초청-01

Deep phenotyping pipeline: from phenotype definition to machine learning-based model interpretation

Sungyul Chang[1], Unseok Lee[1] and Hyoung Seok Kim[1]*

[1]Convergence Research Center for Smart Farm Solution, KIST Gangneung Institute of Natural Products, 679 Saimdang-ro, Gangneung, Gangwon 25451, Republic of Korea

[Introduction]

Recent advances in plant phenotyping suggest expanding our focus to utilization of digital data through the data translation and modeling for characterization of biological phenomenon. Here, we demonstrate that deep phenotyping pipeline, time course phenotyping of principle features representing dynamic behavior of plant, feature selection and data analysis with machine learning (ML) increased efficiencies of phenotyping data classification and prediction for geometric traits and different chemotypes in heterogeneous populations of rosette wild plant, *Lactuca denticulate* (Houtt.)Maxim.

[Materials and Methods]

Here, we present new plant phenotyping pipeline, consisted of following frameworks, definition of phenotypic features for target traits, time-course phenotyping over growing periods, model generation through training, model validation at independent experimental set, and selection of important features with interpretation of data to understand biological mechanisms. This pipeline has been applied for classification and characterization of three heterogeneous populations of *Lactuca denticulate* (Houtt.)Maxim [13]that undergoes adaption in physically separated locations. Random forest(RF) models successfully differentiated high-dimensional phenotyping data sets with the information on important features that allows how combination of morphological features in rosette can be associated with biomass accumulation and variations in chemotype.

[Results and Discussions]

Model composed of architectural features, selected as principle features by random forest(RF), are more effective to classify different heterogeneous populations in comparison to the model based on rosette growth indicators such as projected area, convex hull and perimeter. Deep phenotyping coupled with RF model allows prediction of the projected area and convex hull from principle features at organ level with high accuracy. Accumulated data through time-course phenotyping during continuous growth stages improve resolution of model interpretations, suggesting multi-dimensional data matrix helps to differentiate complex phenotyping data. Random forest model also successfully classify two chemotype groups, as same as the populations are separated to two clustering groups based on the pattern of chemical profiles by hierarchical method. This result indicates that geometricfeatures of rosette are associated with the differences in secondary metabolite profiles. Our model-assisted deep phenotyping approach provides insight into interpreting phenotyping data to use in agricultural and ecological studies as well as establishing phenotyping strategies through the identification of principle phenotypic features.

*Corresponding author: E-mail. hkim58@kist.re.kr