

Genomic prediction in plants using machine learning

Kozlov K.*, Bankin M., Samsonova M.

Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia

*kozlov_kn@spbstu.ru

Key words: bread wheat; rye; chickpea; genomic prediction; machine learning

Motivation and Aim: Today, genomic selection (GS) has become an established methodology in many plant breeding programs with the main application of reducing the length of breeding cycles [1]. Many genomic prediction (GP) methods have been proposed as there is no universal best method that can be used under all circumstances. The accuracy of GP depends on a series of factors including the quality and pre-processing of the phenotypic data, the platform used to obtain genomic information, the population mating design, the intrinsic genetic architecture of the trait, the genetic structure of the population, how the genotype-by-environment interaction is dealt with, and prediction method [2]. Nonparametric machine learning methods, e. g. Support Vector Machine, Random Forest, Gradient Boosting Machine provide tremendous flexibility to adapt to complicated associations between data and output [3]. Recently convolutional neural networks (CNN) have been intensively studied and applied in genome-based breeding [4]. These methods automatically identify latent patterns or features from data by sequentially stacking several layers. CNN are widely used in image classification [5]. Large genomics data can be encoded as artificial image objects (AIO) by considering individual genetic variants as pixels [6]. Consequently, AIO could be used by CNN on regression and classification tasks [7]. Here we used three datasets of bread wheat, rye and chickpea to compare the performance of several parametric and nonparametric methods in the regression task, i. e. in predicting phenotype from genomic data.

Methods and Algorithms: The rye and bread wheat datasets were generated by the Russian Bread consortium. Rye dataset consisted of 35208 SNPs, which were called in 451 samples phenotyped for 29 different morphological, productivity and biotic stress traits. The bread wheat dataset contained 16432 SNPs called in 960 samples, which were phenotyped for 7 traits. Several GP models were constructed for rye dataset namely BLUP, Bayesian regression with and without kinship matrix or covariance matrix, Extreme Gradient Boosting with and without two principal components. For bread wheat dataset GP models were build using BLUP, Bayesian regression with and without kinship matrix, as well as Extreme Gradient Boosting with two principal components and kinship matrix. Each model was constructed using 90 % of the dataset for training with the 4-fold cross-validation method. The rest of the samples comprising 10 % of the dataset were used to estimate the model accuracy by calculating the Pearson correlation between data and model prediction. All the models were trained for 30 iterations using various combinations of hyperparameters, with 10 different random partitions explored for each combination.

The combination of CNN for feature extraction and Extreme Gradient Boosting for regression was applied to model two main productivity traits of chickpea, 1000 seed weight (TSW) and number of seeds per plant (SNpP). Details on the phenotyping,

genotyping and subsequent analysis were presented in [8]. SNPs were encoded AIO and fed into CNN. Each convolutional kernel slides along the corresponding image channel and the results of processing are combined into one feature map. The size of the filter kernel of each convolutional layer was tuned using a small subset of the available experimental data and the kernel weights are the learning parameters of the model. Each convolutional layer is followed by a subsampling layer, which reduces the dimension of maps in order to enlarge features. Such filtering helps, among other things, to avoid overfitting. The formation of a new feature map is based on the max pooling operation, which is performed by selecting the maximum value from a subsample of a given size. The contribution of individual SNP to the model prediction was estimated by the feature attention map calculated as the weighted sum of the convolution layers of the network.

Results: Genomic prediction models were built for 15 rye traits. For some traits, e. g. Width of the third subflag leaf, several models resulted in accurate predictions, while for others (e. g., Weight of 1000 seeds) only one model was accurate. The maximum accuracy equal to 88 % was attained with the Extreme Gradient Boosting method for the Length of the third subflag leaf.

Predictive models were constructed for all wheat traits. The accuracy of the models for Brown rust, Yellow rust, Heading, Booting, Plant height and Yield exceeded 50 %, the best accuracy of 72.5 % was attained for the Yellow rust trait.

The accuracy of genomic prediction models for the TSW and SNpP traits of chickpea was 85 % and 84 %, respectively. Functional characterization of 200 SNPs with the highest contribution to the TSW trait prediction showed that 91 of them are within known genes involved in general energy metabolism, photosynthesis, abiotic stress response, as well as in seed formation and quality. Likewise, 61 of 200 SNPs with the highest contribution to the SNpP trait prediction were within 50 known genes. 31 SNPs were found in the non-coding region of genes (introns) and 30 in the coding region (exons). Interestingly, 3 of these SNPs were found in gene encoding boron transporter 2. Boron plays essential roles in formation and stability of the plant cell wall through the cross-linking of the rhamnogalacturonan II [9].

Conclusion: The non-parametric methods yielded superior results than parametric methods across all traits and datasets. The best accuracy of 88 % was achieved for the Length of third subflag leaf on the rye dataset using the Extreme Gradient Boosting model. Accurate prediction models can be generated by using CNN in combination with AIO for feature extraction, however further studies are necessary to provide new valuable and unbiased information for assessing their methodological potential for GP.

Funding: The authors thank the Russian Bread Consortium, which provided access to datasets on rye and bread wheat. The study is supported by the Russian Science Fund (Grant No. 22-46-02004).

References

1. Meuwissen T.H.E., Hayes B.J., Goddard M.E. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*. 2001;157:1819-1829. doi 10.1093/genetics/157.4.1819
2. de Los Campos G., Hickey J.M., Pong-Wong R., Daetwyler H.D., Calus M.P. Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics*. 2013;193(2):327-345. doi 10.1534/genetics.112.143313
3. Montesinos-López O.A., Montesinos-López A., Mosqueda-Gonzalez B.A. et al. A zeroaltered Poisson random forest model for genomic-enabled prediction. *G3 (Bethesda)*. 2021;11(2):jkaa057. doi 10.1093/g3journal/jkaa057
4. Montesinos-López O.A., Montesinos-López A., Pérez-Rodríguez P. et al. A review of deep learning applications for genomic selection. *BMC Genomics*. 2021;22(1):19. doi 10.1186/s12864-020-07319-x

5. Pook T., Freudenthal J., Korte A., Simianer H. Using Local Convolutional Neural Networks for Genomic Prediction. *Front Genet.* 2020;11:561497. doi 10.3389/fgene.2020.561497
6. Galli G., Sabadin F., Yassue R.M. et al. Automated Machine Learning: A Case Study of Genomic “Image-Based” Prediction in Maize Hybrids. *Front Plant Sci.* 2022;13:845524. doi 10.3389/fpls.2022.845524
7. Bavykina M. et al. Modeling of Flowering Time in *Vigna radiata* with Artificial Image Objects, Convolutional Neural Network and Random Forest. *Plants.* 2022;11:3327
8. Sokolkova A., Bulyntsev S.V., Chang P.L. et al. Genomic Analysis of Vavilov’s Historic Chickpea Landraces Reveals Footprints of Environmental and Human Selection. *Int J Mol Sci.* 2020;21:3952
9. Saouros S., Mohan T.C., Cecchetti C. et al. Structural and functional insights into the mechanism of action of plant borate transporters. *Sci Rep.* 2021;11:12328. doi org/10.1038/s41598-021-91763-6