

# RESEARCH PROJECT PROPOSAL

## **A PROPOSED MACHINE LEARNING FRAMEWORK FOR THE CLASSIFICATION OF GENES ASSOCIATED WITH ABIOTIC STRESS-RESISTANT TRAITS IN RICE**

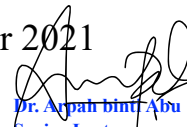
KOK HERNG

17203580/1

BIOINFORMATICS PROGRAMME  
INSTITUTE OF BIOLOGICAL SCIENCES  
FACULTY OF SCIENCE  
UNIVERSITI MALAYA

SUPERVISOR: DR. ARPAH BINTI ABU

DATE: 12<sup>th</sup> November 2021



Dr. Arpah Binti Abu  
Senior Lecturer  
Institute of Biological Sciences, Faculty of Science  
Universiti Malaya, 50603 Kuala Lumpur  
+603 7967 6742  
arpah@um.edu.my

## **1. Field of Research**

Computational Bioinformatics

## **2. Topic of Research**

Machine learning-based predictive model in classification genes and SNPs that are related to abiotic stress-resistant traits in rice

## **3. Statement of Problem**

The discovery of contributing molecular mechanisms linked to abiotic stress-resistant traits in rice requires the identification of relevant genes among thousands of candidate genes predicted by high throughput genomics or traditional linkage analysis. However, the experimental validation of high number of candidate genes are proved to be time- and resource-consuming. Therefore, efficient computational approaches are needed to address this problem.

In silico techniques such as UCSC Xena and The Broad GDAC Firehose ((Khurshed, Molenaar, & Noorden, 2019), and with the technological advances in genomics such as new generation sequencing technologies and functional genomics, high dimensional annotations for individual genes which provide information describing gene-gene interactions and networks can be generated. The availability of such vast amounts of information, however, poses additional challenges, which include, inter alia, the need for the integration of heterogeneous data from multiple sources and the extraction of most important information from the high dimensional feature space.

Network-based gene prioritization such as Knowledge Network Gene Prioritization (Kimmel & Visweswaran, 2013) and GenePANDA (Yin, Chen, Wu, & Tian, 2017) are some of the computational algorithms that has been extensively studied not only on *Arabidopsis thaliana*, but also on other crops such as corn and rice. However, some rice gene networks associated with specific phenotypes are different from those in *Arabidopsis* and therefore, a species-specific protein-protein interaction network using machine learning (ML) was developed by Liu et al., 2017 to facilitate the systematic discovery of genes that control specific phenotypes, including important agronomic traits in rice.

In the world of ML, the majority of algorithms are supervised. Supervised ML describes the fitting of a model to data/ a subset of data that have been labelled, which is usually experimentally measured or assigned by humans that derived ultimately from laboratory observations, but often these raw data are pre-processed in some way. Supervised ML come in

three major types: Regression, Classification and Hybrid.

In regression models, the task is to approximate a mapping function from input variable to a continuous output variable (Brownlee, 2017). One of the examples for regression would be Linear Regression, where we create a correlation between a dependent variable, and one or more independent variables using a straight line (regression line). Another example is Support Vector Machine, where we search for a fitting function  $f(x)$ , having deviation less than  $\varepsilon$  from the target ( $y_i$ ) acquired for the relating training data set (Choudhary & Gianey, 2017).

On contrast, classification approximates a mapping function from input variables to discrete output variables (Brownlee, 2017). For example, Logistic Regression is used to predict the probability of an outcome having only two values, also known as binary classification (Choudhary & Gianey, 2017). Besides that, Naïve Bayesian classifiers assign the most likely class to a given example described by its feature vector (Rish, 2001).

Finally, hybrid models can often depend on parts of regression to advise how to do classification, or sometimes the opposite (Burger, 2018). According to Al Amrani, Lazaar, & El Kadiri, (2018), the Random Forest is an ensemble learning method for classification and regression that constructs decision trees and delivers the class that is the mode of the classes output by individual trees. Besides that, a Neural Network utilizes both regression and classification. It is given a list of inputs, then the neural network performs a number of processing steps before returning an output (Burger, 2018).

In general, ML is gaining ground in plant research (van Eeuwijk et al., 2019; Singh et al., 2016, 2018; Mochida et al., 2019; Sperschneider, 2019; Sun et al., 2019; Wang et al., 2020). However, there is still more room for exploration specifically for rice. The advent of ML makes it possible to establish the model causal genes and SNPs that are related to abiotic stress-resistant traits in rice. This will give an insight into specific genes and SNPs that are associated with abiotic stress-resistant traits. This study aims to establish the ML framework using supervised classifiers by utilizing existing rice information and knowledge for the classification of such candidate genes and SNPs, which can improve the efficiency and accuracy of data analysis. Results obtained from this study can be extended to further development of a decision support system in rice gene and SNP validation, as well as application in genome editing.

#### 4. Research Aim and Objective

This study aims to establish the predictive model for the classification of genes associated with abiotic stress-resistant traits in rice. Specifically, the objectives of this study are:

- 1) To construct the supervised machine learning-based predictive model of causal genes that are related to abiotic stress-resistant traits.
- 2) To validate the accuracy and efficiency of the predictive models.

#### 5. Research Methodology

##### 5.1. Data acquisition

Rice (*Oryza sativa*) will be collected from MARDI as well as public databases, knowledge sources, and scientific literature. The following are the data that will be used in this study and its sources.

- Genome data

A list of stress-resistant and susceptible rice genomes and genes will be extracted from the RSS database. The Malaysian rice genomes with different stress-resistant and susceptible traits in the analysis will also be included.

- Expression data

The gene expression (RNA-seq, microarray) data will be extracted from ArrayExpress, Expression Atlas and Rice Expression Database.

- QTL & GWAS

The QTL and GWAS data that related to abiotic stress studies will be extracted from HapRice, SNP-Seek, IC4R, RiceVarMap, RiceDiversity databases.

- Known/Functional genes

Several rice databases have been developed to curate and provide information on previously known genes that are associated with abiotic trait. This known/functional gene will be extracted from those databases (i.e. MBKBASE, funRiceGenes)

Next, the collected data will be cleaned, validated, and curated into the genes associated with the abiotic stress-resistant trait in rice.

## 5.2. Materials

### 1. Software specification

The Integrated Development Environment (IDE) chosen for R will be RStudio, with packages such as tidyverse (for data visualization and transformation) and caret (for model construction).

### 2. Hardware specification

Table 5.2 shows the specification of the laptop that will be used in the construction of the predictive model.

Table 5.2: Hardware specification

Hardware	Specification
Processor	AMD Ryzen 7 3750H @ 2.30 GHz
RAM	16GB
Hard disk	500GB

## 5.3. Proposed solution / methodology

A predictive model will be constructed using supervised learning. Supervised learning tasks aim to predict outputs (classifying gene and SNP, either a discrete label, in the case of classification or a numerical value, in the case of regression) for a given object, given a set of input features (abiotic stress-resistant traits) that describes the object. Supervised learning optimizes a predictive model by fitting its parameters to perform well on labelled training data consisting of inputs and corresponding known outputs. The resulting models then will be used to make predictions for new, unseen test data. Convolutional neural network (CNN), autoencoder, random forests (RF), support vector machine (SVM) and logistic regression will be studied in this project.

Generally, the selection of algorithms to construct the models will be based on the curated data (will be known as features later). Besides that, other factors such as heterogeneity and redundancy of the data, and the presence of interaction and nonlinearities will be considered too. The purpose to employ feature selection methods is to look into the features which are most significant and meaningful in the modeling process, in order to obtain more accurate prioritization. Recursive feature elimination (RFE), Correlation-based feature selection (CFS) and Pearson's correlation coefficient (PCC), are the methods that will be tested in this project.

#### 5.4. Gantt Chart

Refer to the **Appendix 1**.

### 6. Research Significances

The ML framework will provide an insight into specific genes and SNPs that are associated with abiotic stress-resistant traits in rice. It will allow swift screening of the potential genes and SNPs. Therefore, it is highly valuable for biologists, geneticists, rice breeders as well as agriculture sectors for the development of new and improved rice varieties.

Besides that, it will assist in resource optimization and cost reduction in the field of selection-assisted reproduction (MAS) and genome editing can be utilized through the development of a one-stop digital platform.

Moreover, it will promote innovation in the agriculture sector through precision biotechnology methods.

### 7. References

- Al Amrani, Y., Lazaar, M., & El Kadiri, K. E. (2018). Random Forest and Support Vector Machine based Hybrid Approach to Sentiment Analysis. *Procedia Computer Science*, 127, 511-520.
- Brownlee, J. (2017). Difference Between Classification and Regression in Machine Learning. Retrieved from <https://machinelearningmastery.com/classification-versus-regression-in-machine-learning/>
- Burger, S. V. (2018). *Introduction to Machine Learning with R: Rigorous Mathematical Analysis* (1st ed.): O'Reilly Media.
- Choudhary, R., & Gianey, H. K. (2017, 14-15 Dec. 2017). *Comprehensive Review On Supervised Machine Learning Algorithms*. Paper presented at the 2017 International Conference on Machine Learning and Data Science (MLDS).
- Khurshed, M., Molenaar, R. J., & Noorden, C. J. v. (2019). A simple in silico approach to generate gene-expression profiles from subsets of cancer genomics data. *BioTechniques*, 67(4), 172-176.
- Kimmel, C., & Visweswaran, S. (2013). An Algorithm for Network-Based Gene Prioritization That Encodes Knowledge Both in Nodes and in Links. *PLOS ONE*, 8(11), e79564.
- Liu, S., Liu, Y., Zhao, J., Cai, S., Qian, H., Zuo, K., . . . Zhang, L. (2017). A computational interactome for prioritizing genes associated with complex agronomic traits in rice (*Oryza sativa*). *The Plant Journal*, 90(1), 177-188.
- Mochida, K., Koda, S., Inoue, K., Hirayama, T., Tanaka, S., Nishii, R., & Melgani, F. (2018). Computer vision-based phenotyping for improvement of plant productivity: a machine learning perspective. *GigaScience*, 8(1).
- Rish, I. (2001). *An empirical study of the naive Bayes classifier*. Paper presented at the IJCAI 2001 workshop on empirical methods in artificial intelligence.

- Singh, A., Ganapathysubramanian, B., Singh, A. K., & Sarkar, S. (2016). Machine Learning for High-Throughput Stress Phenotyping in Plants. *Trends in Plant Science*, 21(2), 110-124.
- Singh, A. K., Ganapathysubramanian, B., Sarkar, S., & Singh, A. (2018). Deep Learning for Plant Stress Phenotyping: Trends and Future Perspectives. *Trends Plant Sci*, 23(10), 883-898.
- Sperschneider, J. (2020). Machine learning in plant–pathogen interactions: empowering biological predictions from field scale to genome scale. *New Phytologist*, 228(1), 35-41.
- Sun, S., Wang, C., Ding, H., & Zou, Q. (2019). Machine learning and its applications in plant molecular studies. *Briefings in Functional Genomics*, 19(1), 40-48.
- van Eeuwijk, F. A., Bustos-Korts, D., Millet, E. J., Boer, M. P., Kruijer, W., Thompson, A., . . . Chapman, S. C. (2019). Modelling strategies for assessing and increasing the effectiveness of new phenotyping techniques in plant breeding. *Plant Science*, 282, 23-39.
- Wang, H., Cimen, E., Singh, N., & Buckler, E. (2020). Deep learning for plant genomics and crop improvement. *Current Opinion in Plant Biology*, 54, 34-41.
- Yin, T., Chen, S., Wu, X., & Tian, W. (2017). GenePANDA—a novel network-based gene prioritizing tool for complex diseases. *Scientific Reports*, 7(1), 43258.

Appendix 1

TASK	Semester 1, 2021/2022														Semester 2, 2021/2022													
	WEEKS																											
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Lab Work / Field Work																												
Data gathering and extraction																												
Data cleaning																												
Data curation and integration																												
Features selection implementation																												
Models building implementation																												
Models testing and evaluation																												
Research Project Proposal																												
Submission of proposal draft																												
Submission of proposal																												
Oral Presentation																												
Mid-term review																												
Seminar presentation																												
Thesis																												
Submission of first draft																												
Submission of second draft																												
Submission of final draft																												
Submission of Materials for Examination																												
Thesis																												
Log Book																												
Poster																												