High-throughput phenotyping and machine learning applications

Sven Warris[1,2], Rick van de Zedde[2]
[1]Applied Bioinformatics, Wageningen University & Research, The Netherlands
[2]Netherlands Plant Eco-phenotyping Center, Wageningen University & Research, The Netherlands
E-mail: sven.warris@wur.nl, rick.vandezedde@wur.nl

1. Introduction

Linking genomic information to important traits and other phenotypic data is key in many research fields. This includes genome-wide association studies and eQTL analyses in both animal and plant species. Our main focus is to optimize plant breeding targeted on disease resistance and climate change adaptation. To include as much genetic diversity in these analyses, a large amount of genomic and genotypic information needs to be collected. And this also requires phenotypic data of several individuals of as many different genotypes and crosses as possible[1,2]. Here we present the combination of both HTP genomics and HTP phenotyping as the foundation for new machine learning approaches to facilitate these new developments.

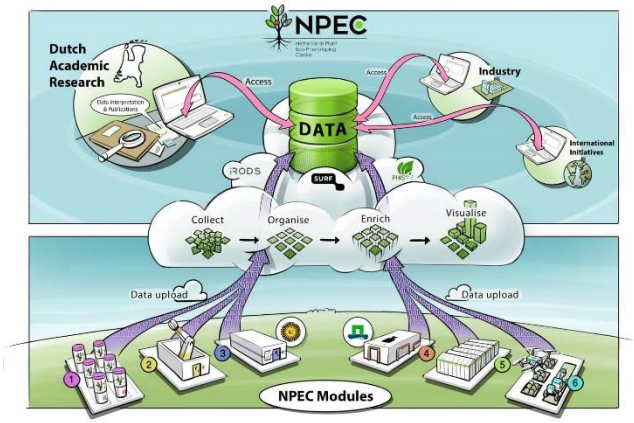

Figure 1: NPEC facility with the data flow from the 6 modules: **1** Ecotrons, **2** Plant-Microbe Interactions, **3** Multi-Environment climate chamber, **4** High-Throughput Phenotyping climate chamber, **5** Greenhouse and **6** Open-Field.

2. Approach

Recent developments in genomics allow us to create high quality reference genomes of for example different cultivars using long-read sequencing. Together with high throughput resequencing and targeted sequencing approaches produce large-scale genotypic datasets of up to thousands of individuals. Available compute infrastructure and approaches such as pangenomics[3] allows us to store and analyse these complex sets.

With the start of the Netherlands Plant Eco-phenotyping Center (Figure 1) we can now phenotype hundreds of plants in the greenhouse up to thousands in the field and climate rooms. Phenotypic information includes plant height, NDVI, greenness, biomass and leaf area, measured several times a day. Data are collected through platforms such as drones, sophisticated scales (PlantArray) and 3D imaging systems (PlantEye, Maxi-Marvin, Figure 2).

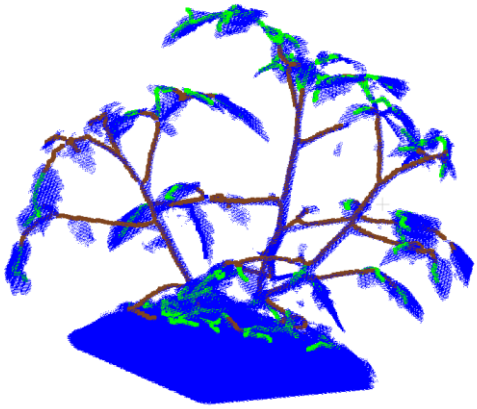

Figure 2: 3D model of a tomato plant from the Maxi-Marvin

3. Results & Discussion

Machine learning approaches, such as deep learning, generally require large amounts of data for training and validation[4]. Especially in data set with a high degree of variability as it is the case with plant genomics. Current developments in technologies for both genomics and phenotyping allow us to expand existing approaches as well as develop new ones to link phenotypes to genotypes. ML combined with pangenomics tools will present new ways of identifying, for example, presence/absence variation and multi-loci associations. Key traits we are interested in to investigate are usually highly variable on the genome level, such as disease resistance and species (in-)compatibility, very diverse in phenotype, such as environmental tolerances, or a combination of both.

References

1. Xiao, Q., Bai, X., Zhang, C. & He, Y. Advanced high-throughput plant phenotyping techniques for genome-wide association studies: A review. *J. Adv. Res.* **35**, 215–230 (2022).
2. Hu, Z. L., Park, C. A. & Reecy, J. M. Bringing the Animal QTLdb and CorrDB into the future: Meeting new challenges and providing updated services. *Nucleic Acids Res.* **50**, D956–D961 (2022).
3. Sheikhizadeh, S., Schranz, M. E., Akdel, M., de Ridder, D. & Smit, S. PanTools: representation, storage and exploration of pan-genomic data. *Bioinformatics* **32**, i487–i493 (2016).
4. Shi, W., van de Zedde, R., Jiang, H. & Kootstra, G. Plant-part segmentation using deep learning and multi-view vision. *Biosyst. Eng.* **187**, 81–95 (2019).