A Machine Learning Approach to Improve Yield Prediction Model with Multi-omics Data

Hayato Yoshioka*†[1,2], Tristan Mary-Huard[1,3], Julie Aubert[1], and Hiroyoshi Iwata[2]

[1]Mathématiques et Informatique Appliquées – AgroParisTech, Université Paris-Saclay, Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement, Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement : UMR0518 – France

[2]Graduate School of Agricultural and Life Science, The University of Tokyo – Japon

[3]Génétique Quantitative et Evolution - Le Moulon (Génétique Végétale) – AgroParisTech, Université Paris-Saclay, Centre National de la Recherche Scientifique, Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement, Centre National de la Recherche Scientifique : UMR8120, Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement : UMR0320 – France

Résumé

Genomic prediction, originally designed to predict plant traits using genome-wide marker genotype data, has now broadened to incorporate various omics datasets for improved prediction models. By leveraging different types of omics data to investigate rhizosphere microorganisms, we aim to better understand the symbiotic relationships between plants and these microbial communities. This holistic approach not only enhances our predictive capabilities but also reveals the intricate interactions shaping plant-microbe dynamics in agricultural settings. However, the complexity of the omics network poses challenges in fitting predictive models, which could be addressed by integrating machine learning techniques. In this study, we develop a method for combining multi-omics data to predict yield based on the genome and rhizosphere meta-genome and meta-metabolome data. In our two-step modeling approach, we initially trained a model to predict meta-metabolomic profiles as intermediate traits using plant genomic and meta-genomic data. Subsequently, we developed a separate model to predict yield based on the meta-metabolome profiles generated by the first model, in conjunction with plant genomic and meta-genomic data. A notable characteristic of this modeling approach is its ability to make predictions without relying on the original meta-metabolome data used for model training. Our model can compensate for the metabolomic data, which is difficult to obtain in real cases. To evaluate the performance of machine learning approaches for this prediction task on a multi-omics soybean dataset, we compared Best Linear Unbiased Prediction (BLUP), widely used in genomic prediction, with Random Forest (RF), a simple but highly versatile machine learning method. We demonstrate that incorporating the predicted meta-metabolome profiles improves the accuracy of yield prediction. In particular, the RF model outperforms the BLUP model, highlighting the effectiveness of machine learning in utilizing multi-omics data. Furthermore, the importance analysis of variables in the RF model provides insight into the interactions among various omics datasets.

*Intervenant

†Auteur correspondant: hayato.yoshioka@agroparistech.fr

Mots-Clés: Multi omics, Microbiomics, Metabolomics, Genomic Prediction, Machine Learning, Mixed model equations