement plus adaptables et performants que les modèles ensemblistes.
- Parmi les modèles basés sur des réseaux de neurones, les modèles génératifs de type WGAN-GP parviennent à bien simuler des génotypes artificiels très réalistes selon les analyses (PCA, métriques de distance). C'est une piste prometteuse à explorer pour améliorer l'apprentissage des modèles de prédiction en sélection génomique.

Perspectives

Poursuite de la thématique de recherche dans deux thèses

Les membres du projet poursuivent leur collaboration pour développer le partage des ressources (grands jeux de données, plateforme de calculs ColabIA). Le projet a par ailleurs abouti au financement de deux thèses en IA appliquée à la sélection génomique au sein de l'unité GABI :

- la **thèse de Sihan Xie (DeepSelectGene, 2024-2026)**, financée par le métaprogramme Digit-Bio, qui vise à développer des méthodes d'apprentissage pour les espèces pour lesquelles il n'existe des données génotype-phénotype que pour quelques milliers d'animaux.
- La thèse de **Fatima Shokor (2022-2025)** financée par APIGENE, vise à développer des modèles d'IA appliqués à la prédiction des phénotypes par le croisement de races bovines.

Infrastructure de calcul dédiée à l'IA

La plateforme ColabIA, dédiée aux applications d'intelligence artificielle, s'inscrit dans la durée. Sa maintenance et son développement se poursuivent grâce à la collaboration active entre les membres du projet GenIALearn et l'unité d'épidémiologie EPIA (Clermont-Ferrand). Un nouvel investissement réalisé en 2024 a permis de renforcer les capacités de calcul GPU et de stockage, ouvrant la voie à l'exploration de modèles plus complexes sur des jeux de données d'apprentissage plus volumineux.

Ce projet interdisciplinaire a favorisé la mise en place de collaborations à la fois entre équipes de l'unité, entre départements d'INRAE, et avec des partenaires externes, notamment le laboratoire IBISC (INRAE – Université d'Évry Val-d'Essonne / Université Paris-Saclay). Cette dynamique collaborative se poursuit au travers de l'amélioration continue de la plateforme ColabIA et de l'encadrement de doctorants.. De plus, les unités impliquées dans GenIALearn ont contribué à la construction du Cluster DATA IA de l'Université Paris-Saclay, dans le cadre du programme France 2030 piloté par l'Agence Nationale de la Recherche et sont aujourd'hui parties prenantes de ce cluster.

Publications

- Xie, S., Tribout, T., Boichard, D., Hanczar, B., Chiquet, J., & Barrey, E. (2025). Deep Generative Models for Discrete Genotype Simulation. BioRxiv, 2025.08.08.669289. https://doi.org/10.1101/2025.08.08.669289
- Shokor, F., Croiseau, P., Gangloff, H., Saintilan, R., Tribout, T., Mary-Huard, T., & Cuyabano, B. C. D. (2024) Deep Learning and GBLUP Integration: An Approach that Identifies Nonlinear Genetic Relationships Between Traits. bioRxiv https://doi.org/10.1101/2024.03.23.585208
- Shokor, F., Croiseau, P., Gangloff, H., Saintilan, R., Tribout, T., Mary-Huard, T., & Cuyabano, B. C. D. (2025). Deep learning and genomic best linear unbiased prediction integration: An approach to identify potential nonlinear genetic relationships between traits. Journal of Dairy Science, 108(6), 6174–6189. https://doi.org/10.3168/jds.2024-26057

Voir l'ensemble des publications et communications issues du projet sur https://digitbio.hub.inrae.fr