UNIVERSITY OF SÃO PAULO

LORENA SCHOOL OF ENGINEERING

LEONARDO FERNANDES SOUZA

Machine Learning applied to the study of interactions between micro-RNA and biotic stress in plants

LORENA

2021

LEONARDO FERNANDES SOUZA

Machine Learning applied to the study of interactions between micro-RNA and biotic stress in plants

Monograph presented to the Biochemical Engineering course of the Lorena School of Engineering of the University of São Paulo, as part of the requirements for obtaining the title of Biochemical Engineering.

Advisor: Prof. Dr. Valdeir Arantes

LORENA

2021

ESTE EXEMPLAR CORRESPONDE A VERSÃO FINAL DO TRABALHO DE CONCLUSÃO DE CURSO DO ALUNO LEONARDO FERNANDES SOUZA, ORIENTADO PELO PROF. DR. VALDEIR ARANTES

ASSINATURA DO ORIENTADOR

Ficha catalográfica elaborada pelo Sistema Automatizado
da Escola de Engenharia de Lorena,
com os dados fornecidos pelo(a) autor(a)

Souza, Leonardo
Machine Learning applied to the study of interactions between micro-RNA and biotic stress in plants / Leonardo Souza; orientador Valdeir Arantes. – Lorena, 2021.
35 p.

Monografia apresentada como requisito parcial para a conclusão de Graduação do Curso de Engenharia Bioquímica – Escola de Engenharia de Lorena da Universidade de São Paulo. 2021

1. Machine learning. 2. Microrna. 3. Biotic stress. 4. Plant-pathogen. I. Título. II. Arantes, Valdeir, orient.

RESUMO

SOUZA, L. F. **Aprendizado de Máquina aplicado ao estudo das interações entre microRNA e estresse biótico em plantas**. 2021. Monografia (Trabalho de Conclusão de Curso II – Engenharia Bioquímica) – Escola de Engenharia de Lorena, Universidade de São Paulo, Lorena, 2021

O desafio de garantir segurança alimentar para o futuro da população mundial é uma das principais preocupações de pesquisadores de biologia de plantas. Devido às mudanças ambientais e o aumento esperado da população global, desenvolver cultivos com alto rendimento e tolerantes à estresse é uma prioridade para os próximos anos. MicroRNAs (miRNAs) são pequenos RNAs não-codificantes, endógenos de 20-22 nucleotídeos codificados por genes de microRNA (MIR genes) que conseguem regular o nível transcricional da planta reprimindo a expressão de seus alvos. Cientistas têm descoberto muitas interações entre microRNA e as respostas das plantas contra estresse biótico, sugerindo a importância dos microRNAs contra o ataque de patógenos. Conforme novas técnicas de alto rendimento vem sendo desenvolvidas, Aprendizado de Máquina (AM) tem emergido como uma ponderosa ferramenta para extrair informação de complexos dados biológicos. Algoritmos de AM podem explorar grandes bancos de dados e estabelecer relações em modelos não lineares sem muito conhecimento sobre os padrões dos dados antecipadamente. Essa revisão almeja validar o potencial use de aprendizado de máquina para analisar microRNAs em plantas sob estresse biótico, predizendo seus alvos e elucidando suas complexas interações de funcionalidade. Além disso, este trabalho deve preencher a lacuna de conhecimento necessário entre os campos da ciência da computação e da biotecnologia para permitir que os pesquisadores apliquem o aprendizado de máquina em seus estudos sobre miRNAs e interações planta-patógeno.

Palavras-chaves: Aprendizado de Máquina, microRNA, Estresse biótico, Planta-patógeno

ABSTRACT

SOUZA, L. F. **Machine Learning applied to the study of interactions between micro-RNA and biotic stress in plants**. 2021. Monograph (Course Conclusion Work II - Biochemical Engineering) – Lorena School of Engineering, University of São Paulo, Lorena, 2021

The challenge of guarantying food security to the future world population is a significant concern of plant biology researchers. Due to environmental changes and the expected increase of the global population, developing high-yielding and stress-tolerant crops are priorities for the following years. Micro-RNAs (miRNAs) are small non-coding endogenous RNAs of 20-22 nucleotides encoded by microRNA genes (MIR genes) which can regulate plant transcriptional levels repressing the expression of their targets. Scientists have described many interactions between microRNAs and plants' responses to biotic stress suggesting their importance against pathogens attacks. As new high-throughput genomic techniques have been developed, Machine Learning (ML) has emerged as a powerful tool to extract information from complex biological data. ML algorithms can explore large datasets and establish relations on non-linear models without knowing much about the data patterns in advance. This review aims to validate the potential use of machine learning to analyze microRNAs in plants over biotic stress predicting their target and elucidating their complex interactions functionality. Furthermore, this work should fill the gap of required knowledge between computer science and biotechnology fields to enable researchers to apply machine learning to their studies about miRNAs and plant-pathogen interactions

Keywords: Machine Learning, microRNA, Biotic Stress, Plant-Pathogen

LIST OF FIGURES

Summary

1 Introduction

Extreme climatic conditions and the increase of the global population have taken attention from plant biology researchers. According to the United Nations Population Division, the global population will touch the mark of 8.3 billion by 2030. Assuring that food, fodder, fiber, and fuel demand will be met for the following years requires the development of high-yielding and stress-tolerant crop varieties. An emerging concern is the climatic variations in some regions of the globe, as being sessile, plants have to face and cope with the environment they live in. This change will lead to shifts in biomes and can affect the sensible relationship between diseases and crops (CHAUDHARY; GROVER; SHARMA, 2021a; THUDI et al., 2021).

Genetic engineering is currently being utilized to enhance desired crop traits. However, since a single trait might be controlled by many genes, manipulating agronomical traits requires genetic modulators that act precisely and target in a specific manner (CHAUDHARY; GROVER; SHARMA, 2021a).

MicroRNA manipulation is emerging as a potential target for genetic engineering to improve the agronomic properties of crops. Regulating miRNAs expression level is an efficient strategy to improve plant responses to environmental stresses (biotic and abiotic), plant growth, and development levels (DJAMI-TCHATCHOU et al., 2017).

Machine Learning (ML) has arisen as a powerful tool to analyze biological data in the last years. ML can investigate many data instances and reveal complex interactions, such as predicting miRNA targets (KURUBANJERDJIT et al., 2013).

This work aims to validate the potential application of Machine Learning to understand the role of miRNAs in plants under biotic stress conditions by a literature review. It will also be structured as a guideline to help researchers

applying Machine Learning tools to their plant biotic stress experiments. It should fill the gap of required knowledge between the fields of computer science and biotechnology.

2 Plant immune system and miRNA

Plants, being sessile, cannot avoid or flee from adverse environmental challenges and contact with other living organisms. To withstand the biotic and abiotic stress they face during the entire lifecycle, plants have developed sophisticated defense mechanisms to notice such attacks and initiate adaptative responses (GIMENEZ; SALINAS; MANZANO-AGUGLIARO, 2018).

2.1 Plants and biotic stress

From seed emergence to mature lifecycle, plants deal with the damage caused by parasites and pathogens, such as viruses, fungi, bacteria, nematodes, or insects. This damage is known as biotic stress and, despite the defense mechanisms evolved for centuries by plants, it is still one of the reasons for significant economic losses from crop fields every year (JEYARAJ et al., 2020).

Plant pathogens can be classified by their different strategies to attack plants: necrotrophs, which kill the plant tissue before feed on its content (e.g., rotting bacteria); biotrophs, which maintain the plant cells alive during infection (e.g., viruses, nematodes, fungi); hemibiotrophs, which parasitize the living tissue for some time and follow with a necrotrophic phase (e.g., oomycetes) (SERGEANT; RENAUT, 2010).

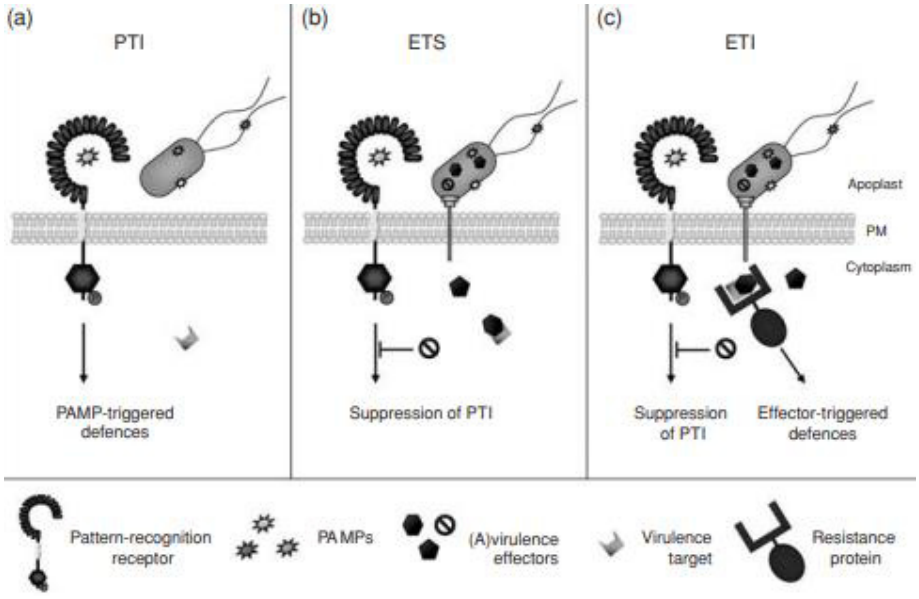
2.2 Plant immune system

Unlike mammals, plants do not own mobile defender cells and do not possess an "antibody" circulatory system. Instead, they depend on each cell performing immune functions and on signals emanating from the infected areas (DMITRIEV, 2003). The first defense line is passive and consists of physical

barriers such as thick cuticles, waxes, and cell walls. Plants also produce constitutive chemical compounds to protect themselves from microbes and herbivores. They are toxic secondary metabolites and proteins that inhibit the functions of pathogens. The passive defense layer is not pathogen-specific and, its amplitude depends on the species and the environmental conditions (SERGEANT; RENAUT, 2010).

The plant immune system consists of two branches (Fig. 1): PAMP-triggered immunity (PTI) and effector-triggered immunity (ETI). The first is based on identifying pathogen-associated molecular patterns (PAMPs) such as flagellin from bacteria, chitin and ergosterol from all fungi, and transglutaminase from Oomycetes. The detection of molecular structures unique to microorganisms by Pattern Recognition Receptors (PRRs) transmits information across the plasma membrane, beginning a host-signaling sequence that triggers pathogen non-specific immune responses (NÜRNBERGER; KEMMERLING, 2018).

Fig. 1: The plant immune system.



(a) PAMPs triggers immune defenses (PTI). (b) Virulent pathogens release effectors that facilitate infections (effector-triggered susceptibility, ETS). Some suppress PTI. (c) Resistance protein recognizes these effectors and trigger ETI.

Font: (NÜRNBERGER; KEMMERLING, 2018).

However, successful pathogens deploy effectors that contribute to their virulence resulting in effector-triggered susceptibility (ETS). Some pathogens have evolved effector proteins capable of suppressing the PTI signaling. As an evolutionary response, the second plant immunity system branch, ETI, recognizes specific effectors released by pathogens either indirectly or by directly "nucleotide-binding sites plus leucine-rich repeat" proteins (NBS-LRR) recognition (JONES; DANGL, 2006).

ETI is a result of gene-for-gene resistance. A pathogen factor that enhances pathogenicity on a specific host plant is counteracted by a plant factor. These interactions are driven by pathogen avr (avirulence) gene loci and alleles of the corresponding plant disease resistance (R) locus. R products recognize avr-dependent signals and trigger the chain of signal-transduction events that culminates in the activation of defense mechanisms and an arrest of pathogen growth. The largest class of R genes encodes a cytoplasmic receptor-like protein, NBS-LRR. When the receptors interact with specific elicitors produced by the avr genes, some form complexes responsible for activating defense pathways, often associated with other kinases (DANGL; JONES, 2001).

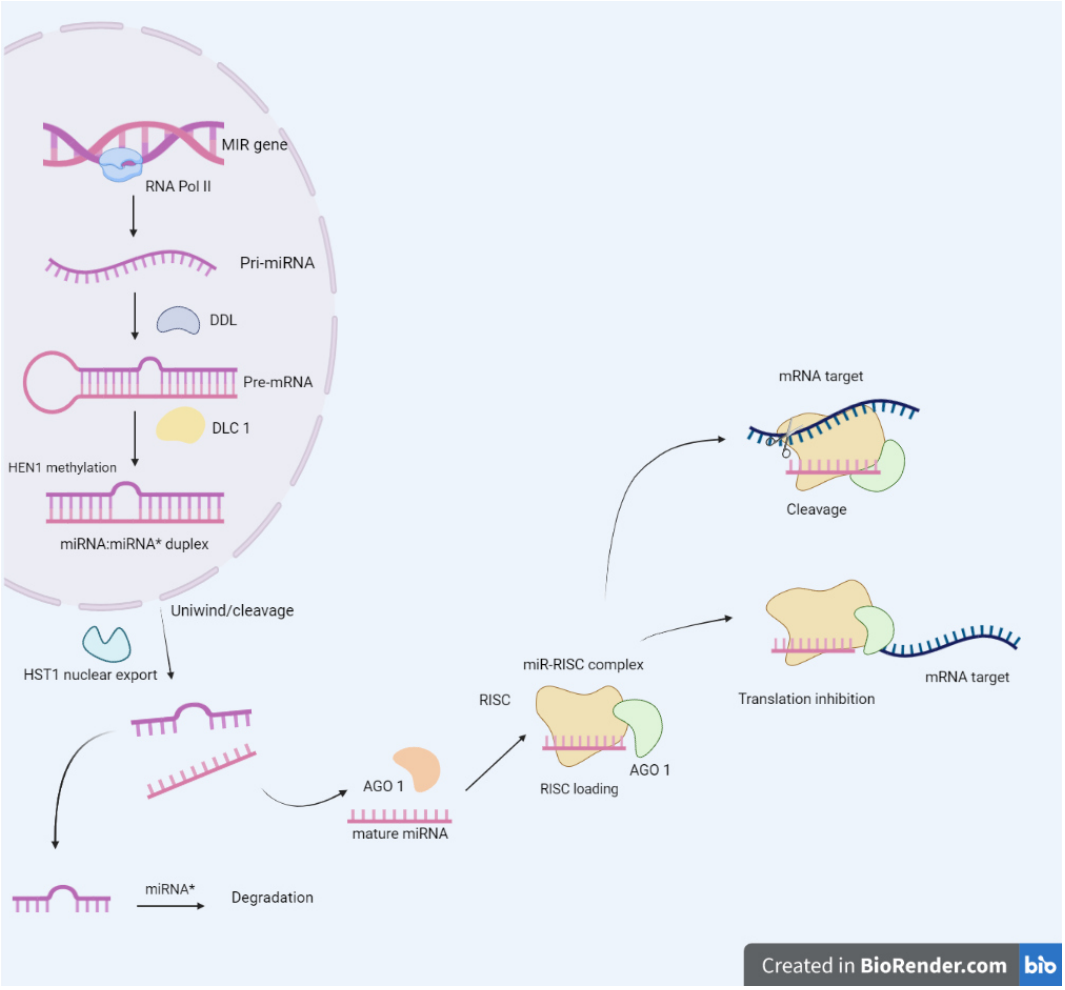
After pathogen recognition, plants start their defense strategy: confine and attack the pathogen with phytoalexins (antimicrobials). Some changes occur in ion fluxes and, the concentration of salicylic acid (SA) increases, which is a signal to establish a non-specific resistance: systematic acquired resistance (SAR). SAR enhances the resistance against a broad spectrum of pathogens during several weeks. Furthermore, the cell produces Reactive oxygen species (ROS) and nitric oxide (NO), which are responsible for initializing a hypersensitive response (HR). An HR leads to apoptosis of cells surrounding the infection, thus preventing its spreading. It also triggers the strengthening of the cells walls and the production of Pathogenesis-related proteins, besides enzymes involved in synthesizing

phytoalexins. Mitogen-Activated Protein Kinases (MAPKs) are activated to control de protein activity level in the vicinity of the invasion site (DMITRIEV, 2003).

2.3 miRNA

Small non-coding RNAs (sRNA) also play an essential role in mediating genes that encode resistant proteins. In addition to being involved in development, growth, cell proliferation, and other variety of biological processes, they target transcriptions factors of defense genes such as MAPKs and NBS-LRRs genes (KULSHRESTHA et al., 2020).

Fig. 2: Micro-RNA biogenesis



Font: Adapted from (CHAUDHARY; GROVER; SHARMA, 2021b).

In plants, two main classes of sRNA can be distinguished: micro-RNA (miRNA) and small-interfering RNA (siRNA). In brief, they differ from each other in their biogenesis and functions (ALI et al., 2020).

The miRNAs are small non-coding endogenous RNAs of 20-22 nucleotides (nt) encoded by microRNA genes (MIR genes) (Fig. 2). The MIR genes are found in intergenic areas or within introns of other genes. RNA polymerase II transcribes a long single-stranded primary miRNA (pri-miRNA) from nuclear MIR genes. The pri-miRNA folds into a stem-loop structure stabilized by RNA-binding protein DAWDLE (DDL), forming a precursor miRNA (pre-miRNA). Subsequently, an endoribonuclease named DICER-like (DCL1) and other proteins such as HYPONASTIC LEAVES 1 (HYL1) and SERRATE (SE) process the pre-miRNA structure and create a miRNA:miRNA* duplex. To protect the recently generated duplex from degradation, a methyltransferase protein, HUA ENHANCER 1 (HEN1), methylate the structure at the 3' terminus and export it into the cytoplasm assisted by HASTY (HST1), an exportin protein (KHRAIWESH; ZHU; ZHU, 2012).

In the cytoplasm, one strand of the duplex (the mature miRNA) is stabilized by an ARGONAUTE 1 (AGO1) protein, while the other strand is degraded by exosomes. The stabilized complex AGO1-miRNA is loaded to an RNA-induced gene silencing complex (RISC) and guides the binding to cognate targets by sequence complementarity. The complex pair with the mRNA target and induces its cleavage or repress its translation. In addition to post-transcriptional control, miRNAs regulate gene expression by causing epigenetic changes such as mRNA poly-(A) tail cleavage and DNA and histone methylation (KHRAIWESH; ZHU; ZHU, 2012).

2.4 miRNA and biotic stress

According to Tang J., Chu C. (2017), 65.9% of miRNA targets are transcription factors (TFs), and 6.6% targets NBS-LRR protein encoded genes indicating the role of miRNAs in diverse gene regulatory networks and plant immune systems.

Several examples of interactions between miRNA and plants over biotic stress have been described in the last decades. For instance, Navarro et al. (2006) showed that *Arabidopsis thaliana* exposed to a flagellin-derived peptide (flg22) induced a miR393 that down-regulates auxin signaling by targeting auxin receptor transcripts increasing bacterial resistance. LI et al. (2014) investigated the rice *(Oryza sativa)* immunity against the blast fungus *Magnaporthe oryzae* and showed that transgenic plants overexpressing miRNA160a and miRNA398b exhibited enhanced resistance to *M. oryzae* up-regulating the expression of defense-related genes. Moreover, Liang et al. (2019) has successfully applied artificial miRNA technology to control cucumber green mottle mosaic virus (CGMMV).

Given the importance of identifying the miRNAs targets for the analysis of miRNAs role in the complex network that regulates stress response, biological approaches alone are not sufficient to solve this problem. Experimental validation of every potential miRNA target is time-consuming and expensive. For a given miRNA, a large number of potential targets may be present. Therefore, there is a need for a strategy able to reduce the number of possible miRNA targets in advance, thus becoming feasible to apply experimental approaches to validate and characterize their functions and effects (GIANSANTI et al., 2019; RIOLO et al., 2020).

In addition to identifying target miRNAs, further investigation into the responses of different plant stress levels and the co-interaction between miRNAs is also a topic of interest. Plants produce miRNAs at distinct levels throughout their

life cycle, and due to the large number of miRNAs that can be involved in a single signaling response, distinguishing how their concentrations alter and their specific contributions when exposed to pathogen attacks is not an easy task.

2.5 Computational approaches

The bioinformatics community has developed several tools for computational analysis underlying the development of miRNA target prediction algorithms. These tools are based on different biological properties of mRNA sequence and miRNA-mRNA interactions. Although they can significantly reduce the number of putative miRNAs, they often result in inconsistent predictions compared to each other, leading to a high level of false positives targets after being validated experimentally. Statistical inference based on Machine Learning has emerged as an integrated approach to limit the number of false positives and strengthen the value of the predictions (RIOLO et al., 2020).

Machine learning can also be used to investigate further plant stress by having the plant miRNA expressions. Rather than linear methods, which generally cannot describe data with multiple inputs and synergies, ML algorithms learn the complex non-linear patterns from training data and predict the stress condition of unknown plant samples in addition to miRNA interactions (ASEFPOUR VAKILIAN, 2020).

3 Biological Data

Recent technological advances in genomics, high-throughput sequencing, imaging, and other omics techniques have led scientists to a new biology era. Now, the scientific community has access to large biological datasets available for sophisticated analysis of complex biologic interactions. This rapid increase in biological data dimension is challenging conventional analysis strategies. With the advances in computational resources and computer science over the past few decades, bioinformatics is evolving as an integrative field between computer science and biology (ANGERMUELLER et al., 2016; AUSLANDER; GUSSOW; KOONIN, 2021; CHICCO, 2017; MA; ZHANG; WANG, 2014a).

The large amount of data introduced by the new technologies has made biologists rethink data analysis strategies and develop new tools to analyze the data. These datasets present the raw material to gain insights into biological systems and complex diseases, but higher-level analyses are needed to explore their potential (CAMACHO et al., 2018; MA; ZHANG; WANG, 2014a).

Machine learning has emerged as an attractive tool in biological research, particularly in the computational biology community. Being able to handle large datasets and extract information from their complexity interaction through accurate statistical models, ML provides next-level analyses allowing new perspectives and novel hypotheses about living systems (CAMACHO et al., 2018; CHICCO, 2017).

3.1 Machine Learning

Machine Learning is a multidisciplinary branch of artificial intelligence that incorporates computer science, statistics, and information theory to simulate human learning by exploring patterns in the data without the need to define them in

advance. ML algorithms can find functional relationships from data that underlying characteristics are unknown or undefined. As commonly described, machine learning is a field of study that gives computers the ability to learn without being explicitly programmed, which means that algorithms keep self-improving to enhance the performance of learning tasks (MA; ZHANG; WANG, 2014b).

Most ML applications involve four steps: data cleaning and pre-processing, feature extraction, model fitting, and evaluation (Fig. 3). Given one sample of data, it is customary to denote all features and covariates as *input* x and label it with its response variable or *output* value y. ML algorithms can be roughly divided into supervised or unsupervised learning (ANGERMUELLER et al., 2016).

Fig. 3: Machine Learning applications common workflow

Font: (ANGERMUELLER et al., 2016).

3.1.1 Supervised learning

Supervised learning (SL) algorithms are trained with labeled data, which means that for a given x, the respective known value y is provided to the model as well. After training, the model should be optimized to predict an unknown output for a provided input. SL problems can also be divided based on their output type values into classification (categorical) or regression (continuous). Examples of SL are Support Vector Machines (SVM), Regression, Random Forest, and Convolutional Neural Networks (CNN) (AUSLANDER; GUSSOW; KOONIN, 2021).

$$To\ \text{estimate:}\ f\ x\ = y \qquad x_1, y_1\ ,\ x_2, y_2\ , \ldots ,\ x_n, y_n\ \rightarrow input \qquad (1.1)$$

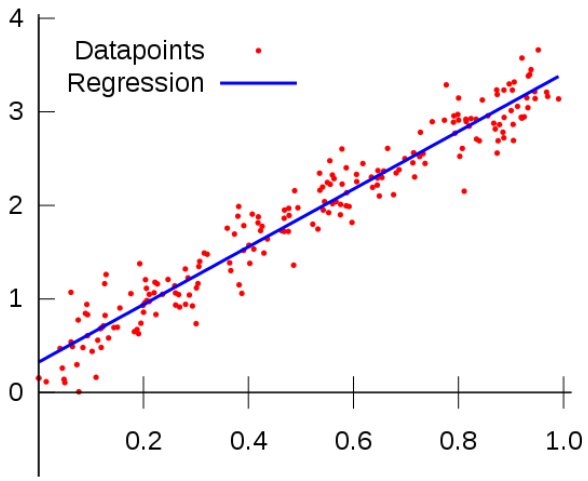
Linear methods

Linear regression is the most widespread machine learning method. The model is based on predicting a value from a linear combination of its given features (Fig. 4):

$$y\ w, x\ = w_0 + w_1 x_1 + \cdots + w_p x_p \qquad 1.2$$

In the equation above, y is the predicted value, $x_1, x_2, \ldots, x_n$ are the observed values from data, the vector $w = (w_1, \ldots, w_p)$ is composed of the optimized coefficients, and w_0 is the interception point. In one dimension, the model resumes to a line equation: $y = ax + b$. The machine goal is to assign the w vector values so that deviations of the real observations from the predict line are minimized (PEDREGOSA et al., 2011).

Fig. 4: Linear regression



Font: (FUMO, 2017).

Logistic regression is a supervised classification algorithm that models the probability of an observation belongs to one of two classes and classifies it by establishing a threshold (decision value).

$$P\ Y = 1|x\ = \frac{1}{\left[e^{-\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j\right)}\right]} \quad 1.3$$

Considering that the output $y = 1$ represents one of the classes, the model estimates that for a given value of x, the observation is related to this class ($y=1$). The purpose of the machine is to estimate the unknown parameters: $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ to classify the data better (LEMESHOW; HOSMER, 1982).

Linear methods can be modified to address more complex problems. However, they usually have the disadvantage of overfitting (Section 3.1.3) requiring regularization, which is a process to keep complex models simple and avoid overfitting) (LEMESHOW; HOSMER, 1982).

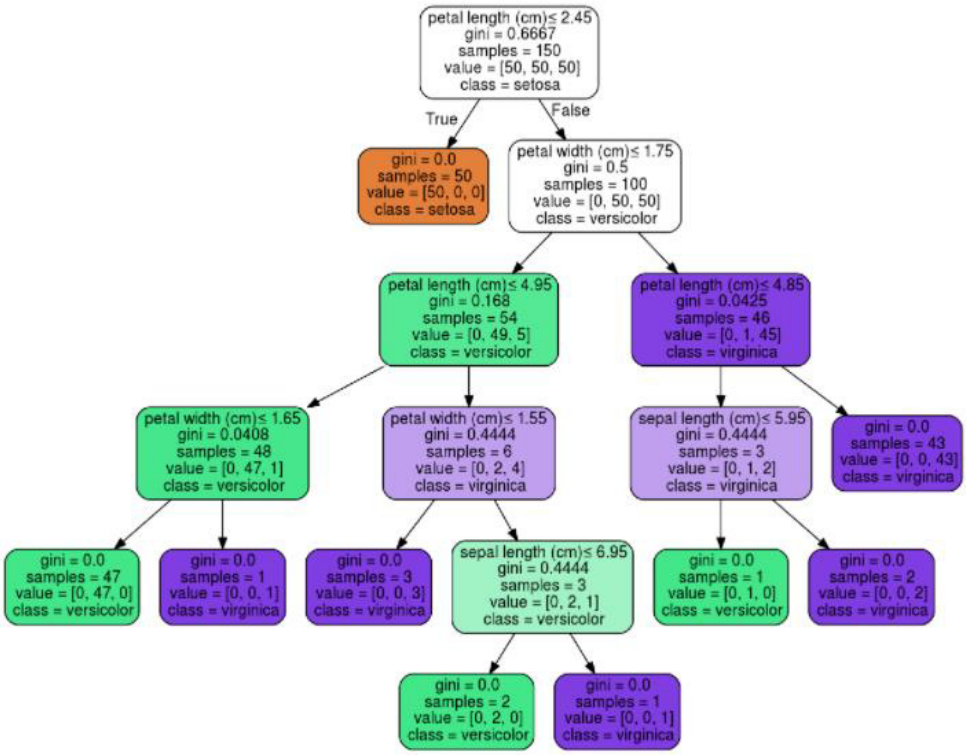
Decision Trees

The decision tree model is a non-parametric supervised learning method used for classification and regression represented by a tree-like structure, where each node has a binary decision criterion based on one or more parameters. The node tests the data and split it into two branches, where each branch represents one of the possible outcomes. The process happens until the stream reaches a leaf node in which the observation is finally classified (ARABNIA; TRAN, 2011).

Depending on which algorithm is applied, decision tree models may vary. For instance, the ID3 algorithm starts with the whole training data and chooses the best feature to use as a criterion to the root node (first node). Then, it splits the set into subsets. If all instances in the subset have the same classification, the process stops for that branch, and a leaf node is returned with that classification. Otherwise, if the subset contains multiple classifications and there are no more features to test, the algorithm will return the most frequent classification. If there are more features to test, the algorithm will recursively call itself again (ARABNIA; TRAN, 2011).

Random Forest (RF) (Fig. 5) model is an example of combining and training multiple decision trees to improve predictive performance and can efficiently be applied to complex and non-linear datatypes (ZHANG; MA, 2012).

Fig. 5: Decision Tree Example



Font: (PEDREGOSA et al., 2011).

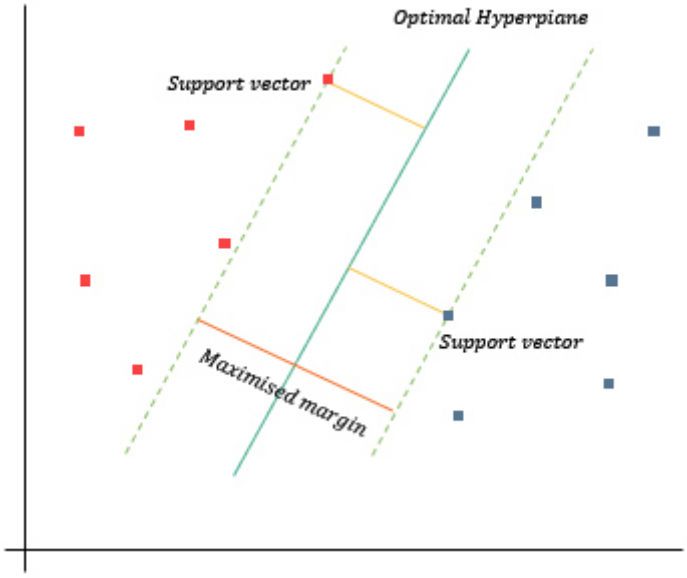
Support Vector Machines (SVM)

Support Vector Machine (SVM) is a supervised learning method initially used for classification, but it can also be modified to solve regression and outliers detection problems (Fig. 6). The algorithm creates a hyperplane in the feature space to linearly separate the observations of different classes. (PEDREGOSA et al., 2011).

Model training consists of optimizing the hyperplane so that the distance between its margins is maximized. The sample points that lay over the margins are the so-called Support vectors. To achieve its goal in non-linear data, the model also

employs kernel functions to implicitly transform the space, enabling the separation in the original feature space. (SCHÄFER; CIAUDO, 2020).

Fig. 6: Optimal Hyperplane and Support vectors



Font: Author.

SVM is an effective tool in high dimensional spaces, large datasets, and cases where the number of dimensions is greater than the number of samples, achieving high accuracy predictions (SCHÄFER; CIAUDO, 2020).

3.1.2 Unsupervised Learning

Unsupervised learning (UL) algorithms are used when the labels on the input data are unknown. These methods can identify hidden patterns in unlabeled data by themselves, without the need for output labels. Methods such as hierarchical clustering and principal components analysis (PCA) are used to cluster subsets of data or reduce its dimensions directly. Unsupervised techniques can be advantageous as a first step before training a supervised learning model, for example, reducing the number of relevant features.

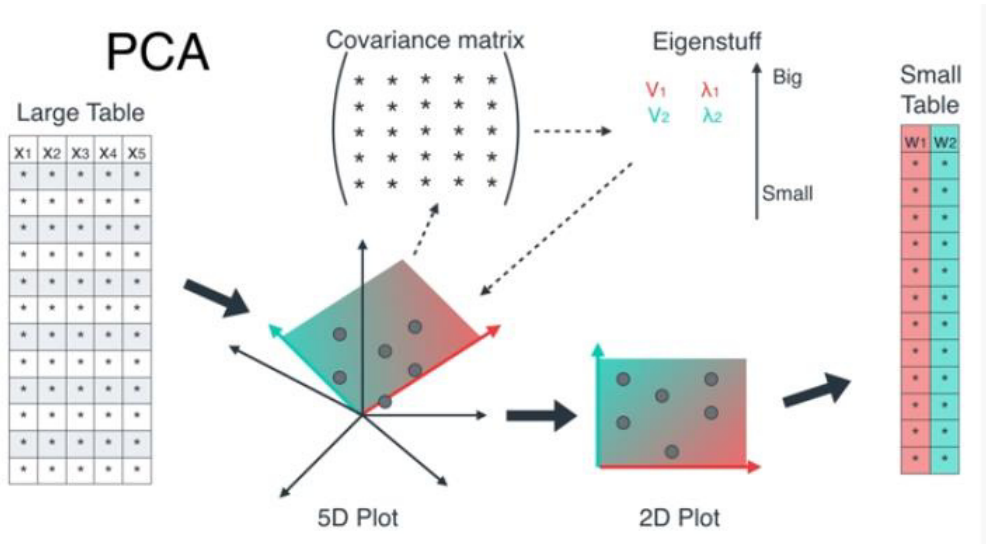
$$To\ \text{estimate:}\ f\ x\ = y \qquad x_1\ ,\ x_2\ ,\ldots,\ x_n\ \rightarrow input \qquad (1.4)$$

Principal components analysis (PCA)

Principal components analysis (PCA) is an unsupervised learning feature extraction procedure, which extracts a small set of directions (principal components; PCs) to represent the data and reduce its dimension (Fig. 7). It reveals underlying structures in data and derives a low-dimensional set of features from a large set of variables. PCA assumes that a small number of principal components can significantly explain the variance of a complete dataset.

PCA arises from quantifying the importance of each dimension for describing the variability of a data set. The algorithm results in a sorted array of new features such that the first principal component is the direction of the data which the observations vary the most. The output of PCA is often employed as an input of a supervised model in place of the entire set of input features. Additionally, PCA is commonly used to visualize complex datasets with multiple parameters (SHLENS, 2014).

Fig. 7 Principal Component Analysis (PCA)



Font: (SERRANO, 2019).

3.1.3 Deep Learning

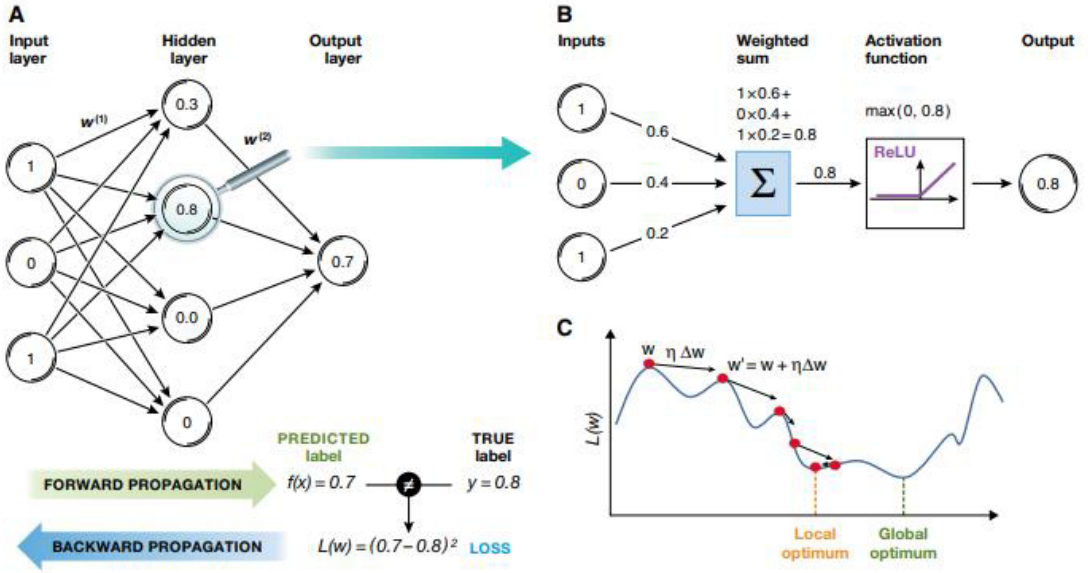
Deep learning (DL) is a subfield of Machine Learning. Different from ML classic algorithms, deep neural networks (DNN) can circumvent the manual extraction of features by learning them directly from data, which means that no pre-defining features need to be set based on prior knowledge. Moreover, DNN can capture non-linear dependencies at multiple genomic scales (ANGERMUELLER et al., 2016).

Artificial Neural Networks (ANN) consists of layers of interconnected compute units (neurons) (Fig. 8). The network receives data in an input layer, which are transformed in a non-linear way through multiple hidden layers, and finally computes an output in the output layer. Each neuron is connected to all neurons from the previous layer.

Each neuron receives multiple inputs, computes a weighted sum, and applies a non-linear function to calculate its output. The weights are the free parameters to be optimized to capture the models' representation. Learning minimizes a loss function, which is the difference between the output layer and the true label. This loss is backward propagated through the network to compute the gradients of the loss function to weights and update their value to move along the direction of steepest descent $\mathrm{d}w$ multiplied by a learning rate (ANGERMUELLER et al., 2016).

Alternative structures include convolutional neural networks (CNN), recurrent neural networks (RNN), and autoencoders (ANGERMUELLER et al., 2016).

Fig. 8: Artificial Neural Networks



Font: (ANGERMUELLER et al., 2016).

3.1.4 Data preparation

Training data is the key component to the success of every machine learning application. The model performance is directly related to the data quality, thus, efforts on collecting, labeling, cleaning, and normalizing data are worth it and can significantly improve the prediction accuracy (ANGERMUELLER et al., 2016; CHICCO, 2017).

The first critical point is the size of the dataset. Sufficient training data must be available to fit complex models. The ideal situation is to have at least ten times more data instances than data features (CHICCO, 2017).

The second critical point is the pre-processing of the dataset. Arranging the dataset as a good input implies randomly shuffling data instances to avoid any possible trend that can be incorrectly learned by the model, besides discarding all inconsistent, corrupted, and inaccurate data. Biological data is often collected from multiple experiments, which happen under different conditions. These conditions need to be taken into account when applying data from different sources to train

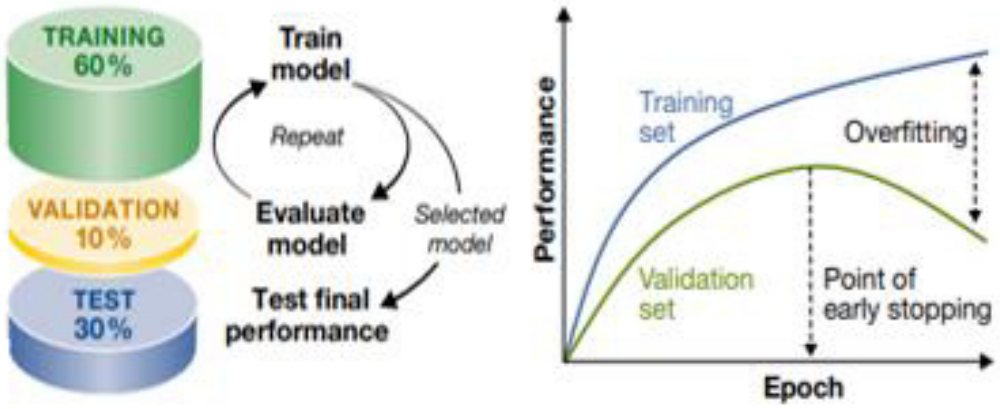
the algorithm. Normalization of numerical datasets can also help to accelerate the training and to put the whole dataset in a common frame (CHICCO, 2017).

Another frequent problem in computational biology is the imbalanced datasets. In an imbalanced dataset, one class is over-represented in relation to the other. As consequence, the machine may learn better how to recognize the over-represented classes, but it may have difficulty in classifying the data as the less-represented classes. To undergo this problem, the solution can be simply collecting more data, or removing data elements from the over-represented class. Other techniques to tackle the imbalanced data problem can be explored depending on the uniqueness of the data (CHICCO, 2017).

To assure that the model will be able to predict unseen data, ML models need to be trained, selected, and tested on independent data sets. A common practice is to split the data into 3 parts: 60% for training, which is used to learn different hyper-parameters[1], 10% for validation, which is used to assess these parameters, and 30% for a test, which is used to evaluate the model with the best performance on unseen data. Employing independent data sets helps to check for overfitting. Overfitting happens when the model learns intrinsic characteristics exclusively from the training data and loses its capacity of predicting new occurrences (Fig. 9). Comparing the model performance over the training set with the validation set can indicate overfitting, as shown in Fig 8 (ANGERMUELLER et al., 2016).

[1] Hyper-parameters of a machine learning algorithm are higher-level properties of the algorithm statistical model, which can strongly influence its complexity, its speed in learning, and its application results (ANGERMUELLER et al., 2016).

Fig. 9: (i) Training, Validation and Test model. (ii) Overfitting performance



Font: (ANGERMUELLER et al., 2016).

3.1.5 Evaluation the model

After training the model with the training dataset, evaluating its performance on a validation dataset is an essential step. For this task, there are several statistical scores to measure the model performance. For regression models, the root-mean-square error (RMSE) and the coefficient of determination (R^2) are standard scoring metrics (SCHÄFER; CIAUDO, 2020). For supervised binary classification problems, Matthews correlation coefficient (MCC) can score how well the classifier is doing on both the negative and positive elements with a value between -1 (worst value) and +1 (best value) (CHICCO, 2017).

$$MCC = \frac{TP \;.\; TN - FP \;.\; FN}{\sqrt{TP + FP \;\;.\;\; TP + FN \;\;.\;\; TN + FP \;\;.\;\; TN + FN}} \qquad 1.5$$

TP: True positive; TN: True negative; FP: False positive; FN: False negative

3.2 ML-based miRNA target prediction

Advances in next-generation sequence (NGS) technology have enabled improved methods to identify miRNAs. Elucidating how miRNAs regulate plants

responses is highly dependent on the recognition of their target molecules (SRIVASTAVA et al., 2014).

The development of miRNA target prediction algorithms can be based on characteristics of mRNA and miRNA-mRNA interactions or Machine Learning statistical inference (RIOLO et al., 2020).

Machine learning algorithms can be applied to model miRNA interactions and predict miRNA targets. To identify relevant interactions, ML techniques have been employed to build models based on experimental observations. It's important to note that these data may contain errors but, it's the ground truth of the Machine Learning model (SCHÄFER; CIAUDO, 2020).

ML techniques are trained using experimental data previously validated to match miRNA-mRNA interactions by distinguishing positive and negative examples from data. These predictions are based on biological features such as Seed Type, Binding Free Energy, Supplementary Binding, Target Site Conservation, Target Site Accessibility, Target Site Position, and Target Site Abundance. Positive data instances are collected from validated experiments with proven biological significance, while negative examples are usually generated artificially to fill the database (SCHÄFER; CIAUDO, 2020).

Examples of predictions tools based on Machine Learning are TargetScan (Linear regression), miSTAR (Logistic regression and random forest), MiRTarget (SVM), deepTarget (Recurrent Neural Network with Autoencoder) (SCHÄFER; CIAUDO, 2020).

Trained machines reported in literature usually combine different features to make predictions, resulting in methodologies with distinct limitations.

The best approach to obtain good results on predicting targets is applying a combination of tools that are derived from different predictions assumptions, thus ensuring a good balance of sensitivity and specificity (RIOLO et al., 2020).

Validating the selected putative miRNA from computational approaches includes four requirements: miRNA and mRNA need to be co-expressed in the same organism, they also need to specifically interact with each other, miRNA must affect protein expression, and biological function (RIOLO et al., 2020).

Although plant miRNA tools available are optimized to predict targets in Arabidopsis with high specificity, they may not perform well in non-Arabidopsis species, suggesting that non-conventional features of miRNA-mRNA interaction may exist and should be incorporated in next-generation of algorithms to improve target identification interactions (SRIVASTAVA et al., 2014).

3.3 Predicting plant stress and miRNA interactions

After discovering and isolating miRNA genes, machine learning can also be applied to investigate how miRNAs are expressed in stress conditions and their contribution to plant response toward different levels of plant stress. Asefpour (2020) demonstrated in his study the performance of machine learning as a promising tool to discover aspects of miRNAs' contribution to plant stress responses. He used machine learning to predict plant stress by having the plant miRNA expressions and investigated the contribution of each miRNA to the plant response by using feature selection algorithms. The study was conducted with plants under abiotic stress, but the results suggested that the same approach can be extended to biotic stress.

The main issue to understand miRNA functionality is the fact that miRNAs not only regulate immune response but also growth, reproduction, and other cellular activities. Moreover, they can interact with each other to enhance plant response. Distinguishing if the expression of one miRNA has significantly changed and the connection of these alterations with other miRNAs can be solved by Machine Learning algorithms.

Asefpour (2020) created a database with the concentrations of miRNAs when *Arabidopsis thaliana* was exposed to salinity, drought, cold, and heat. The selected miRNAs have been previous discovered to be involved in *Arabidopsis thaliana* stress response. A feature selection algorithm was used to determine which miRNAs had the most relevant alterations on their expression for each kind of stress. With this reduced number of features, supervised learning algorithms, such as Decision tree and Support Vector Machine, were trained to predict which stress response level corresponded to each combination of miRNAs concentration. SVM was able to predict the output with $R^2 = 0.96$ when measured the concentrations of miRNA-169, miRNA-393, and miRNA-396.

Asefpour (2020) was the first to use machine learning to predict stress using miRNAs expression and consequently identifying the most relevant miRNAs involved in each kind of stress response. ML is a potential alternative to understanding the contribution of each miRNA, which is time-consuming with experimental approaches, and thus, contributes to select key molecules to be a target of plant breeding genetics.

The results described above and with the knowledge about how biotic stress response work, it's highly suggested to new researchers to apply machine learning in biotic stress response experiments.

4 Conclusion

As described in this paper, Machine Learning techniques come as a powerful tool to improve our knowledge about plants and their interactions with the environment. They can significantly reduce the number of experimental assays in predicting miRNA by mitigating false positives and highlight the functional interaction between plant-produced miRNAs under stressful conditions by measuring their concentration levels. Both information about miRNA targets and their functional characterization are essential for the development of resistant transgenic plants.

Although ML techniques have not been intensively used to identify the role of miRNAs in biotic stress specifically, this review has shown the potential use in this area. Since most of the plant immune system relies on signaling paths that depend on transcriptional factors and NB-LRR protein, which miRNAs mediate their activation, finding the miRNAs targets and assessing miRNAs concentrations can reveal a lot about plant-pathogen interactions.

Choosing Machine Learning algorithms to study miRNA can deal with the way their data is available and their complexity. Prediction is based not only on genomic sequence complementation but on other biological features as free energy and conservation. Furthermore, miRNA characterization involves multiple correlations which the underlying patterns are unknown yet.

Machine Learning algorithms available up to now can explain complex data patterns, but they are highly dependent on data quality. ML models will naturally achieve better results as more experimental data become available and spread throughout the scientific community. More data about economically relevant plants need to be collected in order to refine their datasets and enable a more accurate performance when applying ML algorithms.

BIBLIOGRAPHY

ALI, M. et al. Biotic stress triggered small RNA and RNAi defense response in plants. **Molecular Biology Reports**, v. 47, n. 7, 19 jul. 2020.

ANGERMUELLER, C. et al. Deep learning for computational biology. **Molecular Systems Biology**, v. 12, n. 7, 29 jul. 2016.

ARABNIA, H. R.; TRAN, Q.-N. (EDS.). **Software Tools and Algorithms for Biological Systems**. New York, NY: Springer New York, 2011. v. 696

ASEFPOUR VAKILIAN, K. Machine learning improves our knowledge about miRNA functions towards plant abiotic stresses. **Scientific Reports**, v. 10, n. 1, 20 dez. 2020.

AUSLANDER, N.; GUSSOW, A. B.; KOONIN, E. V. Incorporating Machine Learning into Established Bioinformatics Frameworks. **International Journal of Molecular Sciences**, v. 22, n. 6, 12 mar. 2021.

CAMACHO, D. M. et al. Next-Generation Machine Learning for Biological Networks. **Cell**, v. 173, n. 7, jun. 2018.

CHAUDHARY, S.; GROVER, A.; SHARMA, P. C. MicroRNAs: Potential Targets for Developing Stress-Tolerant Crops. **Life**, v. 11, n. 4, 28 mar. 2021a.

CHAUDHARY, S.; GROVER, A.; SHARMA, P. C. MicroRNAs: Potential Targets for Developing Stress-Tolerant Crops. **Life**, v. 11, n. 4, 28 mar. 2021b.

CHICCO, D. Ten quick tips for machine learning in computational biology. **BioData Mining**, v. 10, n. 1, 8 dez. 2017.

DANGL, J. L.; JONES, J. D. G. Plant pathogens and integrated defence responses to infection. **Nature**, v. 411, n. 6839, jun. 2001.

DAVID FUMO. **Linear Regression — Intro To Machine Learning #6**.

DJAMI-TCHATCHOU, A. T. et al. Functional Roles of microRNAs in Agronomically Important Plants—Potential as Targets for Crop Improvement and Protection. **Frontiers in Plant Science**, v. 8, 22 mar. 2017.

DMITRIEV, A. P. Signal Molecules for Plant Defense Responses to Biotic Stress. **Russian Journal of Plant Physiology**, v. 50, n. 3, 2003.

GIANSANTI, V. et al. Comparing Deep and Machine Learning Approaches in Bioinformatics: A miRNA-Target Prediction Case Study. In: [s.l: s.n.].

GIMENEZ, E.; SALINAS, M.; MANZANO-AGUGLIARO, F. Worldwide Research on Plant Defense against Biotic Stresses as Improvement for Sustainable Agriculture. **Sustainability**, v. 10, n. 2, 2 fev. 2018.

JEYARAJ, A. et al. Utilization of microRNAs and their regulatory functions for improving biotic stress tolerance in tea plant [*Camellia sinensis* (L.) O. Kuntze]. **RNA Biology**, v. 17, n. 10, 2 out. 2020.

JONES, J. D. G.; DANGL, J. L. The plant immune system. **Nature**, v. 444, n. 7117, nov. 2006.

KHRAIWESH, B.; ZHU, J.-K.; ZHU, J. Role of miRNAs and siRNAs in biotic and abiotic stress responses of plants. **Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms**, v. 1819, n. 2, fev. 2012.

KULSHRESTHA, C. et al. Elucidating micro RNAs role in different plant–pathogen interactions. **Molecular Biology Reports**, v. 47, n. 10, 9 out. 2020.

KURUBANJERDJIT, N. et al. Prediction of microRNA-regulated protein interaction pathways in Arabidopsis using machine learning algorithms. **Computers in Biology and Medicine**, v. 43, n. 11, nov. 2013.

LEMESHOW, S.; HOSMER, D. W. A REVIEW OF GOODNESS OF FIT STATISTICS FOR USE IN THE DEVELOPMENT OF LOGISTIC REGRESSION MODELS1. **American Journal of Epidemiology**, v. 115, n. 1, jan. 1982.

LI, Y. et al. Multiple Rice MicroRNAs Are Involved in Immunity against the Blast Fungus *Magnaporthe oryzae*. **Plant Physiology**, v. 164, n. 2, fev. 2014.

LIANG, C. et al. Artificial microRNA-mediated resistance to cucumber green mottle mosaic virus in Nicotiana benthamiana. **Planta**, v. 250, n. 5, 6 nov. 2019.

MA, C.; ZHANG, H. H.; WANG, X. Machine learning for Big Data analytics in plants. **Trends in Plant Science**, v. 19, n. 12, dez. 2014a.

MA, C.; ZHANG, H. H.; WANG, X. Machine learning for Big Data analytics in plants. **Trends in Plant Science**, v. 19, n. 12, dez. 2014b.

NAVARRO, L. et al. A Plant miRNA Contributes to Antibacterial Resistance by Repressing Auxin Signaling. **Science**, v. 312, n. 5772, 21 abr. 2006.

NÜRNBERGER, T.; KEMMERLING, B. Pathogen-Associated Molecular Patterns (PAMP) and PAMP-Triggered Immunity. In: **Annual Plant Reviews online**. Chichester, UK: John Wiley & Sons, Ltd, 2018.

PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

RIOLO, G. et al. miRNA Targets: From Prediction Tools to Experimental Validation. **Methods and Protocols**, v. 4, n. 1, 24 dez. 2020.

SCHÄFER, M.; CIAUDO, C. Prediction of the miRNA interactome – Established methods and upcoming perspectives. **Computational and Structural Biotechnology Journal**, v. 18, 2020.

SERGEANT, K.; RENAUT, J. Plant Biotic Stress and Proteomics. **Current Proteomics**, v. 7, n. 4, 1 dez. 2010.

SHLENS, J. A Tutorial on Principal Component Analysis. **Google Research**, abr. 2014.

TANG, J.; CHU, C. MicroRNAs in crop improvement: fine-tuners for complex traits. **Nature Plants**, v. 3, n. 7, 30 jul. 2017.

THUDI, M. et al. Genomic resources in plant breeding for sustainable agriculture. **Journal of Plant Physiology**, v. 257, fev. 2021.

ZHANG, C.; MA, Y. (EDS.). **Ensemble Machine Learning**. Boston, MA: Springer US, 2012.