B-122

QTL analysis and phenotype prediction with Machine Learning in soybean

*Tetsuya Yamada**, Institute of Crop Science, National Agriculture and Food Research Organization, Ibaraki, Japan
Yohei Nanjo, Institute of Crop Science, National Agriculture and Food Research Organization, Ibaraki, Japan
Koji Takahashi, Institute of Crop Science, National Agriculture and Food Research Organization, Ibaraki, Japan
Motoki Takahashi, Institute of Crop Science, National Agriculture and Food Research Organization, Ibaraki, Japan

DNA marker assisted selection (MAS) is commonly used in soybean breeding in Japan. Many traits which can be supported by MAS are shattering-resistance, disease-resistance, stress-tolerance, cadmium-accumulation, protein-subunit, maturity, and so on. Although the conventional QTL analysis could estimate the locus of major QTL for these traits statistically, this method spends a lot of time and cost. And, because the conventional QTL analysis uses all samples for estimation of QTL regions, this method does not left samples for validating the accuracy of candidate QTLs. Thus, researchers should validate the accuracy with additional examination in the next year. These days, machine-learning techniques are widely used for solving problems. In this research, estimation of DNA markers in QTL region with popular machine-learning libraries was conducted.

RILs (n=152) derived from crossing between soybean cultivar “Tachinagaha” and breeding line “Tohoku129” were used for QTL analysis. QTL analysis was conducted for green stem disorder (GSD), days to first flowering and protein content with “qtl” package in “R” language. GSD is a trait indicating a gap of maturity between pod and stem.These traits were also examined with machine-learning library “Scikit-learn” in “Python” language. “Support vector machine regressor (SVR)”, “Random Forrest Regressor”, “KNeighbor Regressor”, “Lasso” and “Ridge” methods contained in “Scikit-learn” were used for prediction for the traits and compared for the accuracy among these methods. The accuracy was the highest with “Lasso”, and the second with “SVR”. Estimated major QTL regions were corresponding to those estimated by conventional QTL analysis. These machine-learning models could estimate major QTL for GSD and other phenotypes from rough genotype of whole genome. And prediction of phenotypes of RILs by these models could validate the existence of the QTLs at the same time. Therefore, these methods might complement or alternate the conventional QTL analysis.