Stage Master 2
2022

Application du Deep Learning à la sélection génomique: Apprendre une nouvelle représentation des données génomiques.

Unité d'accueil

UMR1313 Génétique Animale et Biologie Intégrative
Domaine de Vilvert – Bât 211
78350 Jouy-en-Josas
Direction : Mathilde DUPONT-NIVET

Equipe Biologie Intégrative et Génétique Equine (BIGE)
Animateur : Eric Barrey, DR2
Ingénieur de recherche : Anne Ricard, IR

Thématique de recherche de BIGE et projet GenIALearn :
Web : http://www6.jouy.inra.fr/gabi/les-Recherches/Equipes-de-recherche/BIGE

Co-encadrement Equipe IBISC – UEVE – Université Paris-Saclay : Pr Blaise Hanczar, PhD

Sujet

Objectif du projet :

L'objectif du stage est d'apprendre des réseaux de neurones prédisant le phénotype d'individus à partir de leurs données génomiques. Afin de réduire la dimension des données (50K à 670K SNP), nous nous intéresserons en particulier aux méthodes d'apprentissage de représentation des données afin d'extraire des caractéristiques pertinentes pour la prédiction.

Méthodologies

Afin de construire un modèle prédictif performant, il est nécessaire que le réseau de neurones capture les relations entre la sortie et les variables d'entrée mais aussi entre les variables. Pour cela nous nous inspirerons des travaux de représentation de connaissance dans le domaine du traitement du langage naturel qui sont désormais largement utilisés tel que Word2Vec mais aussi des méthodes génératives tel que les VAE. L'idée est d'apprendre une nouvelle représentation des SNP plus compacte et plus informative. Cette représentation aura appris les relations entre SNP de manière non supervisée à partir de larges jeux de données. Cette nouvelle représentation sera utilisée à l'entrée des modèles prédictifs classiques. Des connaissances du domaine comme la localisation des SNP, pourront être intégrées dans l'architecture du réseau de neurones afin de facilité son apprentissage.

Les données massives qui seront traitées seront des caractères de production laitière de vaches laitières et leurs génotypes analysés sur des puces de génotypage de haute densité. A terme, l'applications sera la

prédiction génomique de la production laitière multi-caractères grâce aux informations massives de génotypage.

Profil de formation

- Master ou école d'ingénieur en informatique, mathématique appliquée ou bioinformatique.
- Solides compétences en apprentissage automatique et en particulier dans les modèles de réseau de neurones. Programmation en Python, Tensorflow ou Pytorch.
- Notions de génétique statistique, génomique e bioinformatique : modèle mixte, GWAS, variabilité génétique, génotypage, SNP

Responsables à contacter : lettre de motivation et CV

Eric Barrey: eric.barrey@inrae.fr
Blaise Hanczar: blaise.hanczar@ibisc.univ-evry.fr
Anne Ricard: anne.ricard@inrae.fr