Data Science for Genome Based Optimization in Agriculture: EU Research Data Alliance and Biopotential of Croatia Cultivars

Kurtanjek, Ž.

Croatian Academy of Engineering
zelimir.kurtanjek@gmail.com,

Abstract: The development of high-throughput sequencing methodologies resulted in omic big data sets in biotechnology (pharmaceuticals, agriculture, and food technology) reflecting information on biological potential of large number of varieties of worldwide importance. Omic databases specific to Croatian cultivars, autochthonic plants and animal varieties are also being developed. Comparative advantages and opportunities for the improvement of Croatian cultivars and species can be advanced by inclusion and systemic comparative analysis within EU project RDA (Research Data Alliance) for open data access to H2020 research projects. The bottleneck for the application and discovery of optimal biopotential from big data is the use of advanced machine learning (ML) algorithms supported by thorough validation criteria. The ML algorithms (elastic nets, boosted decision tree forests, and deep learning are applied here for Diversity Array Technology (DArT) genotypization for production of *Triticum aestivum* (wheat bread making cultivars).

Keywords: Research Data Alliance, Boosted decision trees, Elastic nets, DArT genotypization, *Triticum aestivum*

Introduction

The world is facing a number of interrelated global factors which present a risk for its sustainability. Two of the factors are exponentially growing population, presently estimated at seven billion and predicted to ten billion in 2050, and the global climate warming with an estimated average temperature rise of 2.5 ^{0}C . These global risks directly affect food production which is the main economic and political stability factor. The key underlying factor is availability of water needed for agriculture which is the primary input to food chain. Food production must be adapted

to global warming, lack of water for agriculture and animal protein productions, and increased levels of soil and water pollution. Croatia also faces the same problems related to climate changes affecting agricultural production and over a long period of time increased temperatures and extended periods of drought weather (Fig. 1). These challenges must be addressed with potentials of molecular based life sciences entering a new era of data science with big data (petabyte) of sequencing data and advances in new machine learning algorithms based on the broad availability of computer clusters (Hadoop) with Tensor low (TPU) processing computer architecture. Croatia needs to explore these advanced technologies and develop new potentials of autochthonic plant and animal varieties for higher yield and quality (wheat, corn, olive, pork, tangerines, fish farming, marasca cherries, etc.). It is especially interesting to explore the biopotential (bioprospecting) Croatian Adriatic Sea and rare autochthonic plants for pharmaceutical production.

Fig. 1 – Distribution of global aridity land, A: 2011, B: prediction 2050, (Dai. A., 2011, J. Farrant, 2016)

High-throughput sequencing technology has brought life science into a "big data" era with an unrivalled explosion of genomics, transcriptomics, proteomics, and metabolomics. The falling cost (<$1,000 per genome) and increasing speed (<1 day per genome) of high throughput sequencing lead to the snowballing data at petabyte level. However, it is still difficult to transfigure such "Big Data" to valuable biological insights into regulation of metabolic pathway activities. The gap between omics data and cell phenotypes is one of the biggest challenges for achieving "Data-to-Insight". In recent years, a rapid development of artificial intelligence, especially deep learning, provides novel options to overcome this challenge.

The key factor to cope with challenges by harnessing global biological potentials is large scale open data integration. For Croatia data science projects are an important integration into EU project Research Data Alliance (RDA). It is an open access structured database divided into working groups. The working groups IG Agricultural data and WG Wheat data Interop (Hologne O., 2017) are applicable for improvement of Croatian cultivars. The RDA main objectives are: coordination of worldwide research effort to build research infrastructure for wheat genomics, physiology, breeding, and computer science experts in data science and bioinformatics.

Materials and Methods

Molecular markers are technologies which represent invaluable research tool for understanding the genetic control of various traits. They have frequently been utilised in quantitative trait loci (QTL) mapping studies, and applied in breeding programmes through marker-assisted selection. In Croatia there are two scientific institutions, Agricultural Institute of Osijek and Centre of Excellence for Biodiversity and Molecular Plant Breeding in Zagreb, leading in research to improve existing cultivars and by bioprospecting to harness Croatia biopotential. The application of molecular markers resulted in genome associated wide studies (GWAS) as a part of big data science in agriculture. Diversity Array Technology (DaRT) (Jaccoud D, et al., 2001) is commonly molecular marker technology applied in molecular agriculture studies. Due to global aspects of these studies open access databases (RDA) with experimental data accessible for analysis were created. Here are the applied data available from International Maize and Wheat Improvement Centre (CIMMTY, 2017). A similar study for Croatian selected winter wheat cultivars is available (Novoselović D, et al, 2016). The dimension of the applied big data set is 1276x599x8 (DaRT markers x number of wheat breeding lines x number of phenotypes) containing in total about 6 million data points.

The objective of this study is to apply different algorithms of big data analytics to develop predictive models for each of wheat phenotypes, to determine individual DaRT marker importance, and by use of computer simulation assist breeding programs for optimization of new varieties with improved phenotype properties. The applied methodologies are elastic nets (Efron and Hastie, 2016), neural networks (Chollet and Allaire, 2018), and decision trees (Kurtanjek, 2016, 2017). Algorithms for data science available in R are applied (R Core Team, 2018). The main keys of the applied algorithms are feature (DaRT) space regularization and extensive (multifold) bootstrap validation. The elastic nets are linear regression models with use of combined L_1 and L_2 penalty functions to reduce the dimension of the model linear parameter space by systemic testing. The standard linear model, with x denoting DaRT markers is given by

$$\hat{y} = \sum_{i=0}^{i=N} \beta_i \cdot x^{i-1} + E \tag{1}$$

with the objective function

$$F_{object}(\beta) = SSE(y - \hat{y}) + \lambda_1 \cdot L_1 + \lambda_2 \cdot L_2 \tag{2}$$

Neural networks and decision trees are the principal nonlinear models applied in big data science. Deep learning neural networks harness potential of multilayer structure and big data for regularization of inference prediction space,

$$\hat{y} = NN(x) + E \tag{3}$$

while decision trees apply averaging of large random ensemble decisions

$$\hat{y} = \frac{1}{N} \cdot \sum_{i=1}^{i=N} DT_i(x) \; + E \tag{4}$$

Elucidation of decision rules is improved by regularization accounting for ensemble dimension N_T, size of the trees N and purity of the tree leafs

$$\min F_{object.}(X, DTF, \omega) = SSE(y, \hat{y}) + \phi(DTF, \omega) \tag{5}$$

$$\phi(DTF, \omega) = \gamma \cdot N_T + \frac{1}{2} \cdot \lambda \cdot \sum_{i}^{N_T} \omega_i^2 \tag{6}$$

Bootstrapping and optimization based analytical evaluation of gradients are the most important powers of the boosted decision tree forests.

Results and Discussion

DArT profiles of samples (breeding lines) are binary records (0 and 1) corresponding to the presence of marker diversity (SNP and/or methylation) referenced to a standard. A sample of such a record of 1,279 markers is shown in Fig. 2. Most of the profiles are mutually poorly correlated, with the average absolute value of R = 0.1. However, there is a significant correlation between subsets of the profiles and particular phenotype properties. The dimensions of the DArT subspaces determined by the elastic nets for the phenotypes protein content and area grain yield (t/ha) are presented in Fig. 3.

The corresponding sub-dimensions are in the ranges 82-215 and 60-150. Relatively large dimensions of the phenotype subspaces are reflections of complexity of gene

Fig. 2 – A sample distribution of DArT 1279 markers of a single wheat breeding line accession. Markers with diversities are depicted as white lines.

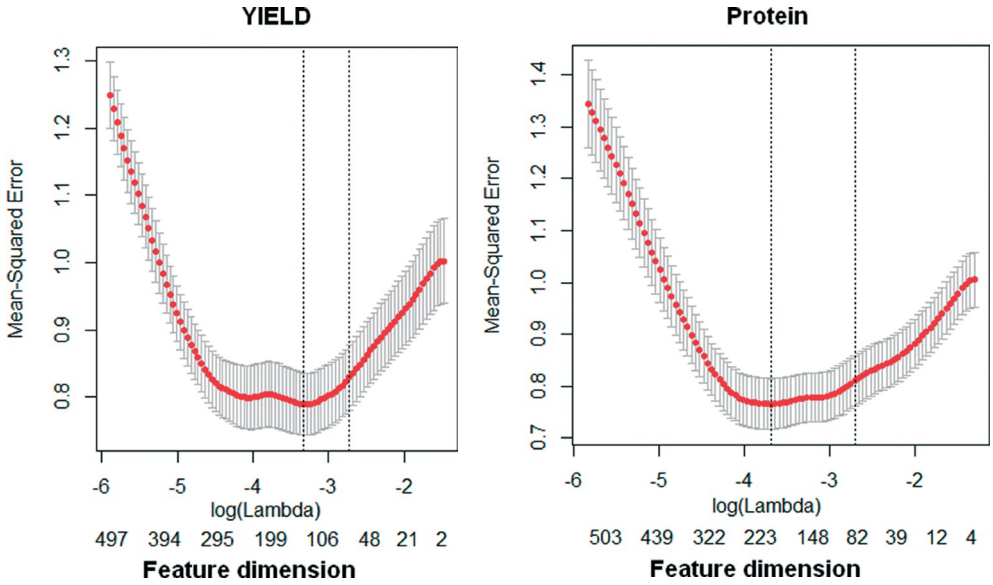

Fig. 3 – Validations of dimensions DArT feature spaces by elastic nets.

interactions responsible for individual property. The obtained correlation coefficients for the elastic net models for the prediction of protein content and yield are R=0.8 and R=0.72 respectively. The unexplained variance by the elastic nets is due to nonlinear interactions (gene level synergism). Considerable improvements for phenotype predictions, accounting for epigenesis, are obtained by multilayer nonlinear interactions, embedded into decision trees and/or deep learning models. Pre-

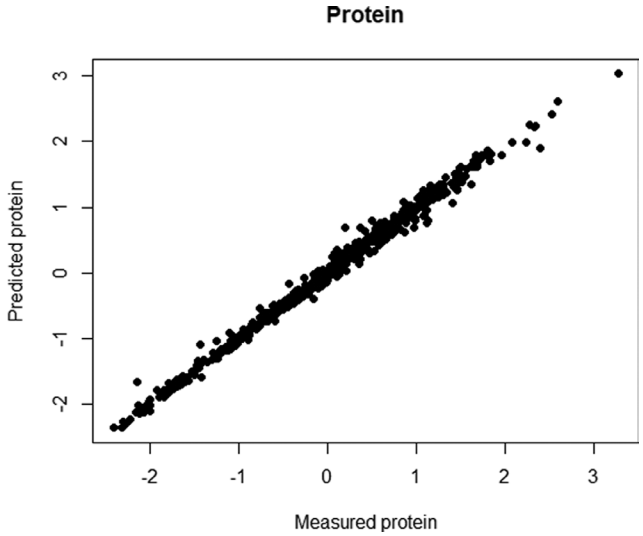

Fig. 4 – Correlation line between DecisionGS model for wheat protein predictions and experimental data.

diction accuracy is increased to 92 % of accounted variance (determination factor). High accuracy of decision trees model (decision tree genetic selection algorithm, abbreviated as DecisionGS) for protein content is depicted in Fig. 4. High accuracy for specific wheat property prediction models enables their application for computer simulation and optimization of breeding process. A simulation program for random exchange of gene (DArT) exchange between parent breeding lines and corresponding prediction of progeny phenotype was developed.

Fig. 5 shows an example of simulated breeding and optimization of protein production by area. As the first parent is selected the wheat breeding line with high protein content but low productivity, and as the second parent the line with low protein content but high yield. DArT correlation between the parents is R = 0.84. 100 randomly generated progenies with corresponding distribution of the optimization objective to gain maximum of protein production per unit are simulated. The median of the distribution is about 5 % higher compared to each of the parent breeding pool. The optimal progeny is shown in Fig. 5 with 15 % increase in the protein yield per area.

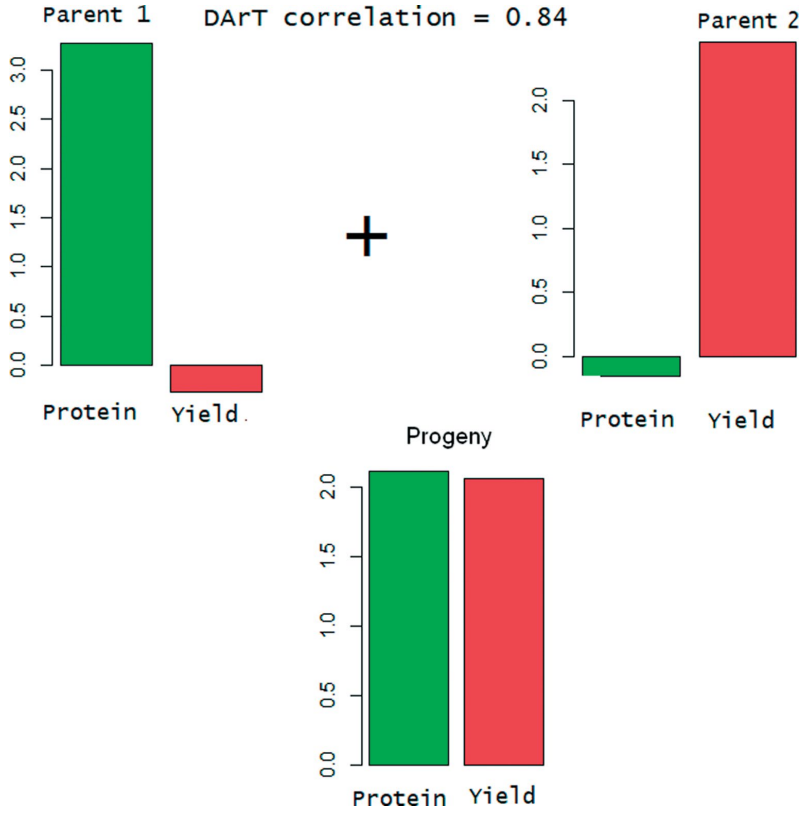

Fig. 5 – Prediction of wheat breeding progeny with optimal protein content and grain yield per area (t/ha).

Conclusions

The application of data science and big data omics studies of the Croatian biopotential are the strategic technologies needed to adapt to global risks due to population increase and global warming. The objectives should be an increase in the national production and food quality with emphasis on specific biopotential of plant and animal varieties in Croatia.

The application of the advanced models and algorithms from data science is essential to harness knowledge for the optimization of production from omic big data sets. The integration and open access of big data, such as the project EU Research Data Alliance, is the key factor for excellence in science projects.

The optimization of wheat production based on computer assisted genomic simulation of breeding improvement is being developed here. A big data set is applied from international science cooperation projects on DArT genotypization to derive elastic nets and boosted decision trees models for prediction of specific wheat phenotypes. The derived DecisionGS model predicts protein content and yield with an accuracy of 92 %. The high accuracy of the model enabled the simulation and optimization of a breeding program resulting in 15 % increase of protein production per area.

Acknowledgements

I would like to thank to Valentina Španić and Daniela Horvat, the members of Agriculture Institute, Osijek, Croatia, for cooperation and valuable assistance.

References

Dai A., (2011). Drought under global warming: a review, WIREs Climate Change , Vol. 2,(1), 45-65.

Farrant J., (2016), https://archive.org/details/JillFarrant_2015G ,(last accessed: March 5, 2018)

Hologne O., (2017), RDA Working and Interest groups Perspective from a participant, Source http://www.inra.fr (last accessed: March 5, 2018)

CIMMYTY (2017) Source http://www.cimmyt.org/

Jaccoud D., Peng K., Feinstein D., Kilian A. (2001), "Diversity Arrays: a solid state technology for sequence information independent genotyping", Nucleic Acids Research, 25(4) 25.

Novoselović D., Bentley A.R,, Šimek R,, Dvojković K.,Sorrells M.E, Gosman N., Horsnell R.,, Drezner G,, Šatovič Z. (2016),"Characterizing Croatian Wheat Germplasm Diversity and Structure in a European Context by DArT Markers", *Front. Plant Sci. 7:184.*

Efron B., Hastie T. (2016), „*Computer Age Statistical Inference – Algorithms, Evidence, and Data Science*“, Cambridge University Press, New York,

Chollet F., Allaire, (2018), “Deep Learning R”, Manning, New York, USA

Kurtanjek Ž., (2016)„Systems Analysis of Ensemble of Decision Trees for Modeling and Process Control“, Proceedings AIChE Annual metting, San Francisco, USA

Kurtanjek Ž. (2017), „Big data analytics in food technology“, Food Chemistry and Technology, Baltimore, MA, USA

Kurtanjek Ž. (2017), „Genome wide big data analytics: case of yeast stress signature detection“, *The EuroBiotechnology Journal*, 1(4) 264-270

R Core Team. R: A language and environment for statisticalcomputing. R Foundation for Statistical Computing, Vienna. Austria, Source https://www.R-project.org (last accessed: March 5, 2018).