Application of support vector regression for prediction of Grey Leaf Spot resistance using high density molecular data.

ORNELLA, L.[(1)]; PEREZ, P.[(2)];TAPIA, E.[(1)] ;CROSSA, L.[(2)]

(1)Rosario, Argentina.
Centro Internacional Franco Argentino de
Ciencias de la Información y de Sistemas (CIFASIS);

(2)Texcoco, Mexico.
Centro Internacional de Mejoramiento
de Maíz y Trigo (CIMMYT);

Grey leaf spot (GLS) is a serious yield-reducing disease of maize (*Zea mays*) in many parts of the world. The causal organism associated with the disease is *Cercospora zeae-maydis*. Pandemic in Africa, GLS is now recognized as one of the most significant yield-limiting diseases of maize worldwide (Ward et al., Plant Sci. 83: 884-895). If multigenic, the nature of a resistance impedes the use of MAS (marked assisted selection) for the generation of reliable, resistant genotypes. A valuable alternative is Genomic selection (GS), because GS estimates breeding values of individuals using the sum of all marker effects, many limitations associated with detecting significant marker trait associations are bypassed and the total genetic variation for the traits of interest can be better captured by the markers (Heffner et al., Cro Sci. 50:1681-1690) .

Prediction of genetic values can be carried out using parametric or nonparametric approaches. Parametric models, such as BayesA (Meuwissen et al, Genetics 157: 1819-1829) , Bayesian Lasso (Crossa et al., Genetics 186: 713-724) or Ridge Regression (Piepho, Crop Sci. 49: 1165-1176) are the most commonly used. However, they are not flexible enough to incorporate complex gene action (e.g., dominance or epistasis). Support vector machines are considered a state-of-the-art machine learning algorithms for classification and regression (SVR). Due to their theoretical foundations, SVR is very suitable for GS applications. Besides, unlike parametric models, no strict assumption is made regarding the form of the genotype–phenotype relationship. Rather, this relationship is driven primarily by the data.

Long et al. (TAG 123:1065-74) proposed ε-insensitive SVR (ε-SVR) for predict GEBVs (genomic estimated breeding values) in wheat. Previously, Maenhout et al. (TAG 115:1003–1013) used ε-SVR for predict maize hybrid performance. Both utilized information from less 2000 to make the prediction. In this work we proposed SVR to predict GLS but using high density marker panels. SNPs (Single Nucleotide Polymorphism) data from Illumina 55K chip and GLS susceptibility information of a set of 300 inbreds determined in 4 different trials were used for train SVR with linear and radial basis function kernels. GLS traits were analyzed using an ordinal scale from 1 (no disease) to 5

(complete infection). The Box-Cox transformation was applied to the original trait to make their distribution more symmetric.

Regression, using two alternative kernels (linear and radial basis function), was performed using the RegSMOImproved class of the WEKA Workbench. Parameters of the algorithms : C (linear kernel) and C, γ (RBF kernel) were optimized by logarithmic grid search of base 2 over a extensive range of values, each point of the grid was evaluated by internal 5-CV cross-validation on the training set using Pearson's correlation as a criterion for success. Epsilon parameter was not optimized (default value = 0.001).

Evaluation of predictive capability was performed by means of repeated hold out (Kim, Comput stat data anal 53: 3735:3745). Data was randomly partitioned into a training set (90 % of the complete set of lines) and a test set (10 % of the set). This was repeated 50 times by sampling lines at random. In order to facilitate the comparison between kernels, we used the same partitions for both linear and RBF kernel.

Success of prediction was measure using Person's Correlation coefficient between observed and predicted values. We also evaluated the predictive mean square error (PMSE). Results are presented in Table 1.

Comparison of performance was statistically validated using the non-parametric Wilcoxon (one tail) matched pairs test, which does not depend on normality assumptions (W.W Daniel, Applied Nonparametric Statistics).

Table 1-Predictive performance, determined as Pearson's correlation coefficient and PMSE, of Support vector regression with radial basis function kernel (SVR-RBF) or linear kernel (SVR-Lin) for GLS resistance on 4 maize datasets using high density molecular data.

Data	No. of lines	No. of markers	Correlation			PMSE		
			SVR-Lin	SVR-RBF	p-value Wilcoxon*	SVR-Lin	SVR-RBF	p-value Wilcoxon*
GLS1	272	46374	0.1969	0.2097	0.3677	1.2855	0.9809	<0.001
GLS2	280	46374	0.4044	0.4252	0.0105	0.8790	0.8276	<0.001
GLS4	261	46374	0.4959	0.5208	0.2307	0.7937	0.7398	<0.001
GLS6	281	46374	0.2632	0.2776	0.1489	1.0690	1.0300	<0.001

* p-value of Wilcoxon (one tail) matched pairs test.

CONCLUSION

Success of prediction was highly variable and dependent of the trials. Radial basis function kernel always outperformed linear kernel in both correlation and PMSE although this difference was only statistically significant in PMSE. Heffner *et al .(I.c.)* states that if net merit exceeds 0.50, GS could greatly outperform conventional MAS in terms of gain per unit time and cost. An improved method of optimization of parameters or more number of training examples should made SVR useful for Grey leaf spot - GS in breeding programs.