Intelligent Environments 2019
A. Muñoz et al. (Eds.)

doi:10.3233/AISE190031

Predictive Characterization of ICARDA Genebank Barley Accessions Using FIGS and Machine Learning

Zainab AZOUGH [a,b], Zakaria KEHEL [b], Aziza BENOMAR [a], Mostafa BELLAFKIH [a] and Ahmed AMRI [b]

[a] *INPT, Rabat, Morocco*
[b] *ICARDA, Rabat, Morocco*

Abstract. The International Center for Agricultural Research in the Dry Areas (ICARDA) has a unique germplasm collection of barley, among many other crops that it holds in its genebank. This collection contains landraces and barley wild relatives and most of them are georeferenced. Distribution of genetic resources is a core genebank activity aiming at responding to requests from various users including breeders, researchers, farmers, etc. ICARDA has developed over the last decade an efficient approach for better targeting adaptive traits called the Focused Identification of Germplasm Strategy (FIGS). FIGS approach links adaptive traits to environments (and associated selection pressures) through filtering and machine learning and it focuses on accessions that are most likely to possess trait specific genetic variation. In this paper, we present a work of predictive characterization on ICARDA barley collection using the FIGS approach and its algorithms combining several machine learning methods, and using several characterization traits. Most of the studied traits have shown a high predictability. Outcomes from this analysis are then used to make a predictive characterization of the entire ICARDA barley collection by assigning probabilities of each trait to the non-evaluated accessions.

Keywords. barley, characterization, FIGS, machine learning

1. Introduction

Genebanks worldwide hold collections of plant genetic resources for long-term conservation and maintain crop diversity for current and future use by crop improvement research, direct use and training. Most of genebanks are facing major problems of size and organization. Some collections have grown so large making their main activities which are the conservation and the use of the genetic diversity challenging. Another challenging aspect of plant genetic resources conservation is the lack of information about accessions specifically a precise evaluation information. This is mainly due to the challenge of evaluation the entire collection of a genebank. However, several genebanks have done a great job on maintaining and curating passport information. But passport data does not help a user of the genebank to discern which accession in a database is potentially containing the trait of interest. The concept of core and mini-core collections have been proposed as a strategy that allows the use of small portion of a germplasm collection to represent

the total collection. Core collections [1],[2] should be dynamic and need to be adjusted when additional germplasm and new information become available. The remaining accessions in the collection should however still be conserved as a secondary source of diversity. Concerns about core collections include rendering the reserve collection more vulnerable to loss, lack of representation of rare, endemic alleles, and poor relation to the specific needs of users. To address the latter concern, specialized core collections have been established around a particular trait, region, or type of material.
For adaptive traits, core and mini-core collections may not capture the needed diversity [3]. An alternative to random selection and core collections, the use of the focused identification of the germplasm strategy (FIGS), which is a trait-based approach, assists genebank managers identify desired genetic material with high probability of having the sought trait. In the last 10 years, ICARDA in collaboration with Vavilov institute in Russia and GRDC-Australia have invested in the development of FIGS that uses germplasm collection site agro-climatic and edaphic information to predict adaptive traits. The premise behind this approach is that the environment under which wild material and landraces will drive the evolution and selection of adaptive traits that could be of use to plant breeders. It seeks to determine and quantify relationships between collection site agro-climatic conditions and the presence of specific traits, such as disease resistance or heat resistance. FIGS has been successfully used to identify sources of resistance for several useful traits for breeding globally such as Sunn pest in wheat in Syria, Russian wheat aphid in bread wheat [4], abiotic stresses, such as drought adaptation in Vicia faba L. [5], resistance to stem rust in bread and durum wheat in [6] and [7], and stem rust and stripe rust in accessions of wheat landraces in [8] and [9]. FIGS is also as an efficient tool of linking genebank accessions to a trait of interest [10].

In this paper, we present a work of a predictive characterization on for ICARDA barley collection building on the FIGS approach by means of:

1. Assessing machine learning predictability for barley collection's characterization traits
2. Using the modeling outcomes to make a predictive characterization of the entire ICARDA barley collection by assigning probabilities to non-evaluated accessions.

2. Materials and Methods

2.1. Datasets Description: Accessions and Traits

ICARDA accessions database contains more than 32000 barley accessions including around 2400 wild relatives, distributed worldwide but collected mainly from the Fertile Crescent, North Africa, Ethiopia, East Europe and South East Asia (see Fig.1). ICARDA barley collection is ranked the second globally and represents 18% of the barley accessions conserved worldwide. More than 40 traits are used at ICARDA, as part of the genebank conservation effort, to characterize barley accessions including phenology, growth habit, morphology, yield components and some diseases. In this study, we used eight characterization traits as presented in Table 1. Table 1 showed a description of the traits that we are using for modeling in this study. The number of accessions evaluated is however greater than the number of geographic sites as we have multiple accessions

per sites in several cases. The sites are characterized by geographic coordinates (Longitude and Latitude). These traits were used as a response or dependent variable in the modeling.

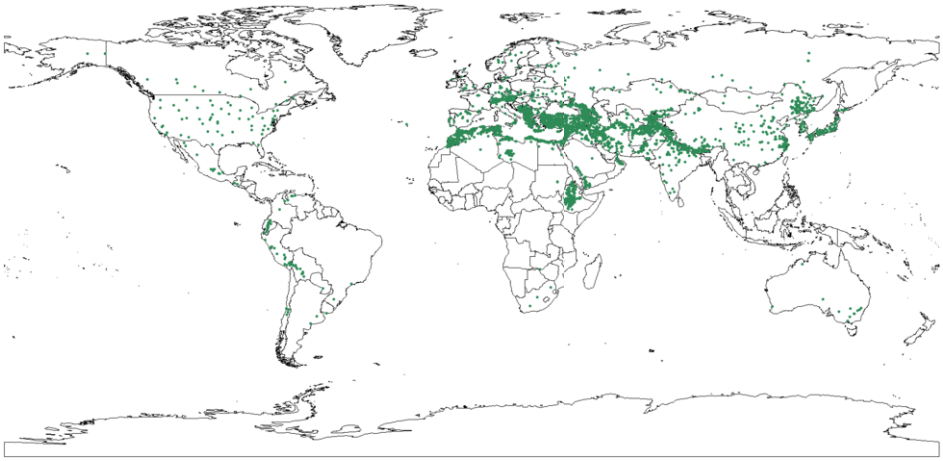

Figure 1. ICARDA barley accessions

Table 1. Characterization traits used for barley modeling

Trait	Accessions evaluated	Unique site evaluation	trait values
Days to heading	15027	3779	early, late
Days to maturity	15012	3775	early, late
Kernel weight	5413	1881	low, high
Productive tillering capacity	10286	2844	low, high
Kernel covering	18220	4562	naked, covered
Kernel row number	20256	4610	six-rowed, two-rowed
Growth class	18376	3932	winter, spring
Yellow rust	1683	756	resistant, susceptible

2.2. *Predictors: WorldClim and ENVIREM Data*

In the modeling, we used environmental data from WorldClim[1] and Envirem[2] databases as predictors.

WordClim Data WorldClim is an open access database providing global climatic layers describing past climatic profiles of collection sites and intended for spatial modeling or mapping. It includes average, monthly minimum and maximum temperatures, precipitation and bioclimatic variables [11].

[1] https://worldclim.org/
[2] https://envirem.github.io/

ENVIREM Data ENVIronmental Rasters for Ecological modeling (ENVIREM) is an open database of climatic and topographic variables used in species distribution modeling and other applications. The database contains 41 variables including aridity and potential evapotranspiration [12]. An example of rasters from the ENVIREM database is represented in the figure 2.

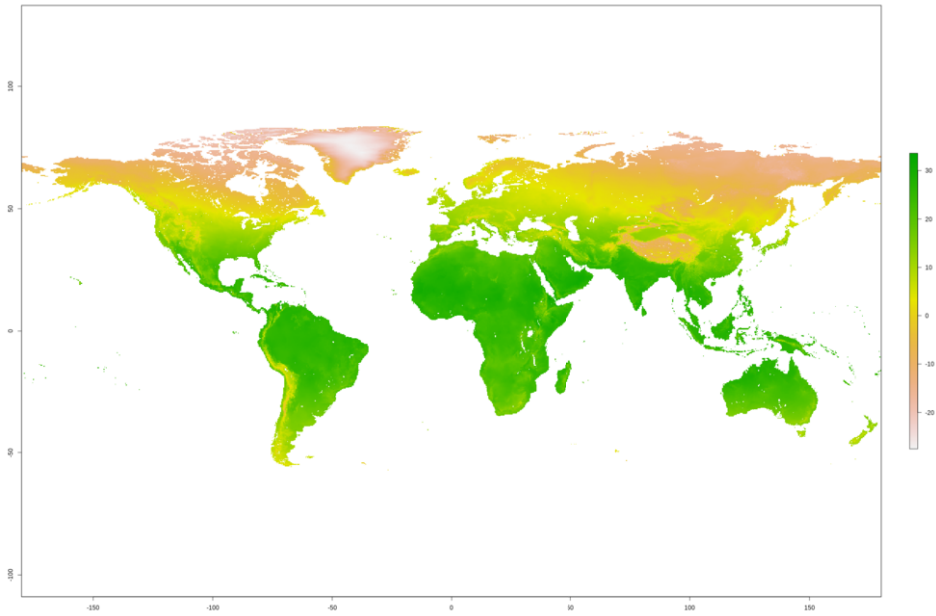

Figure 2. Annual Mean Temperature Distribution

2.3. Approach

The approach we used in this work is FIGS, which is a method based on two distinct pathways. The first pathway is using filtering when no evaluation data is available. This approach mimics the adaptation patterns of a trait and applies same selection pressure exerted on plants by evolution to develop a best subset containing accessions with high probability of having the adaptive traits. The second pathway is the machine learning approach used when partial evaluation of the collection is available. The machine learning algorithms find a function that links adaptive traits, environments (and associated selection pressures) with genebank accessions.

In the modeling, the following machine learning algorithms were used: K-nearest neighbours(KNN)[13], Support Vector Machines(SVM)[14], Random Forest(RF)[15], Artificial Neural Networks(NNET)[16] and Bagged Carts(BCART)[17]. Each machine learning model was tuned to select best tuning parameters using a training set and then the best model was selected between different machine learning models based on several metrics including accuracy, specificity and Kappa. These metrics were computed on the test set.

Then each trait was predicted for non-evaluated ICARDA barley accessions and we as-

signed probability of having the trait. The variable importance on the training set for all predictors were extracted from the best model. The importance of a predictor from a machine learning algorithm is generally calculated based on the increase in the model's prediction error after permuting the predictor.

In this study, R language was used, caret library was used for machine learning[18], rworldmap for geographic plotting[19].

3. Results

All studied traits showed high to medium predictability with a modeling accuracy between 0.757 for productive tillering capacity and 0.968 for kernel covering. Random forest(RF) model was selected as the best model for 6 traits while BCART was selected as the best model for two traits, see Table 2. Kappa was high for all traits, ranging from medium (between 0.5 and 0.6) and substantial (more than 0.6) showing that our predictions are not happening through random predictions. Sensitivities and specificities were also high for all traits validating the high predictability of the studied traits using environmental characterization of the landrace collection sites.

Table 2. Summary of modeling results

Trait	Best model	Accuracy	Kappa	Sensitivity	Specificity
Days to heading	BCART	0.789	0.575	0.752	0.822
Days to maturity	RF	0.772	0.543	0.784	0.761
Kernel weight	BCART	0.847	0.687	0.868	0.819
Productive tillering capacity	RF	0.757	0.514	0.667	0.847
Kernel covering	RF	0.969	0.759	0.711	0.99
Kernel row number	RF	0.862	0.634	0.935	0.667
Growth class	RF	0.865	0.533	0.484	0.97
Yellow rust	RF	0.815	0.561	0.872	0.686

Using the best model from machine learning modeling, the non-evaluated accessions for each trait are predicted using class probabilities at the threshold of 0.5. The predicted probabilities were plotted as a histogram (see Fig.3) to assess the patterns of class probabilities and show the power of the modeling in distinguishing between classes of a trait. Different patterns were found for different traits in separating between classes. Growth class and productive tillering capacity showed a clear separation between predicted classes while the models for the remaining traits had a medium capacity in separating between classes. Model for kernel weight and yellow rust had almost no capacity in separation between trait's classes. We extracted the variable importance from the best model for each trait. Figures 4 and 5 show two examples for kernel weight and covering. Potential evapotranspiration (PET) of the driest month, Precipitation of Driest Month and distance to rivers showing also the availability of water were important factors influencing whether an accession has a high or low kernel weight. On the other hand, mean temperature of wettest quarter and precipitation of warmest quarter were the climatic variables the most influencing whether a barley accession has a covered or naked kernel. Figure 6 is showing a map of the trait and the predicted characterization for the entire

barley collection at ICARDA for trait kernel row number.

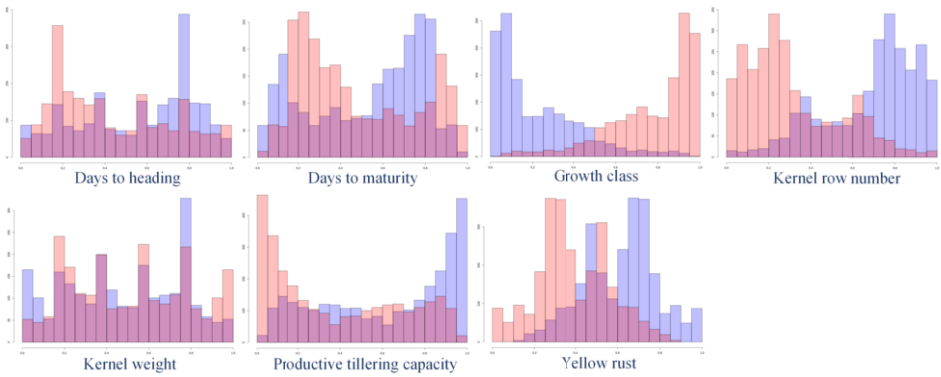

Figure 3. Predicted probabilities histograms for the studied trait (blue and orange are for positive and negative class respectively)

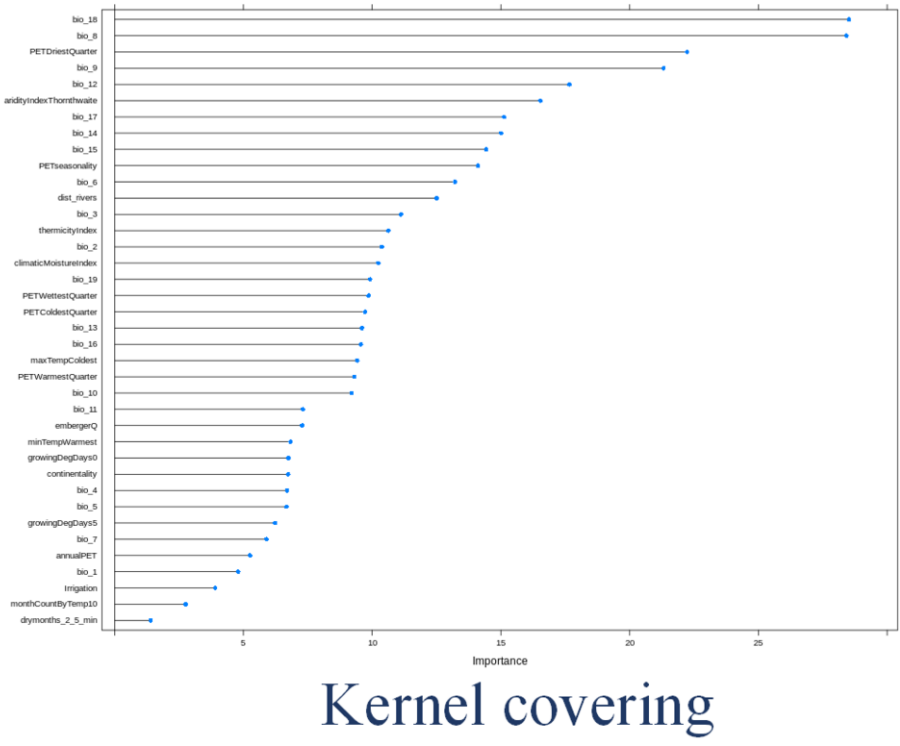

Figure 4. Variable importance graph for the kernel covering model

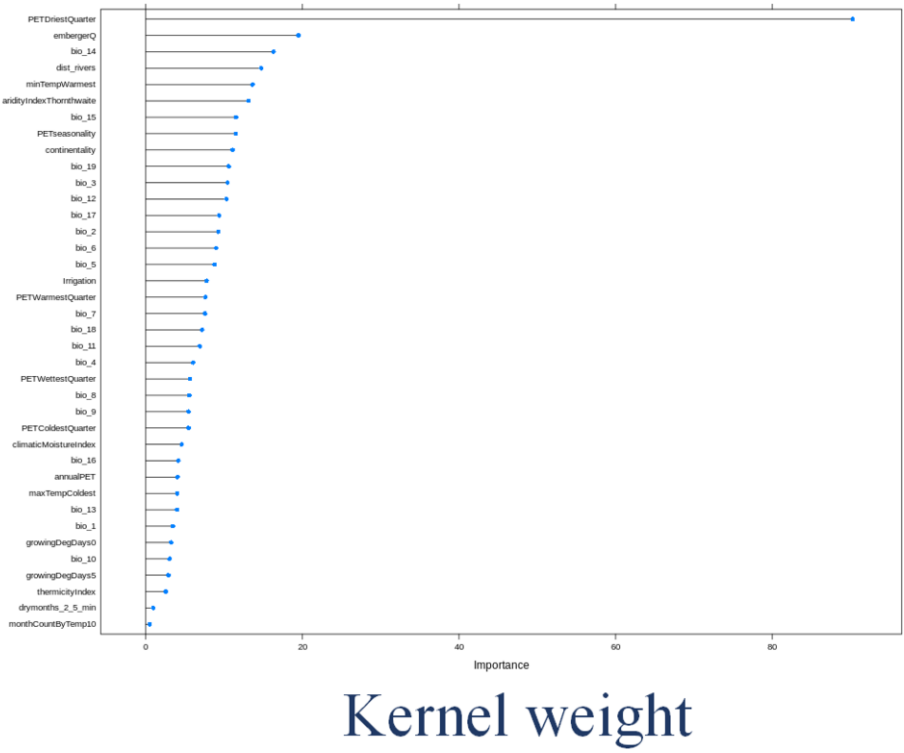

Figure 5. Variable importance graph for the kernel weight model

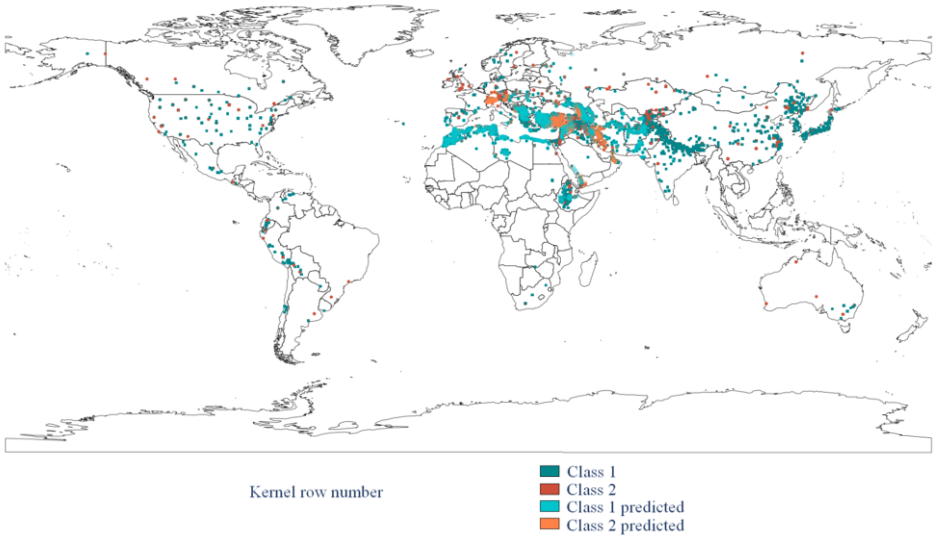

Figure 6. Classes and predictions map

4. Conclusion and Future Work

This work shows that FIGS approach through machine learning algorithms were efficient in finding environmental signals for the characterization traits measured as part of conservation efforts done by the ICARDA genebank. The high predictability of models for all traits was used later in predicting the non-evaluated accessions hold at ICARDA to assign probabilities for the characterization traits to accessions and stored in ICARDA genebank database and used as predictive characterization. This will assist genebank managers at ICARDA on replying more precisely to seed requests from users. This work will be enhanced by using molecular markers techniques to finetune further FIGS approach.

References

[1] Upadhyaya, H.D., Ortiz, R., Bramel, P.J. et al., Development of a groundnut core collection using taxonomical, geographical and morphological descriptors, Genetic Resources and Crop Evolution, (2003), https://doi.org/10.1023/A:1022945715628
[2] Upadhyaya, H D and Bramel, P J and Singh, S, Development of a chickpea core subset using geographic distribution and quantitative traits, Crop Science, (2001)
[3] Gepts, P., Plant Genetic Resources Conservation and Utilization, Crop Sci., (2006)
[4] El Bouhssini, M. , Street, K. , Amri, A. , Mackay, M. , Ogbonnaya, F. C., Omran, A. , Abdalla, O. , Baum, M. , Dabbous, A. and Rihawi, F., Sources of resistance in bread wheat to Russian wheat aphid (Diuraphis noxia) in Syria identified using the Focused Identification of Germplasm Strategy (FIGS), Plant Breeding, (2011), doi:10.1111/j.1439-0523.2010.01814.x
[5] Khazaei H, Street K, Bari A, Mackay M, Stoddard FL, The FIGS (Focused Identification of Germplasm Strategy) Approach Identifies Traits Related to Drought Adaptation in Vicia faba Genetic Resources, PLoS ONE, (2013), doi:10.1371/journal.pone.0063107
[6] Thormann I, Parra-Quijano M, Rubio Teso ML, Endresen DTF, Dias S, Iriondo JM, Maxted N., Predictive characterization methods for accessing and using CWR diversity, In: Maxted N, Dulloo ME, Ford-Lloyd BV, eds. Enhancing crop genepool use. Capturing wild relative and landrace diversity for crop improvement, Boston: CABI International, (2016)
[7] Endresen, D. T. F., K. Street, M. Mackay, A. Bari, A. Amri, E. De Pauw, K. Nazari, and A. Yahyaoui, Sources of Resistance to Stem Rust (Ug99) in Bread Wheat and Durum Wheat Identified Using Focused Identification of Germplasm Strategy, Crop Sci., (2012), doi:10.2135/cropsci2011.08.0427
[8] Bari, A., Street, K., Mackay, M. et al., Focused identification of germplasm strategy (FIGS) detects wheat stem rust resistance linked to environmental variables, Genet Resour Crop Evol, (2012), https://doi.org/10.1007/s10722-011-9775-5
[9] BARI, A., AMRI, A., STREET, K., MACKAY, M., DE PAUW, E., SANDERS, R., NAZARI, K., et al., Predicting resistance to stripe (yellow) rust (Puccinia striiformis) in wheat genetic resources using focused identification of germplasm strategy, The Journal of Agricultural Science, (2014)
[10] Noelle L. Anglin, Ahmed Amri, Zakaria Kehel, and Dave Ellis, A Case of Need: Linking Traits to Genebank Accessions, Biopreservation and Biobanking, (2018)
[11] Fick, S.E. and R.J. Hijmans, Worldclim 2: New 1-km spatial resolution climate surfaces for global land areas, International Journal of Climatology, (2017)
[12] Title P.O., Bemmels J.B., ENVIREM: an expanded set of bioclimatic and topographic variables increases flexibility and improves performance of ecological niche modeling, Ecography, (2018)
[13] S.B. Kotsiantis, Supervised machine learning: a review of classification techniques, Informatica, (2007)
[14] C.-W. Hsu, C.-C. Chang, C.-J. Lin, A Practical Guide to Support Vector Classification, Department of Computer Science, National Taiwan University, Taipei, Taiwan, (2010)
[15] Breiman, L. , Random forests, Machine Learning, (2001)
[16] R. Rojas, Neural Netwoks: A Systematic Introduction, Springer-Verlag, Berlin (1996)
[17] A. Kołcz, N-tuple Network, CART, and Bagging, in Neural Computation, (2000)

[18] Kuhn, Max., Building Predictive Models in R Using the caret Package, Journal of Statistical Software, (2008)
[19] South, Andy, rworldmap: A New R package for Mapping Global Data, (2011)