Dynamic metabolomic prediction based on genetic variation for *Hordeum vulgare*

P. Nemčeková[1] and J. Schwarzerová[1,2]

[1]Department of Biomedical Engineering, Faculty of Electrical Engineering and Communication, Brno University of Technology, Brno, Czech republic

[2]Molecular Systems Biology (MOSYS), University of Vienna, Vienna, Austria

E-mail: xnemce05@vutbr.cz, Jana.Schwarzerova@vut.cz

Abstract—*Hordeum vulgare*, like many other crops, suffers from the reduction of genetic diversity caused by climate changes. Therefore, it is necessary to improve the performance of its breeding. Nowadays, the area of interest in current research focuses on indirect selection methods based on computational prediction modeling. This study deals with dynamic metabolomic prediction based on genomic data consisting of 33,005 single nucleotide polymorphisms. Metabolomic data include 128 metabolites belonging to 25 Halle exotic barley families. The main goal of this study is creating dynamic metabolomic predictions using different approaches chosen upon various publications. Our created models will be helpful for the prediction of phenotype or for revealing important traits of *Hordeum vulgare*.

Keywords— Machine learning, Single nucleotide polymorphism, genomic prediction, *Hordeum vulgare*

1. INTRODUCTION

Genomic prediction is one of the best computational approaches for acquisition of information that is important to breed crops [1]. Genomic predictions are also one of the best methods for selection of the most perspective plants for breeding or for optimization of growing healthy plants [2]. Thanks to that, plants can grow faster with reduced usage of water. Thus, this methodology has a huge potential in ecology and biotechnology. In addition, the same methodology can be useful for early disease detection in biomedicine.

Currently, research focuses on corss-linking genotypes and phenotypes. The main direction of interest focuses on metabolomics with metabolite concentration in the spotlight. This is due to the fact, that metabolites represent the response of biological processes in organisms. The genotype-metabolotype association is sensitive to environmental changes [3]. The first step in current research leads to creating and improving techniques that will have a potential to describe the relationship between genome and metabolome.

Halle Exotic Barley (HEB) population [4] was developed using nested association mapping (NAM), where exotic alleles were placed into a genetic background of German spring barley. Genotyping with the Infinium iSelect 9k SNP chip was used for obtaining the genomic information [5], that was used in this study. The metabolomic data taken from study by Gemmer et al. [6], were processed using MassHunter Qualitative Analysis software. This resulted in data for 1,307 lines where 158 metabolites (METs) were defined. In this study, we focused on the prediction of metabolite concentration related to the HEB family based on genomic variance represented as single nucleotide polymorphism (SNP). This study uses genomic and metabolomic dataset of HEB population mentioned above and brings three models for metabolomic prediction based on genetic variation.

2. DATASET

Hordeum vulgare [7] has wide usage in the food industry as it is the 4th most important crop in the world. According to climate changes, the genetic diversity of barley has been reduced. The genomic dataset was taken from the study by Maurer et al. [8]. The study by Gemmer et al. [6] presented metabolomic information which is related to the above mentioned genomic dataset. Thus, these metabolite data were

used in our study, too.

Genotyping was performed by 50k Illumina Infinium iSelect SNP Array [8]. Genomic data include 1,363 measures of *Hordeum vulgare* nested association mapping population HEB-25. In metabolomic data, 1,361 concentrations of 1,419 were connected to one of the HEB families, which were analyzed in our study.

3. METHODS

3.1. Pre-processing

Firstly, data were reduced by a number of HEB lines and united by accession number for genomic and metabolomic values. The final SNP table includes 3,877 SNP markers, 13 metabolites and 1,307 HEB lines. The genomic dataset was transformed from scaling $\langle\ 0,\ 2\ \rangle$ to $\langle\ -1,\ 1\ \rangle$ for optimization of the input for predictions.

3.2. Dynamic metabolomic prediction

In general, the metabolomic prediction based on genetic variation used static genomic information such as SNP table. This study focused on the dynamic metabolomic prediction using three different prediction approaches, namely: support vector regressor (SVR), least absolute shrinkage and selection operator (LASSO) and sparse partial least squares regression (sPLS), chosen based on publications [9, 10].

Each of the prediction method was implemented for a different setting of parameters. The different rate of training and testing data in ratio from 9:1 to 6:4 was applied. In this study, parameter alpha in LASSO model is set to 0.3.

sPLS is a method based on covariance that combines spectral decomposition technique and multiple regression analysis. The main principle of this method relies on searching for linear combinations of variables. In this study, the number of components of the sPLS model is set to 10. Due to that, the dimensionality of the data is reduced.

SVR uses a non-linear radial basis function and a polynomial kernel function. The principle of SVR consists of transforming data into a hyperplane represented by support vectors. In this study, the regularization parameter C is set to 1 and the ε-insensitive region is set to 0.1. All created and used scripts are available on Github: 'PetraNemcekova/Genomic_Prediction'.

4. RESULTS AND DISCUSSION

The study brings three prediction models based on SNP using three different approaches. Evaluation parameters represent mean absolute error (MAE) and mean squared error (MSE) [11].

Summary of results of all approaches using SVR, LASSO and sPLS are shown in Table I. The best predicting SVR model is trained on 70% of the training dataset. The MAE of the SVR model is 0.188 and MSE is 0.058.

Table I: Summary of results of trained prediction models using SVR, LASSO and sPLS with different percentage of data used for training.

	SVR		**LASSO**		**sPLS**	
	MAE	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**
10% test	0.198	0.058	0.206	0.058	0.413	5.743
20% test	0.203	0.107	0.202	0.069	0.546	0.851
30% test	0.188	0.058	0.272	1.996	0.511	0.576
40% test	0.198	0.074	0.256	1.513	0.279	0.576

The best evaluation parameters of the SVR model is connected to metabolite MET25, MAE is 0.067 and MSE is 0.005. The worst predicted results, MAE is 0.198 and MSE is 0.088, were found for disaccharide MET122. SVR is an appropriate regression model for metabolomic prediction based on genomic variation.

The best evaluation parameters of LASSO models are 0.206 MAE and 0.058 MSE. The standard deviations

(SD) are 0.089 and 0.025. The minimal MAE is associated with the metabolite MET39, the minimal MSE is connected to metabolite MET88. LASSO was proved to be a suitable regression model for metabolomic prediction based on SNP.

In sPLS models, the best division of training and testing data is ratio 7:3 according to obtained values of MAE and MSE. The best sPLS model has 0.511 MAE with 0.933 SD and MSE is 0.576 with 17.767 SD.

The minimal error in the sPLS model is connected to the metabolite MET88, with the MAE being 0.017. According to MSE, the best-predicted evaluation is associated with MET39 with the value of MSE being 0.0007.

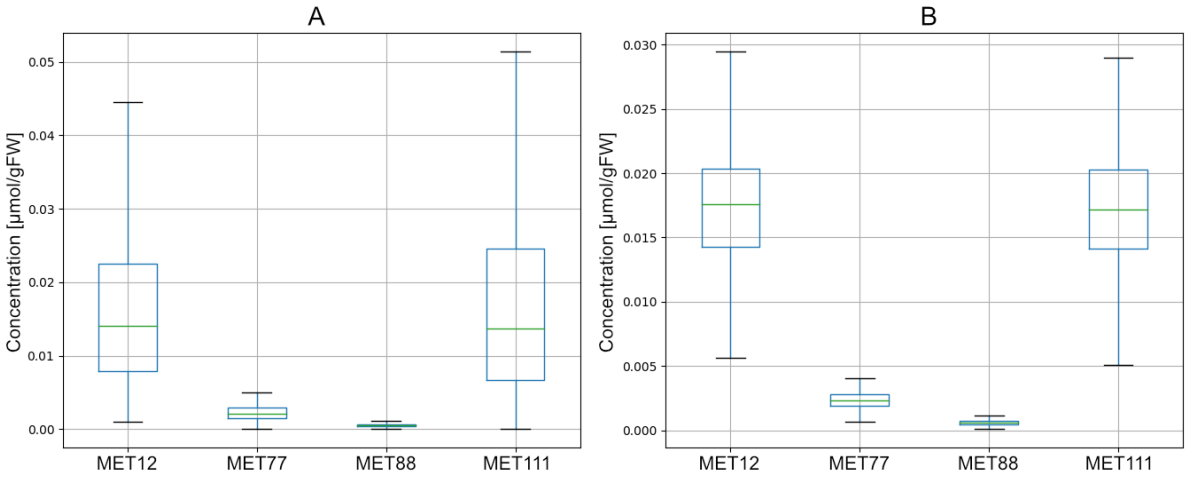

Figure 1: Box plots of original and predicted values by sPLS model. A section shows boxplots of original values and B section represents boxplots of predicted values using sPLS. The boxplots show the statistics of ascorbic acid MET12, ribonic acid MET77, hydroxylase inhibitor MET88 and transthyretin MET111.

Figure 1 shows boxplots of original and predicted data using the sPLS method of 4 randomly chosen metabolites. The values in these boxplots are visualized in the same numerical order. Because of that, evaluation parameters such as MAE and MSE are sufficient for the description of models.

The model with the best evaluation parameters was created using SVR. Thus, this model was selected and modeled using cross-validation with 5 repeats and the number of splits 5. Thanks to that, we obtained the final model with better accuracy. The evaluation parameters of 5-cross-validation have the value of MAE 0.150 and the MSE 7.608.

5. CONCLUSION

Nowadays, ecology research of plants has huge potential due to the combination of laboratory analysis and new computational algorithms. Thanks to that, we can open new way of possibilities based on prediction modeling. This can help with prediction of disease phenotypes or with selection of the most perspective plants in an early development of the crops. One of these major ways is a dynamic metabolomic prediction based on genetic variation. The reason for this is that especially linking of genotype and metabolite is sensitive to environmental changes such as climate changes. This study focused on dynamic metabolomic prediction based on genetic variation in *Hordeum vulgare*. It is a representative of crops which have a main role in the food industry and it is necessary to improve the performance of its breeding. The main goals of our study are creating and comparing prediction models that can be used for describing the relationship between genome and metabolome. These models can be used for the prediction of phenotype or revelation of the important traits of *Hordeum vulgare* using the genotype-metabolotype association. In total, we modeled three prediction models based on three different modeling approaches such as LASSO, sPLS and SVR. All models were evaluated using MSE and MAE. The model that connected to the best evaluation parameters was created using the SVR method. A cross-validation was used to improve this model. This study brings three models that can be used for prediction of metabolomic information based on genetic variation. Thanks to that, this study partly fills the blank space in solving this problematics and opens new opportunities for predicting the genotype-phenotype association. The

whole methodology was implemented in Python and is available on Github: 'PetraNemcekova/Genomic _Prediction'.

ACKNOWLEDGMENT

This work has been supported by grant FEKT-K-21-6878 realised within the project Quality Internal Grants of BUT (KInG BUT), Reg. No. CZ.02.2.69 / 0.0 / 0.0 / 19_073 / 0016948, which is financed from the OP RDE.

REFERENCES

[1] L. M.Zingaretti et al, "Exploring Deep Learning for Complex Trait Genomic Prediction in Polyploid Outcrossing Species", *Front. Plant Sci.*, 2020, doi 10.3389/fpls.2020.00025.

[2] M. E. Goddard, B. J. Hayes, "Genomic selection", *J Anim Breed Genet*, 2007, doi 10.1111/j.1439-0388.2007.00702.x.

[3] A.-M. Waldvogel et al., "Climate Change Genomics Calls for Standardized Data Reporting", *Front. Ecol. Evol.*, 2020, doi 10.3389/fevo.2020.00242.

[4] P. Herzig et al., "Contrasting genetic regulation of plant development in wild barley grown in two European environments revealed by nested association mapping", *Journal of Experimental Botany*, 2018, doi 10.1093/jxb/ery002.

[5] A. Maurer et al, "Modelling the genetic architecture of flowering time control in barley through nested association mapping", *BMC Genomics*, 2015, doi 10.1186/s12864-015-1459-7.

[6] M. R. Gemmer et al, "Can metabolic prediction be an alternative to genomic prediction in barley?", *PLoS One*, 2020, doi 10.1371/journal.pone.0234052.

[7] S. Abbo et al, "Agricultural Origins: Centers and Noncenters; A Near Eastern Reappraisal", *Critical Reviews in Plant Sciences*, vol. 29, no. 5, p. 317-328, 2010, doi 10.1080/07352689.2010.502823.

[8] A. Maurer et al., "50k Illumina Infinium iSelect SNP Array data for the wild barley NAM population HEB-25", *e!DAL*, 2019, doi 10.5447/ipk/2019/20.

[9] S. Wang et al., "Identification of optimal prediction models using multi-omic data for selecting hybrid rice", *Heredity*, 2019, doi 10.1038/s41437-019-0210-6.

[10] C. Du et al., "Genomic selection using principal component regression", *Heredity*, 2018, doi 10.1038/s41437-018-0078-x.

[11] R. Sun et al., "Prediction of Liver Weight Recovery by an Integrated Metabolomics and Machine Learning Approach After 2/3 Partial Hepatectomy", *Front Pharmacol*, 2021, doi 10.3389/fphar.2021.760474.