PhD proposal at the University of Angers : Machine learning for Exploring Subdominance in PolyPLOID Genomes

Research unit : IRHS (Institute of Research in Horticulture and Seeds)

Co-supervision : LAREMA (Department of Mathematics at the University of Angers)

Scientific context :
Whole Genome Duplications (WGD), which are very common in plants, seem to coincide with periods of extinction or global change. Apple underwent a WGD 27 Mya ago (Lallemand et al., 2023), and since footprints of this WGD can still be found in the genomes of modern varieties, the apple tree is an interesting organism for studying the evolution of gene families after WGD (Daccord et al., 2017). From a general perspective, understanding the role of duplicated chromosomes and their contribution to phenotype development is a major challenge in the context of climate change.

Biological questions :

Two duplicated genes are called ohnologs if they are the consequence of a WGD event. In Lallemand et al. (2003), we have shown, thanks to a bioinformatic approach, that there exists an imbalance between ohnolog fragments. Some chromosomal fragments contribute more than their ohnologs to the phenotypic variation. We have named this phenomenon chromosomal subdominance. During this phd, we will tackle the following biological questions:

- Can we confirm and capture this imbalance through genomic prediction, which takes into account allelic variations between individuals?
- Can we take advantage of the knowledge of this imbalance to predict the phenotype more accurately ?

Mathematical questions :

We will also tackle the following mathematical questions:

- Can we consider probability distributions within a neural network, and use them as prior distributions on parameters, in the same way as what is done in mixed model tradionally used by geneticists in genomic prediction ? Such model would lead to a better understanding of the bias of Artificial Intelligence and a better understanding of the decisions made by algorithms. It would also help to improve predictions.

- Can we introduce a genomic version of Random Forests taking into account the genomic covariance between individuals while building classification trees ?

Main steps of the phd:

To begin with, a simulation study will be carried out using the REFPOP apple population (Jung et al, 2020, 2022). The phenotype will be simulated by considering various possible links (additivity, epistasis, dominance, non-linearity ...) between phenotype and genotype at QTLs (i.e. locations of the genome responsible for the variation of quantitative trait). In terms of machine learning, the preferred methods will be Genomic BLUP, random forests, Lasso, Elastic-Net, SVM, RKHS and neural networks. For each simulated trait architecture, we'll extract the best statistical learning method able to capture the imbalance between ohnologs.

In a second step, we will try to improve existing statistical methods in genomic prediction, taking advantage of this imbalance. Given the proximity between mixed models in genomics and spatial statistics, we will built on recent mathematical results in spatial statistics (Wikle and Zammit-Mangion 2023) to improve existing methods in genomic prediction. For instance, we will focus on neural networks and on random forests. In neural networks, Chen et al. (2021) introduced DeepKriging, a deep neural network where the spatial dependency is modeled by adding an extra layer to approximate the spatial process using a basis of functions. For random forest, Saha et al. (2021) suggested, in order to build a decision tree, to replace the least-squares criterion at each node split by an optimization taking into account the spatial correlation structure induced by a Gaussian process.

In order to be more familiar with these new methods, we will consider their associated packages : RandomForestsGLS (Saha et al., 2021), and the Python code of DeepKriging (https://github.com/aleksada/DeepKriging). We will try to improve Deep Kriging (Chen et al., 2021) and the random forests (Saha et al., 2021), by elaborating new mathematical formulas dedicated to genomics. The goal is to reduce the prediction error, and to quantify the information loss (in terms of prediction accuracy) when the two ohnologs are not included in the prediction model (cf. Rabier and Grusea 2021, in another context).

Skills :
- Statistical learning (Random forest, Neural networks, Lasso ...), high-dimensional data analysis, mixed model
- R or Python
- Evolutionary biology would be a plus

To apply :
Apply at https://theses.doctorat-bretagneloire.fr/vaame/campagne-2024
before 05/17/2024

Contact :
Charles-Elie Rabier : charles-elie.rabier@univ-angers.fr
Claudine Landès : claudine.landes@univ-angers.fr
Fabien Panloup : fabien.panloup@univ-angers.fr

References :
- Bartlett, P. L., Montanari, A., Rakhlin, A. (2021). *Deep learning: a statistical viewpoint. Acta numerica*, 30, 87-201
-Belkin, M., Hsu, D., Ma, S., Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32), 15849- 15854
- Chen, W., Li, Y., Reich, B. J., Sun, Y. (2021). Deepkriging: Spatially dependent deep neural networks for spatial prediction. *Statistica Sinica*:10.5705/ss.202021.0277
- Fan, J., Ma, C., Zhong, Y. (2021). A selective overview of deep learning. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 36(2), 264.
- Jung, M., Keller, B., Roth, M., Aranzana, M. J., Auwerkerken, A., Guerra, W., ... Patocchi, A. (2022). Genetic architecture and genomic predictive ability of apple quantitative traits across environments. *Horticulture research*, 9, uhac028.
- Lallemand, T. et al. (2023), Insights into the Evolution of Ohnologous Sequences and Their Epigenetic Marks Post-WGD in Malus Domestica, *Genome Biology Evolution*, 15(10): evad178
- Rabier C-E, Grusea S (2021): Prediction in high dimensional linear models and application to genomic selection under imperfect linkage disequilibrium, *Journal of the Royal Statistical Society Series C*, 70(4), 1001-1026
- Saha, A., Basu, S., Datta, A. (2021). Random forests for spatially dependent data. *Journal of the American Statistical Association*, 118(541), 665-683.
- Wikle, C. K., Zammit-Mangion, A. (2023). Statistical deep learning for spatial and spatiotemporal data. *Annual Review of Statistics and Its Application*, 10, 247-270.
- Zingaretti, L. M., Gezan, S. A., Ferrão, L. F. V., Osorio, L. F., Monfort, A., Muñoz, P. R., ... Pérez-Enciso, M. (2020). Exploring deep learning for complex trait genomic prediction in polyploid outcrossing species. *Frontiers in plant science*, 11, 25.

Titre de la Thèse : Machine Learning pour l'exploration de la sous-dominance dans les génomes polyploïdes

Laboratoire d'accueil : IRHS (Institut de Recherche en Horticulture et Semence)

Coencadrement : LAREMA (Laboratoire Angevin de Recherches en MAthématiques)

Contexte scientifique :
Les duplications entières du génome (WGD), très fréquentes chez les plantes, semblent correspondre à des périodes d'extinction ou de changement global. Le pommier a subi une WGD datée à 27 Mya, et comme des traces de cette WGD persistent dans les génomes des variétés actuelles de pommiers, le pommier est un organisme de choix pour étudier l'évolution des gènes et des familles de gènes post-WGD. D'une manière générale, la compréhension du rôle des chromosomes dupliqués et de leur contribution à l'élaboration du phénotype est un enjeu majeur dans le contexte de changements climatiques.

Questions biologiques :
Deux gènes dupliqués sonts dits ohnologues si ils résultent d'un évènement de WGD. Dans Lallemand et al. (2003), nous avons montré, par une approche alliant bioinformatique et méta-analyse, qu'il existait un déséquilibre entre fragments ohnologues : certains fragments de chromosome contribuent plus que leurs ohnologues à la variation phénotypique des individus. Nous avons baptisé ce phénomène sous-dominance chromosomique. Les questions scientifiques posées sont les suivantes :

- Peut-on confirmer et capter ce déséquilibre grâce au machine learning et notamment la prédiction génomique?
- Peut-on exploiter la connaissance de déséquilibre afin de prédire plus finement le phénotype ?

Questions mathématiques :

- Ne pourrait-on pas au sein même d'un réseau de neurones injecter des lois de probabilités comme a priori sur certains paramètres, à l'instar des modèles mixtes traditionnellement utilisés par les généticiens en prédiction génomique ? Une telle modélisation permettrait une meilleure compréhension du biais de l'Intelligence Artificielle (et des décisions de l'algorithme), et de pouvoir ainsi améliorer les prédictions.
- Ne pourrait-on pas améliorer les prédictions à travers une version génomique des Forêts aléatoires prenant en compte la covariance génomique entre individus lors de la construction des arbres de classification ?

Principales étapes de la thèse et méthodologie envisagée :

On procèdera tout d'abord à une étude par simulation in silico à partir de la population de pommiers publiée par Jung et al (2022). On dispose ainsi de données SNPs de haute-densité (303 329 SNPs) pour 534 individus répartis dans six pays européens. On simulera le phénotype en considérant différents liens possibles (e.g additivité, épistasie, dominance, non linéarité ...) entre phénotype et génotype aux QTLs (QTL= position du génome ayant une influence sur la variation d'un caractère quantitatif). En termes de machine learning, les méthodes privilégiées seront le Genomic BLUP, les forêts aléatoires, le Lasso, l'Elastic-Net, les SVM, les RKHS, les réseaux de neurones. Pour chaque architecture de trait simulée, on pourra ainsi extraire la meilleure méthode d'apprentissage statistique capable de capter le déséquilibre entre fragments ohnologues.

Dans un deuxième temps, on cherchera à améliorer les méthodes statistiques existantes en prédiction génomique, afin d'améliorer la prédiction du phénotype tout en exploitant le déséquilibre. Etant donnée la proximité entre les modèles mixtes en génomique et en statistique spatiale, on s'inspirera de récents résultats mathématiques en statistique spatiale (Wikle et Zammit-Mangion, 2023). A titre d'exemple on pourra s'intéresser aux réseaux de neurones et aux forêts aléatoires. Dans le cadre des réseaux de neurones, Chen et al. (2021) ont introduit DeepKriging, un réseau de neurones profond où la dépendance spatiale est modélisée par l'ajout d'une couche supplémentaire permettant d'approximer le processus spatial à l'aide d'une base de fonctions. Pour les forêts aléatoires, Saha et al. (2021) proposent, afin de construire un arbre de décision, de remplacer à chaque fractionnement de noeud, le critère de moindres carrés par une optimisation prenant en compte la structure de corrélation spatiale induite par un processus Gaussien.

Afin de se familiariser avec ces nouvelles méthodes, on prendra en main les packages associés : RandomForestsGLS (Saha et al., 2021), et le code Python de DeepKriging (https://github.com/aleksada/DeepKriging). On cherchera à améliorer Deep Kriging (Chen et al., 2021) et les Forêts aléatoires (Saha et al., 2021), en développant des formules mathématiques propres à la génomique. On pourra notamment s'intéresser à l'erreur de prédiction, et également quantifier mathématiquement la perte d'information (en termes de précision de prédiction) lorsque les 2 chromosomes ohnologues (issus de la duplication entière du génome) ne sont pas inclus dans le modèle de prédiction (cf. Rabier et Grusea 2021, dans un autre contexte).

Compétences scientifiques et techniques requises pour le candidat :

- Apprentissage statistique (Forêts aléatoires, Réseaux de neurones, Lasso ...), Statistique en grande dimension, Modèle mixte
- Maîtrise des langages de programmation en R et/ou Python
- Des connaissances en évolution ou en biologie végétale seraient un plus

Pour candidater :
Déposer un dossier à
https://theses.doctorat-bretagneloire.fr/vaame/campagne-2024
avant le 17/05/2024

Personnes à contacter :
Charles-Elie Rabier : charles-elie.rabier@univ-angers.fr
Claudine Landès : claudine.landes@univ-angers.fr
Fabien Panloup : fabien.panloup@univ-angers.fr

Bibliographie :
- Bartlett, P. L., Montanari, A., Rakhlin, A. (2021). *Deep learning: a statistical viewpoint. Acta numerica*, 30, 87-201
-Belkin, M., Hsu, D., Ma, S., Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32), 15849- 15854
- Chen, W., Li, Y., Reich, B. J., Sun, Y. (2021). Deepkriging: Spatially dependent deep neural networks for spatial prediction. *Statistica Sinica*:10.5705/ss.202021.0277
- Fan, J., Ma, C., Zhong, Y. (2021). A selective overview of deep learning. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 36(2), 264.
- Jung, M., Keller, B., Roth, M., Aranzana, M. J., Auwerkerken, A., Guerra, W., ... Patocchi, A. (2022). Genetic architecture and genomic predictive ability of apple quantitative traits across environments. *Horticulture research*, 9, uhac028.
- Lallemand, T. et al. (2023), Insights into the Evolution of Ohnologous Sequences and Their Epigenetic Marks Post-WGD in Malus Domestica, *Genome Biology Evolution*, 15(10): evad178
- Rabier C-E, Grusea S (2021): Prediction in high dimensional linear models and application to genomic selection under imperfect linkage disequilibrium, *Journal of the Royal Statistical Society Series C*, 70(4), 1001-1026
- Saha, A., Basu, S., Datta, A. (2021). Random forests for spatially dependent data. *Journal of the American Statistical Association*, 118(541), 665-683.
- Wikle, C. K., Zammit-Mangion, A. (2023). Statistical deep learning for spatial and spatiotemporal data. *Annual Review of Statistics and Its Application*, 10, 247-270.
- Zingaretti, L. M., Gezan, S. A., Ferrão, L. F. V., Osorio, L. F., Monfort, A., Muñoz, P. R., ... Pérez-Enciso, M. (2020). Exploring deep learning for complex trait genomic prediction in polyploid outcrossing species. *Frontiers in plant science*, 11, 25.