ISSN: EN TRÁMITE
https://doi.org/10.59741/p05c6793

Identification of Plants miRNAs from deep RNA-seq data using a multilayer perceptron

Identificación de microRNAs de plantas a partir de datos de secuenciación profunda de RNA empleando un perceptrón multicapa

Recibido:
04/11/2024
Aceptado:
18/03/2025
Publicado:
19/03/2025

*Marco A. Juárez-Verdayes† and Javier Montalvo-Arredondo†**

Molecular Bioengineering and Bioinformatics Laboratory. Departamento de Ciencias BásicasUniversidad Autónoma Agraria Antonio Narro, Calzada Antonio Narro 1923, CP 25315. Buenavista, Saltillo Coahuila
Marco A. Juárez-Verdayes, orcid: https://orcid.org/0000-0002-3948-4036
Javier Montalvo-Arredondo, orcid: https://orcid.org/0000-0003-4651-5387
† These authors contributed equally.

*corresponding author:
buitrejma@gmail.com

Abstract

Micro-RNA (miRNA) – mediated transcript degradation is a layer of gene regulation at the post-transcriptional level that has important roles in plants. Some traits of plants that are of interest to the food industry are tightly regulated by this molecular mechanism. In México, Fabaceae – family plants represent one of the main food sources. Accordingly, it is important to study this layer of regulation to improve crop and seed production yields, nonetheless, one of the pressing concerns is the miRNAs loci identification. The basic and ancillary criteria, sometimes are not enough evidence for identifying miRNA loci. Artificial intelligence (AI), such as convolutional neural networks (CNN), have shown excellent predictive performance in identifying miRNAs loci, however, some of these CNN are complex and difficult to train and run. A multi-layer perceptron (MLP) model has been proposed for identifying pre-miRNAs sequences; it processes 180 feature information, however, the analysis is limited by the feature calculation, because it is computationally intensive. In this work, we proposed the use of AI based on a multi-layer perceptron (MLP) model which isn't complex and easy to train, we also proposed the use of k-mer frequencies to extract information from nucleotide and secondary-structure representation sequences. We tested several features of MLP models such as activation functions between layers and the number of dropout layers. The best-fitted models showed 84-90% of sensitivity and 98 to 100% of specificity when they were evaluated with testing datasets. We tested the predictive performance of the best-fitted models on real deep RNA-seq data. In conclusion, in this paper, we present an MLP-based AI capable of identifying pre-miRNAs sequences from the Fabaceae family plants using deep RNA-Seq data, these AIs showed sensitivity values of 80-85% and specificity values of 90-95%.

Resumen

La degradación de transcritos mediada por micro-RNAs (miRNAs) es una capa de regulación génica a un nivel post-transcripcional que desempeña funciones importantes en plantas. Algunos rasgos fenotípicos que son de interés para la industria alimenticia están fuertemente regulados por este mecanismo molecular. En México, plantas de la familia Fabaceae representan una de las principales fuentes de alimentación. En consecuencia, es importante estudiar esta capa de regulación para mejorar los cultivos y la producción de semillas, sin embargo, uno de los tópicos más difíciles es la identificación de los loci miRNAs. Los criterios básicos no son evidencia suficiente para la identificación correcta. Recientemente, inteligencias artificiales (IA) basada en redes convolucionales (CNN) han mostrado un excelente poder predictivo en la identificación de los loci miRNAs, sin embargo, algunas de estas CNN profundas son complejas y difíciles de entrenar y ejecutar. Se ha propuesto un modelo basado en la arquitectura de perceptrón multicapa (MLP) para la identificación de loci de miRNAs, sin embargo, su rendimiento es limitado debido a las altas capacidades de cálculo necesarias para procesar las secuencias. En este artículo mostramos como las IA basadas en modelos simples (MLP) son una opción más ágil y fácil de entrenar, Esto, debido en parte, a el uso de frecuencias de k-meros para extraer información de las secuencias de nucleótidos y de la representación de la estructura. En este artículo evaluamos diferentes características de los modelos MLP como, funciones de activación, y capas "dropout". Los modelos más adecuados mostraron una sensibilidad del 84-90% y una especificidad del 98-100% cuando fueron sometidos a prueba con los datos de evaluación. Adicionalmente, evaluamos los modelos con secuencias de transcritos ensambladas y obtuvimos valores de sensibilidad del 80-85% y especificidad del 90-95%.

ISSN: EN TRÁMITE
https://doi.org/10.59741/p05c6793

Key words:

Fabaceae family, Artificial intelligence, microRNAs, Gene regulation in plants, Post-transcriptional regulation.

Palabras clave:

Familia Fabaceae, Inteligencia artificial, microARNs, Regulación génica de plantas, Regulación post-transcripcional.

INTRODUCTION

Plants life success and the development of traits of agronomic interest are tightly related to precise gene regulation. Several layers of regulation are implicated in these phenomena, however, micro-RNAs (miRNAs)-mediated posttranscriptional regulation has caught the attention because they fine-tune the expression of several gene families that have roles in huge and diverse plant phenomena like pathogen resistance (Yang *et al*., 2021), abiotic stress tolerance (Zhang *et al*., 2022), plant-environment interactions (Song *et al*., 2019), traits development and homeostasis (Dong, Hu and Zhang, 2022). In México, *Fabaceae* – family plants such as *Phaseolus vulgaris* (common bean), have gained interest in agriculture and the food industry because crop deployments and seed production represent one of the main food sources (Centeno-González *et al*., 2021; Shavanov 2021). To improve the yield of crops and the production efficiency of these food sources, it is important to understand the regulatory process during plant development. We think studying the regulatory process mediated by miRNAs in *Fabaceae* plants, could unveil important features of the developmental process and other functional roles that are essential in plants, nonetheless, one of the pressing concerns in studying miRNA regulation, is the miRNA loci (MIR genes) identification and annotation (Meyers, 2008; Rojo-Arias and Busskamp, 2019)

During miRNA canonical biogenesis, DNA-dependent RNA polymerase II (DNA Pol II) is responsible for transcribing the MIR genes that could be coding or non-coding genes, generating transcripts called primary transcripts (pri-miRNAs) which are consecutively processed, mainly by, the endoribonuclease DICER-LIKE 1 (DCL1) enzyme, Serrate (SE) and Hypnoatic leaves 1 (HYL1) cofactors producing the pre-miRNAs hairpins/stem-loop which are translocated from the nucleus to the cytoplasm to be cleaved by DCL1 again to form the miRNA/miRNA* duplex that is methylated by Hua Enhancer 1 (HEN1). Then, the RNA-induced silence complex (RISC) formed together with the 20-24 length mature miRNAs and the ARGONAUTE carrier protein (AGO1). This complex is responsible for targeting transcripts and recruiting nucleases to degrade them or inhibit the translation process (Gangadhar *et al.*, 2021; Zhang *et al*., 2022).

MIR genes are extremely challenging to locate using basic criteria such as the presence of the transcript and the prediction of the stable hairpin formation, and it is because there is a massive quantity of transcripts that form stable hairpins and they aren't miRNAs (Kozomara and Griffiths-Jones, 2010; Rojo-Arias and Busskamp, 2019). Ancillary criteria have been proposed such as conservation between species, target identification, dicer (DCL1) dependence, RNA-dependent RNA polymerase and Polymerases IV/V independence, but they aren't strong evidence in identifying a MIR locus. Evolutionary relationships are strong evidence that can be used to identify MIR loci, however, this approach avoids identifying species-specific MIR loci that aren't conserved between species (Meyers, 2008). To understand this complex layer of regulation, it is important to correctly identify the MIR loci that encode for miRNAs in sequenced genomes.

ISSN: EN TRÁMITE
https://doi.org/10.59741/p05c6793

Several approaches have been proposed to solve this problem, and those that have succeeded are based on convolutional neural networks (CNN). Several works have shown that these artificial intelligence are capable of identifying miRNAs transcripts and MIR loci with high accuracy (Cha *et al*., 2021; Zheng *et al*., 2019; Zhang *et al*., 2024). Some CNNs are complex neural networks, they are composed of several convolutional and pooling layers, and several kernels that preprocess the input data, by a process called convolution to automatically extract different small features or patterns which are passed to a downstream feed-forward neural network also known as multi-layer perceptron to get the probability of predicting if something belongs to a category (Krizhevsky, Sutskever and Hinton, 2012). It is also known that the accuracy of a neural network is related to its size; the bigger it gets, the more accurate it becomes. To build an accurate CNN, it has to be complex and big, and this becomes computationally expensive (Zhao *et al*., 2021). In addition, MLP models have flexibility in data inputs and have a high capacity in recognizing more abstract patterns. Lokuge and coworkers (2022), have proposed an MLP model that processes 180 sequential, structural and thermodynamic features for plant pre-miRNA identification, however, the analysis is constrained for the feature calculation because it is computationally intense.

Simple multi-layer perceptron model (MLP) is less complex than deep CNN and it doesn't have to be deep to get accurate predictions, it is composed of the input layer, several hidden layers, and the output layer. This architecture is equal to the last part of a CNN (Krizhevsky, Sutskever and Hinton, 2012). Thus, in this work, we proposed the use of artificial intelligence (AI) based on the MLP architecture, but instead of using a convolution process to extract patterns from input data, we proposed the use of k-mer frequencies from nucleotide and secondary-structure representation sequence, to extract useful patterns from input data to make predictions on the assembled sequences of deep-sequencing RNA-seq data. This information extraction method isn't computationally expensive.

With this in mind, we built five models (Daphne, Emerald, Florence, Greece and Hilda) based on the MLP architecture, but with different characteristics related to the number of dropout layers and activation functions between layers and in the output layer. They are simple neural networks composed of an input layer that receives input data from putative pre-miRNA assembled sequences, 4 hidden layers (500, 250, 100, 50 neurons) and the output layer of one neuron, if it turns on, it predicts a miRNA, otherwise it turns off. These models were trained in a modest PC with a 4-core 3GHz Ryzen 3 processor and 10Gb of RAM; no special devices, such as GPUs or TPUs were used in this work. The models' predictive accuracy we obtained in this work was similar to the accuracy obtained with CNN reported in (Zheng *et al*., 2019). However, these MLP models have the advantage of simplicity and compactness therefore, they are easy to train, tune and run, even in low-capacity devices. Also, they have flexibility in input data processing and can learn more abstract patterns. In this work, we also evaluated the effect of the models' individual output combinations by average, on models' predictive performance and we observed that some combined.

ISSN: EN TRÁMITE
https://doi.org/10.59741/p05c6793

MATERIALS AND METHODS

1. Datasets preparation

Datasets were constructed with sequences obtained from two sources, (1) unspliced transcript sequences from species belonging to the *Fabaceae* family downloaded from NCBI datasets (https://www.ncbi.nlm.nih.gov/datasets/), and (2) with hairpin pre-miRNA sequences already reported in Plants miRNA Encyclopedia (Guo *et al*., 2020) for those *Fabaceae* species. We build 150 datasets for training models; and 40 datasets for testing.

To build the datasets, we randomly sampled subsequences of size of 60 to 200 nucleotides from unspliced transcripts obtained from reported transcriptomes for Fabaceae plant species (*Glycine max* GCA_000004515.5, *Cicer arietum* GCA_000331145.1, *Cajanus cajan* GCA_000340665.2, *Phaseolus vulgaris* GCA_000499845.2, *Arachis ipaensis* GCA_000816755.2, *Arachis duranensis* GCA_000817695.3, *Arachis hypogea* GCA_003086295.2, *Medicago truncatula* GCA_003473485.2, *Vigna unguiculata* GCA_004118075.2, *Glycine soja* GCA_004193775.2, *Lotus japonicas* GCA_012489685.2, *Arachis stenosperma* GCA_014773155.1, *Vigna umbellate* GCA_018835915.1, *Pisum sativum* GCA_024323335.2). We randomly sampled 4450 – 4480 sequences 150 times for the training datasets and 40 times for the testing datasets. The hairpin pre-miRNA sequences were downloaded from the Plant miRNA Encyclopedia database, where the miRNA loci were annotated by employing in silico predictions and experimental evidence from small RNA-seq and Parallel Analysis of RNA ends (PARE-seq) data (Guo *et al*., 2020). The hairpin pre-miRNA sequences dataset was split into two datasets, one of them with 80% (3813) of the sequences for training and the other with 20% (953) of the sequences for testing. For each training dataset, we concatenated the hairpin pre-miRNA sequence dataset destined for training, and for each testing dataset, we also concatenated the pre-miRNA sequences dataset destined for testing. The randomly sampled subsequences were labeled as negative pre-miRNA sequences, and the sequences of pre-miRNA hairpins were labeled as positive sequences during training.

To transform the information from sequences to multi-layer perceptron input data, the 5-mers frequencies of the nucleotide sequences and two-dimensional structure-sequence representation calculated by RNAfold of ViennaRNA package (Lorenz *et al*., 2011), percentage of guanine and cytosine, minimum free energy and minimum free energy divided by sequence length were calculated. These data were used during training for discriminating between true and false pre-miRNAs sequences. All dataset values were normalized using the Min-Max method where original data are linearly transformed (Henderi *et al*., 2021). Training and testing datasets are available at figshare_link.

2. K-mers frequencies calculation for nucleotide and secondary structure sequence

We computed the total possible k-mers of 5 nucleotides composed by an alphabet of 4 letters [“A”, “C”, “G”, “T”] representing the nucleotides, then the count of the presence of each k-mer on sequence is calculated using a sliding-window algorithm. Then, each k-mer count is divided by the total count of all possible k-mers for each analyzed sequence. We applied a similar approach to process the information of the secondary structure. In that case, the secondary structure is represented by 3 letters [“.”, “(“, “)”]. The dot means that the nucleotide doesn't match any other nucleotide, left parenthesis and right parenthesis means that these nucleotides are aligned in the secondary structure. So the sequence that represents the secondary structure was processed in a similar way to the nucleotides k-mers but in this case, we used 5-mers conformed with letters of an alphabet of 3 letters.

ISSN: EN TRÁMITE
https://doi.org/10.59741/p05c6793

3. Model evaluation methods

The metrics we used to evaluate the models were Sensitivity, Specificity, Positive Predictive Value (PPV) and Negative Predictive Value (NPV). Sensitivity can be defined, in this case, as the probability of identifying positive results on a population of sequences that are known to be miRNAs, and specificity is the probability of correctly identifying negative results on sequences population that are not miRNAs (Trevethan, 2017).

$$\text{sensitivity} = \frac{TP}{(TP+FN)} \qquad \text{Eq. 1}$$

$$\text{specificity} = \frac{TN}{(TN+FP)} \qquad \text{Eq. 2}$$

Meanwhile, PPV and NPV measure the likelihood if a given sequence encodes or doesn't encode for a miRNA, respectively (Trevethan, 2017).

$$\text{PPV} = \frac{TP}{(TP+FP)} \qquad \text{Eq. 3}$$

$$\text{NPV} = \frac{TN}{(TN+FN)} \qquad \text{Eq. 4}$$

We also measured the accuracy as follows.

$$\text{accuracy} = \frac{(TP+TN\,)}{(TP+TN+FP+FN)} \qquad \text{Eq. 5}$$

Where TP means true positives, TN, true negatives, FP, false positives and FN false negatives.

4. Multi-layer perceptron model building, training and testing

Five multi-layer perceptron model architectures were constructed (Daphne, Emerald, Florence, Greece and Hilda) with an input layer of 1271 containers receiving sparse data, four hidden dense-connected layers with 500, 250, 100 and 50 neurons each layer. To avoid data overfit ting, two dropout layers set to 20% were added between the first and second hidden layers for Daphne, Emerald and Florence. Dropout layers of 20% were set between all hidden layers for the Greece model, and Hilda was constructed without any dropout layer between hidden layers. We tested different strategies of neuron activations in hidden layers; for the Emerald model we used sigmoid activation functions, for Florence, Greece and Hilda we used the hyperbolic tangent activation functions. In the case of Daphne, we did not use any activation function in the hidden layers. The output layer contained one neuron and was activated with the sigmoid function for every model (see Table 1).

Table 1. Architecture features of MLP models

	Daphne	Emerald	Florence	Greece	Hilda
Act. function	None	Sigmoid	Tanh[a]	Tanh	Tanh
Dropout	2 layers[b]	2 layers	2 layers	Full[c]	None
Output act. function	Sigmoid	Sigmoid	Sigmoid	Sigmoid	Sigmoid

a. Hyperbolic tangent activation function.
b. Two dropout layers were set at 20% after hidden layer 1 and 2.
c. Full means that this model has dropout layers set at 20% after all hidden layers.

ISSN: EN TRÁMITE
https://doi.org/10.59741/p05c6793

During training, the Adaptive Moment Estimator (Adam) algorithm was used as the optimizer, Binary Cross Entropy was used as the Loss function, and Binary Accuracy was used as a metric to monitor the training. Hyperparameters such as learning rate and weight rate decay were set to $1x10^{-4}$ and $1x10^{-7}$, respectively.

Adam optimizer.

$$v_t = \beta_1 * v_{t-1} - (1-\beta_1) * g_t \qquad \text{Eq. 6}$$

$$s_t = \beta_2 * s_{t-1} - (1-\beta_2) * g_t^2 \qquad \text{Eq. 7}$$

$$\Delta w_t = -n \frac{v_t}{\sqrt{s_t + \epsilon}} * g_t \qquad \text{Eq. 8}$$

$$w_{t+1} = w_t + \Delta w_t \qquad \text{Eq. 9}$$

Where w_{t+1} is the weight for the next iteration, w_t is the current weight, n is the learning rate, Δw_t is the amount of change of weight at time (t), v_t is the exponential average of the gradients, s_t is the exponential average of the gradients raised to the square, and β_1 and β_2 are hyper-parameters set to 0.9 and 0.99 respectively.

$$BCE = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log\big(p(y_i)\big) + (1-y_i) \cdot \log(1-p(y_i)) \qquad \text{Eq. 10}$$

Binary Cross Entropy Loss function.

We performed a mini-batch gradient descent learning with a batch size of 64 and 5 epochs with 5 dataset randomization. For each dataset randomization, 80% of it, was used for training, and the remaining 20% was used for validation. The new model's parameter values were saved if the Binary Accuracy value calculated on validating data was higher than the same metric calculated with the previously saved model's parameters. We iterated this procedure until loss and validation loss values remained unchanged using the 150 training datasets. Model testing was done with new 40 datasets and the best model was selected. To use these models, we calculated the predictions from individual models and the results were rounded. If the result is equal to 1, the transcript was predicted to encode a miRNA, otherwise it was discarded. Model building, training and testing were done with Numpy, Pandas and Tensorflow + Keras Python libraries and with the required dependencies. The training was done on a PC machine with a 3GHz Ryzen 3 processor and 4Gb of RAM.

ISSN: EN TRÁMITE
https://doi.org/10.59741/p05c6793

5. Deep-sequencing RNA data processing

Deep-sequencing RNA data (dsRNA) from *Phaseolus vulgaris* (Accession number SRP074456) and *Medicago truncatula* (Accession number SRP000631) were downloaded from the NCBI SRA archive, accessing them by the links from Plant's miRNA Encyclopedia (Guo et al., 2020). Reads' quality was evaluated with FastQC (Andrews, 2010), reads were filtered out by quality value lower than 24 (for phred33 encoding) and adapters were also cut with Trimmomatic (Bolger, Lohse and Usadel, 2014) with default parameters except for MINLEN which was set to 10. Reads mapping was done with Hisat2 (Kim et al., 2019) using the reference genome sequences of *Phaseolus vulgaris* G19833 (Accession: GCA_000499845.2) and *Medicago truncatula* Jemalong A17 (Accession: GCA_003473485.2). The SAM output files were sorted, indexed and transformed to BAM format with Samtools utilities (Li et al, 2009). BAM files were used to perform an assembly with Python custom scripts (simpleAssembler.py https://github.com/exseivier/mlp-fabaceae/SCRIPTS) and the Pysam Python module. After assembly, sequences were extracted from the GTF output file with Gffread (Pertea and Pertea 2020). These sequences were used in downstream analysis with the multi-layer perceptron models to predict if the assembled sequences encode or don't encode for a miRNA.

RESULTS AND DISCUSSION

1.Multi-layer perceptron model is a computationally streamlined option

Despite the challenging in identifying loci that encode miRNAs in a genome of interest, artificial intelligence based on neural networks has shown good predictive performance. Convolutional neural networks (CNN) have been chosen because they can extract complex miRNAs' features automatically from a sole nucleotide sequence, and they avoid depending on the calculated features such as minimum free energy, sequence length, secondary structure and base composition (Zheng *et al*., 2019; Zhang *et al*., 2024). Convolutional Neural Networks' predictive accuracy, sensitivity and specificity values ranged between 92 to 97, 87 to 97 and 97 to 100% respectively (Cha *et al*., 2021; Zheng *et al*., 2019; Zhang *et al*., 2024). Nonetheless, some of these neural networks are complex because they have several convolutional, kernel and pooling layers where input data is preprocessed, and then the information extracted from the convolution process is flattened and passed through a multi-layer feed-forward neural network also known as multi-layer perceptron (Krizhevsky, Sutskever and Hinton, 2012). Predictive accuracy is highly related to the complexity and size of the neural network; the bigger and more complex it is, the more accurate it is. But, it is also known that big and complex neural networks are difficult to train because they are computationally expensive (Zhao *et al*., 2021).

Another way to obtain information and extract complex features from nucleotide sequences is to use k-mer frequencies. These frequencies were mainly used in alignment-free sequence comparison methods and represent the nucleotide composition on a given sequence. If the size of the k-mer increases, the frequency representation becomes more specific (Zielezinski *et al*., 2017). In this work, we hypothesized that an artificial intelligence based on a simple multi-layer perceptron (MLP) model, would be a streamlined option because its architecture isn't complex and it hasn't to be big to get a good predictive performance (Zhao *et al*., 2021). We also think that k-mer frequencies can be used as input data, and we also dare to propose the use of the sequence information of the secondary-structure representation calculated with RNAfold, and processed as k-mer frequencies as input data. We think we can extract base composition and secondary structure patterns from these input data.

N_k^l∈ {W_1,W_2,…,W_n }   Eq. 11

The expression above shows the total possible k-mers that can be formed for nucleotide sequence and secondary structure sequence where "l" is the alphabet length (4 for nucleotides, and 3 for secondary structure), k stands for the size of the k-mer and n is equal to l^k. For this MLP architecture, we built the input layer as follows: k-mer frequencies for nucleotide sequences (45= 1024 k-mers) were concatenated with additional information (sequence length, guanine and cytosine percentage, minimum free energy (MFE) and MFE / sequence length) and the k-mers frequencies of secondary structure (35=243 k-mers). The whole input layer sums 1271 containers, that were consecutively dense-connected to the 4 hidden layers as described in materials and methods.

To this end we built five neural networks based on a multi-layer perceptron architecture (Daphne, Emerald, Florence, Greece and Hilda) (see Table 1). Results for the training process are shown in (Figure 1). It is easily observed, in almost all models (Figure 1A-C, 1E), except for Greece (Figure 1D), the diminishing of the Loss value and the increment of the Binary Accuracy value for training (blue line) and validating (red line) datasets along the training process that were around from 5k to near 7k of iterations. It took, in average, 30 minutes to reach that level of training in a PC with a 3GHz Ryzen 3 processor with 4 cores and 10Gb of RAM. No graphic card was used in this work.

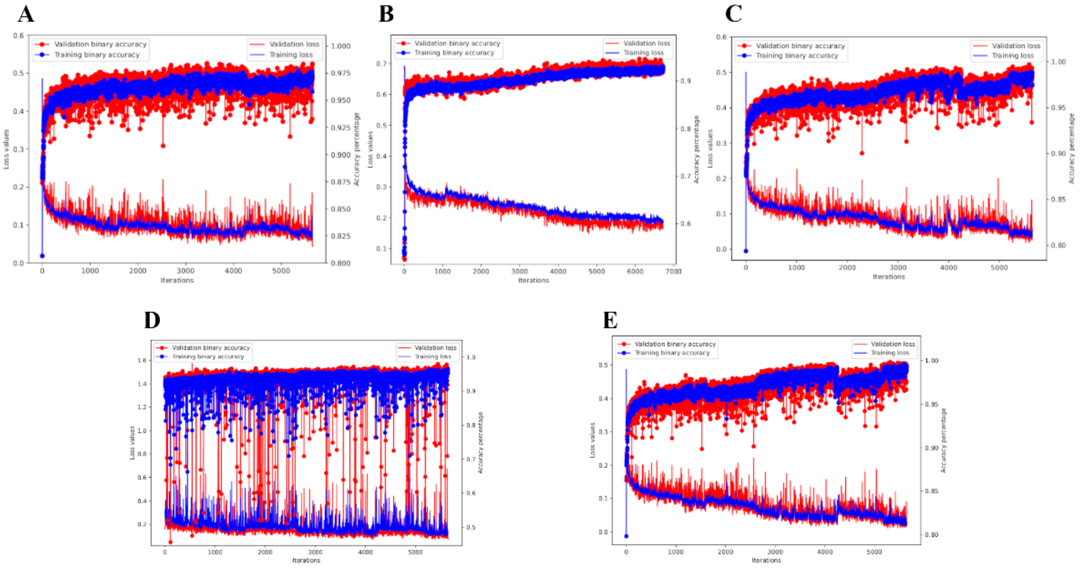

Figure 1. Models' behavitor development along training. . In this figure we show the line plots, that monitor the evolution of model's predictive performance along the training iterations/epochs for (A) Daphne, (B) Emerald, (C) Florence, (D) Greece and (E) Hilda. Red lines show the validation Loss calculated with validating data, and blue lines show the training Loss calculated with training data. Lines with dots represent the Binary Accuracy calculated with training data (blue) and validating data (red).

The performance development during training for Daphne, Florence and Hilda (Figure 1A, 1C, 1E) was similar between them, it was much less chaotic compared to Greece's (Figure 1D) development behavior, although, it was moderately chaotic compared to Emerald (Figure 1B). This result can be explained because Daphne doesn't use any activation function in the hidden layers and Florence and Hilda use hyperbolic tangent activation function in the hidden layers which permits calculating parameters with negative or positive values. Unlike all these models, Emerald uses the sigmoid function in hidden layers that allows calculating only positive values from 0 to 1. We think this feature leads to a less chaotic training process in the case of Emerald. The pattern of behavior for the Greece development can be explained by the excessive use of dropout layers that can introduce bias and uncertainties in the model's predictive performance during training. Nevertheless, the chaotic history of the model development along the training is not related to the final model performance. Finally, we observed no data overfitting during training from the beginning to the end, so that means the Loss values calculated with validating datasets didn't dramatically diverged from Loss values obtained with the training datasets.

ISSN: EN TRÁMITE
https://doi.org/10.59741/p05c6793

To monitor and test the model's predictive performance we saved the models' parameters if they produced a lower validating Loss and higher validating Binary Accuracy values than these values calculated with the previous model's parameters. So we saved a new model each time it happened during the entire training process. These saved models were used to test the predictive accuracy with new datasets the models never saw during training.

2. Models Testing

We evaluated the predictive performance for each saved model's parameters for every model architecture with 40 new datasets. So that means we got 40 different testing Binary Accuracy values for each model's parameters tested. For each model, we plotted the horizontal boxplots of these values. In the ordinate axis, we plotted the Binary Accuracy of the saved model's parameters during training sorted from lowest to highest value, and in the abscise axis, we plotted the Binary Accuracy obtained with the testing datasets (Figure 2). For almost all models, except for Greece (Figure 2D), we observed an expected behavior, it is a positive correlation between the Binary Accuracy observed during training and the Binary Accuracy obtained during testing, and the saturation effect observed at higher values of training and testing Binary Accuracy. These curved shapes observed in (Figure 2), for some models, represent the real continued improvement of the models along the training, until they reached a point where they couldn't improve any more. Testing Binary Accuracy ranged between 89 to 98% which are excellent predictive values for a mathematical model.

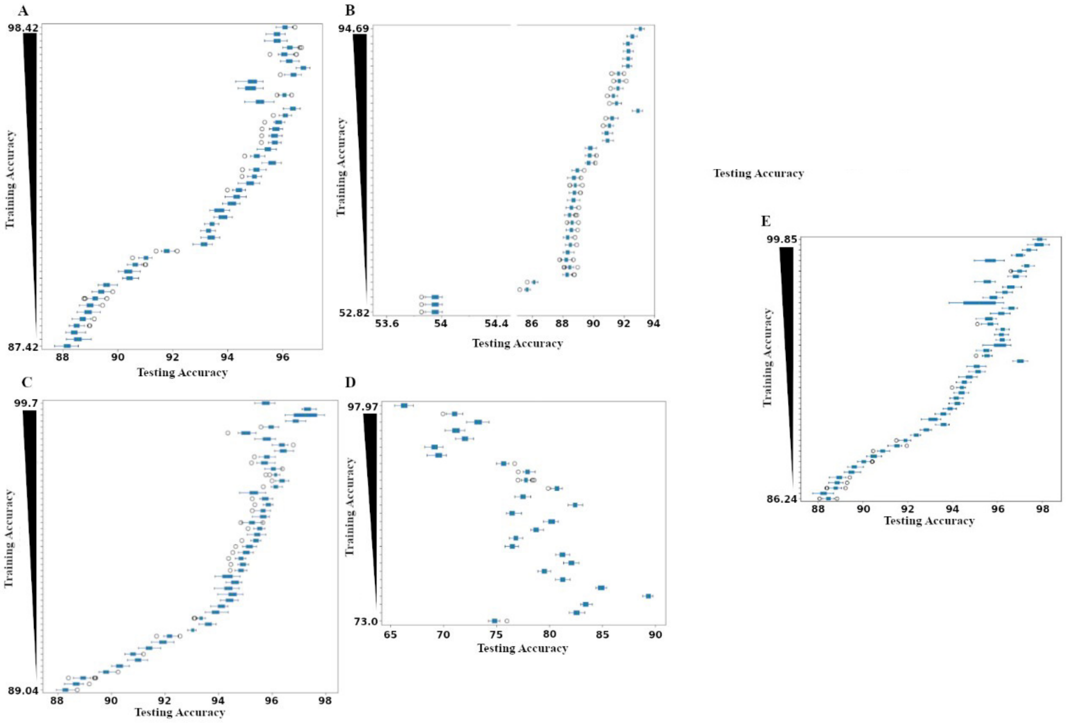

Figure 2. Testing predictive performance. In this figure we show the results of the predictive performance evaluation for every model's parameters, saved during training for (A) Daphne, (B) Emerald, (C) Florence, (D) Greece, (E) Hilda. In the ordinate axis we sorted the saved models' parameters by Binary Accuracy values from the lowest to the highest, and plotted in the abscise axis the testing Binary Accuracy values in horizontal boxplots. In this plot, training accuracy means the Binary Accuracy calculated with the validating datasets during training.

In the case of the Greece model, we observed an eccentric behavior (Figure 2D), which means, it appears, as though it was a negative correlation between the Binary Accuracy observed in the training and the testing Binary Accuracy. But in a closer inspection, it can be seen a quick positive correlation between 73 to 94.26 % of Binary Accuracy observed in training reaching 90% of testing Binary Accuracy, and then a negative correlation was observed. It seems the model improved quic-

ISSN: EN TRÁMITE
https://doi.org/10.59741/p05c6793

kly and then it became less predictive. We think this behavior may be attributed to the excessive use of dropout layers between the 4 hidden layers that caused an uncertain effect.

To better understand the predictive power of the MLP models we calculated the Sensitivity, Specificity, Positive Predictive (PPV) and Negative Predictive Values (NPV) using the 40 testing datasets in the model evaluation process. So, we challenged the best model's parameters for every model (Daphne, Emerald, Florence, Greece and Hilda) to make predictions over these testing datasets and counted the true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). TP means the pre-miRNAs sequences that were predicted as pre-miRNAs, TN means the no pre-miRNAs sequences that were predicted as no pre-miRNAS, FP means non-miRNAS sequences that were predicted as pre-miRNAs and FN means pre-miRNAs sequences that were predicted as no pre-miRNAs.

Accordingly, the sensitivity concept means the accuracy of the model to correctly predict if a sequence is a miRNA given a dataset of miRNAs sequences. Sensitivity values for the selected best models vary between them. Emerald and Greece showed the highest values, Daphne and Florence showed lower values, and Hilda showed the lowest values. That means Emerald and Greece are fitted to correctly predict miRNAs in a population of miRNAs sequences because they got values ranging from 84 to 90% (Figure 3A). The specificity concept means the accuracy of the model to correctly discard sequences as pre-miRNA given a dataset of no pre-miRNA sequences. In this analysis, we observed that Daphne is better fitted to correctly discard sequences as miRNAs compared to other models because it showed the highest values ranging from near 98 to 100% (Figure 3B). These results are confirmed with the PPV and NPV values respectively (Figure 3C, 3D). These results suggest that some models like Emerald and Greece are fitted to correctly predict miRNAs but they are not well fitted to discard sequences as pre-miRNAs. In contrast, Daphne is an excellent model for discarding pre-miRNAs, but has a lower performance in correctly predicting miRNAs, unlike Emerald and Greece. It seems that, during training, each model was specialized to one task that could be in predicting or discarding.

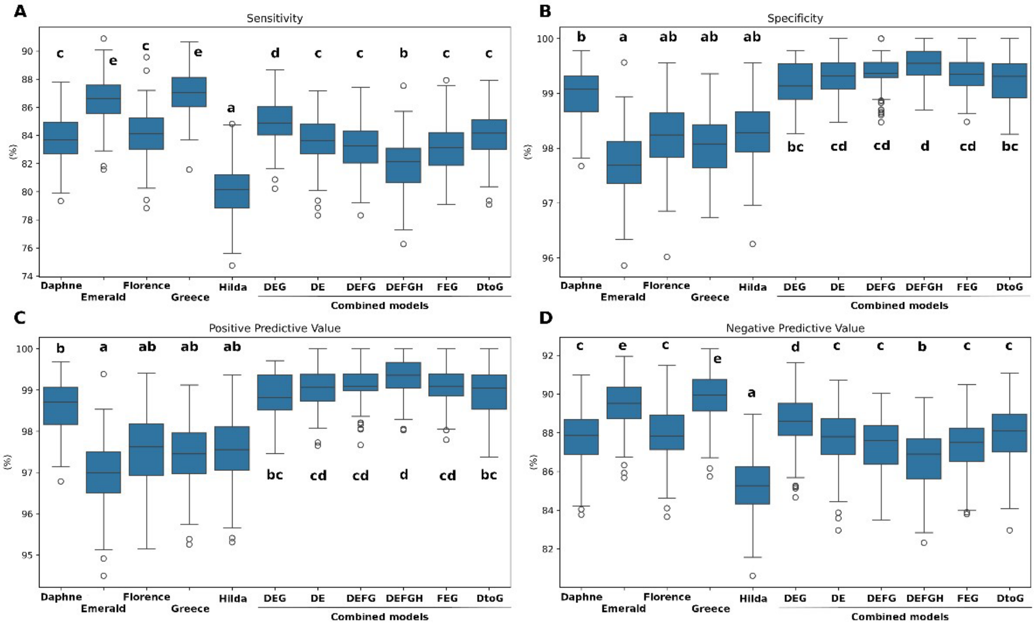

Figure 3. Predictive performance evaluation of individual models and combinations.. In this figure we plot the percentage of (A) sensitivity, (B) specificity, (C) PPV and (D) NPV. Lower case letters tag groups of means of data whose values are significantly different. We used the Tukey HSD method to calculate the minimum significant difference between means. In order to normalize the data, log(odds) values of the percentages were calculated, that means the ln(X/(100-X)) where X is the percentage value.

ISSN: EN TRÁMITE
https://doi.org/10.59741/p05c6793

As we saw earlier, the chaotic history of the evolution of the model's predictive performance is not related to the final model's accuracy. Despite the chaotic evolution of Greece's model, we observed a good performance at predicting miRNAs. In addition, we got, in this work, metric values very similar to the values obtained with the Convolutional Neural Networks proposed by Zheng *et al*. (2019), but with a less complex and more compact neural network. This observation agrees with the idea that MLP architecture using k-mer frequencies of nucleotide and secondary structure sequences as feature information, is a streamlined option.

3. Prediction improvement by models' output combinations

According to the observations about the task specialization, the models have suffered during training, we hypothesized that the average of the individual outputs of the models which have contrasting specialized scenarios would produce a more robust and accurate pre-miRNA sequence prediction. To test this hypothesis, we calculate the average of the Emerald, Daphne and Greece models' output with the following logic: as we observed, Daphne has specialized in discarding miRNAs, and Emerald and Greece have specialized in the predicting miRNAs task. We think the Daphne, Emerald and Greece output combination by the average will render better performance, in addition to the non-overfitted feature of the Greece model caused mainly by the use of excessive dropout layers between the 4 hidden layers. We called "DEG" to this combination. We also built other combinations and we used them as controls in the analysis ("DE", "DEFG", "DEFGH", "FEG", "DtoG", see Table 2).

Table 2. Combinations of the model's individual outputs

Combinations	Daphne	Emerald	Florence	Greece	Hilda
DEG	X[a]	X		X	
DE	X	X			
DEFG	X	X	X	X	
DEFGH	X	X	X	X	X
FEG		X	X	X	
DtoG	X			X	

a. the X's means that this model was combined with the other selected.

In this analysis, we confirmed the prediction of our hypothesis of combining Daphne, Emerald and Greece outputs. In (Figures 3A and 3B), we showed the sensitivity and specificity values for individual models as well as for the combinations. It can be observed that DEG's combined model showed a sensitivity significantly higher than Daphne, but was lower than Emerald and Greece. And the specificity of the DEG combined model is much higher than Emerald and Greece. It seems like the DEG combined model has inherited the features of the individual models and got a better predictive performance. It also can be observed in the values of PPV and NPV (Figure 3C and 3D). The other models showed better specificity but had lower values of sensitivity even lower or at least equal than Daphne and Florence. In general, these other models did not perform as well as DEG combined model did.

4. Predicting miRNAs from deep-sequencing RNA-seq data

Although the "DEG" combined model and other individual models have shown good results, they haven't been tested with real deep-sequencing RNA-seq data. To test the predictive performance on real data, we analyzed deep-sequencing RNA (dsRNA-seq) data obtained from *Phaseolus vulgaris* (common bean), a plant of interest in the agriculture and food industry (Accession number SRP074456) and *Medicago truncatula* (Accession number SRP000631). We downloaded the dsRNA-seq data from the NCBI SRA archive, and the data processing pipeline is described in the materials

ISSN: EN TRÁMITE
https://doi.org/10.59741/p05c6793

and methods section. Once we had the transcripts assembled from dsRNA-seq data, we analyzed them with Dpahne, Emeral and Greece models as well as the "DEG" combined model to predict which of these assembled sequences are pre-miRNAs and which of them aren't. We used as true pre-miRNAs, the stem-loop sequences reported at Plants miRNAs Encyclopedia (PmiREN, Guo *et al.*, 2020). The results of the assembling process and the transcripts prediction are shown in Table 3.

Table 3. Results of assembling transcripts and predictions

Models	**DEG combined model** [d]		**Daphne**	
Plant species	*P. vulgaris*	*M. truncatula*	*P. vulgaris*	*M. truncatula*
Predicted [a]	166	187	182	195
Discarded [b]	1394	2294	1378	2286
Total [c]	1560	2481	1560	2481
Models	**Emerald**		**Greece**	
Plant species	*P. vulgaris*	*M. truncatula*	*P. vulgaris*	*M. truncatula*
Predicted	217	243	198	222
Discarded	1343	2238	1362	2259
Total	1560	2481	1560	2481

a. Assembled transcripts predicted as pre-miRNA sequences.
b. Assembled transcripts discarded as pre-miRNA sequences.
c. Total assembled short transcripts (60 to 200 nucleotides). %GC ranges 10 – 65% for P. vulgaris and 0-78% for M. truncatula.
d. DEG combined model is the average of the Daphne, Emerald and Greece individual outputs.

Results of sensitivity and specificity as well as the percentage of true positive (TP), false positive (FP), true negative (TN) and false negative (FN) predicted with MLP models (DEG, Emerald, Daphne and Greece) on dsRNA data from P. vulgaris is shown in (Figure 4).

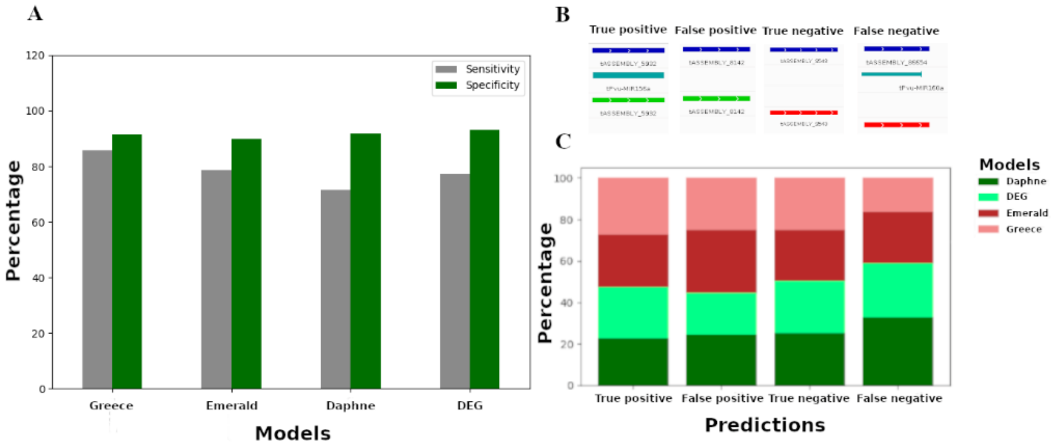

Figure 4. Models' predictive performance on dsRNA data of *P. vulgaris.* In (A) we plot the percentage of sensitivity and specificity for each tested model. (B) Schematic representation of the four possible outcomes: true positive, false positive, true negative and false negative. Horizontal blue bars represent the assembled transcripts; the cyan bars represent the mapped stem-loop sequences of reported miRNAs. Green and red horizontal bars represent the assembled transcripts, that were predicted as miRNA or discarded, respectively. In (C) we plot the percentages of true positives, false positives, true negatives and false negatives in stacked bars for each tested model.

ISSN: EN TRÁMITE
https://doi.org/10.59741/p05c6793

We observed good specificity values for all individual and combined models reaching above 90%, which means a good discarding performance. We also observed that the DEG sensitivity value was higher than Daphne's sensitivity value but lower than Emerald's and Greece's sensitivity values. Although this behavior, in general, was congruent with the models' testing results, we observed a notorious good performance in Greece predictions, showing high specificity values, even similar to Daphne's sensitivity value (Figure 4A). Additionally, we observed, that Daphne's model had the lowest percentage of TP and the highest percentage of FN; this observation correlates with the lowest sensitivity value. Contrastingly, Greece had the highest percentage of TP and the lowest percentage of FN, and this correlates with the highest sensitivity value. Emerald had the highest FP percentages which affected the model behavior diminishing the specificity value. These results suggest that Greece is the best-fitted model to analyze dsRNA data of Fabaceae family plants (Figure 4C). To confirm this suggestion, we evaluated these models with dsRNA data from other related plants, *Medicago truncatula*. So we hypothesized if Greece is the fitted model to accurately predict pre-miRNAs sequences from dsRNA data of Fabaceae plants, we should expect to find similar results on the prediction performance with transcripts from assembled dsRNA data of *M. truncatula.*

Results of sensitivity and specificity values, as well as count percentages for TP, FP, TN and FN for models' predictive performance on dsRNA data of *M. truncatula*, are shown in Figure 5. In this analysis, we observed, that every model had high values of specificity as they did in predictions made on dsRNA data of *P. vulgaris*. However, we observed that the sensitivity value of DEG's combined model was closely similar to Daphne's sensitivity value; it didn't improve and it doesn't correlate with the results obtained in the case of *P. vulgaris.* Unlike Daphne's and DEG's combined model, Greece and Emerald had high sensitivity values, being Greece the one that had the highest value. We also observed percentages of TP, FP, TN and FN with similar tendencies to these results observed in the case of *P. vulgaris*. This means Greece had the higher percentage of TP and the lowest percentage of FN, inversely Daphne model had the lower percentage of TP and the higher percentage of FN. Moreover, we observed that Emerald showed a higher percentage of FP. These results suggest, again, that Greece is the best-fitted model to predict plant miRNAs of the *Fabaceae family*.

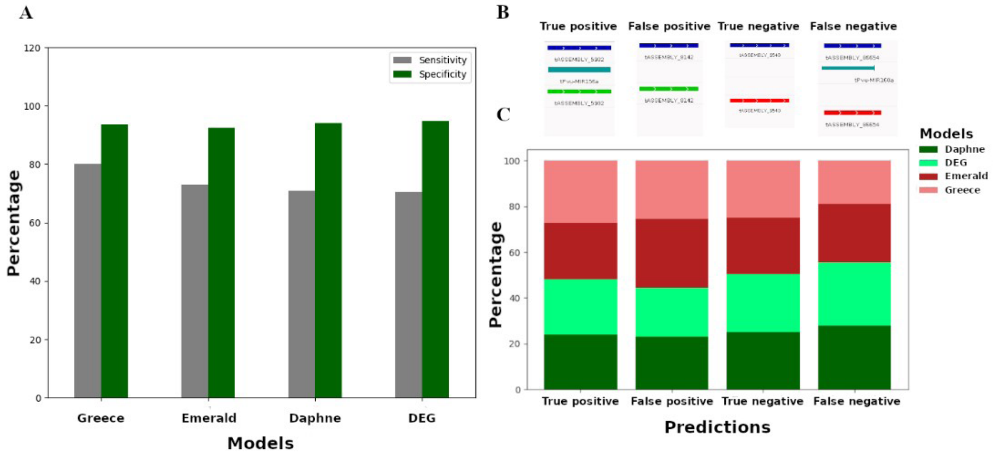

Figure 5. Models' predictive performance on dsRNA data of *M. truncatula* . In (A) we plot the percentage of sensitivity and specificity for each tested model. (B) Schematic representation of the four possible outcomes: true positive, false positive, true negative and false negative. Horizontal blue bars represent the assembled transcripts; the cyan bars represent the mapped stem-loop sequences of reported miRNAs. Green and red horizontal bars represent the assembled transcripts, that were predicted as miRNA or discarded, respectively. In (C) we plot the percentages of true positives, false positives, true negatives and false negatives in stacked bars for each tested model.

ISSN: EN TRÁMITE
https://doi.org/10.59741/p05c6793

Greece and Emerald showed low specificity values and showed high values of sensitivity compared to Daphne's specificity and sensitivity values during the testing stage (Figure 3). DEG combined model showed a balanced performance with high sensitivity values compared with Daphne and high specificity values compared to Greece and Emerald also, during the testing stage (Figure 3). Noteworthy, Greece performed better than Emerald Daphne and combined model DEG achieving high sensitivity values and similar specificity values when predicting miRNAs from deep-sequencing RNA data in *Phaseolus vulgaris* (Figure 4) and *Medicago truncatula* (Figure 5). In summary, Greece showed acceptable values of sensitivity and specificity ranging from 80.00 to 85.72% and 91.46 to 93.57%, respectively, which were closely similar to the values obtained with the convolutional neural network proposed by (Zheng *et al*., 2019).

CONCLUSION

In this work, we developed models of artificial intelligence based on a simple Multi-Layer Perceptron architecture with the ability to identify miRNAs sequences of the *Fabaceae* – family plants from the deep-sequencing RNA-seq data. The Greece model showed a chaotic evolution during training, fast improvement, and the best predictive performance with real data, thus we conclude that the use of dropout layers is important to avoid data overfitting in MLP models, and the chaotic evolution during the model training and testing isn't related to the final predictive performance with real data. Greece showed comparable values of accuracy, sensitivity and specificity to the convolutional neural network approach, but Greece is based on a simple MLP architecture, thus that permits training, tuning and running this model with low-capacity devices. We also conclude that calculating k-mer frequencies is another way to extract patterns and useful information from nucleotide sequences and secondary-structure representation sequences that can be used as input data in MLP-based models. Finally, we think the Greece artificial intelligence can be used in pipelines for the mapping and annotating of miRNAs in the Fabaceae-family genomes, and it can be easily retrained to perform prediction tasks in other plant families.

The models can be accessed in the following link: https://github.com/exseivier/mlp-fabaceae

Conflicts of Interest

"No conflicts of interest have declared the authors".

Author Contributions

Data processing: Marco Adán Juárez-Verdayes, Javier Montalvo-Arredondo.
Research: Marco Adán Juárez Verdayes, Javier Montalvo-Arredondo.
Research conceptualization: Javier Montalvo-Arredondo.
Software: Javier Montalvo-Arredondo.
Draft paper writing: Javier Montalvo-Arredondo.
Manuscript reviewing and editing: Marco Adán Juárez-Verdayes, Javier Montalvo-Arredondo.
Formal analysis: Marco Adán Juárez-Verdayes, Javier Montalvo-Arredondo.

ISSN: EN TRÁMITE
https://doi.org/10.59741/p05c6793

ACKNOWLEDGMENTS

We acknowledge "Coordinación de la División de Ingeniería" and "Departamento de Ciencias Básicas" from "Universidad Autónoma Agrara Antonio Narro" for the facilities infrastructure we used to build, train and test the MLP-based AI models.

REFERENCES

Andrews S. (2010). FastQC: A quality control tool for high throughput sequence data. Available at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 30, 2114–2120. DOI: https://doi.org/10.1093/bioinformatics/btu170.

Cha, M., Zheng, H., Braham, C., Li, X., & Hu, H. (2021). A two-stream convolutional neural network for microRNA transcription start site feature integration and identification. Scientific Reports, 11, 5625. DOI: https://doi.org/10.1038/s41598-021-85173-x.

Centeno-González, N. K., Martínez-Cabrera, H. I., Porras-Múzquiz, H., & Estrada-Ruiz, E. (2021). Late Campanian fossil of a legume fruit supports Mexico as a center of Fabaceae radiation. Communications Biology, 4, 41. DOI: https://doi.org/10.1038/s42003-020-01533-9.

Dong, Q., Hu, B., & Zhang, C. (2022). microRNAs and their roles in plant development. Frontiers in Plant Science, 13, 824240. DOI: https://doi.org/10.3389/fpls.2022.824240.

Gangadhar, B. H., Venkidasamy, B., Samynathan, R., Saranya, B., Chung, I.-M., & Thiruvengadam, M. (2021). Overview of miRNA biogenesis and applications in plants. Biologia, 76, 2309–2327. DOI: https://doi.org/10.1007/s11756-021-00763-4.

Guo, Z., Kuang, Z., Wang, Y., Zhao, Y., Tao, Y., Cheng, C., . . . others. (2020). PmiREN: a comprehensive encyclopedia of plant miRNAs. Nucleic Acids Research, 48, D1114–D1121. DOI: https://doi.org/10.1093/nar/gkz894.

Henderi, Wahyuningsih, T. & Rahwanto, E. (2021). Comparison of min-max normalization and z-score normalization in the K-nearest neighbor (kNN) algorithm to test the accuracy of types of breast cancer. 4(1), 13-20. DOI: https://doi.org/10.47738/ijiis.v4i1.73.

Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nature Biotechnology, 37, 907–915. DOI: https://doi.org/10.1038/s41587-019-0201-4.

Kozomara, A., & Griffiths-Jones, S. (2010). miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Research, 39, D152–D157. DOI: https://doi.org/10.1093/nar/gkq1027.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84 – 90. DOI: https://dl.acm.org/doi/10.5555/2999134.2999257.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Subgroup, 1. G. (2009). The sequence alignment/map format and SAMtools. Bioinformatics, 25, 2078–2079. DOI: https://doi.org/10.1093%2Fbioinformatics%2Fbtp352.

Lokuge, S., Jayasundara, S., Ihalagedara, P., Kahanda, I., & Herath, D. (2022). miRNAFinder: A comprehensive web resource for plant pre-microRNA classification. Biosystems, 215, 104662. DOI: https://doi.org/10.1016/j.biosystems.2022.104662.

Lorenz, R., Bernhart, S. H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., & Hofacker, I. L. (2011). ViennaRNA Package 2.0. Algorithms for Molecular Biology, 6, 1–14. DOI: https://doi.org/10.1186/1748-7188-6-26.

Meyers, B. C., Axtell, M. J., Bartel, B., Bartel, D. P., Baulcombe, D., Bowman, J. L., . . . others. (2008). Criteria for annotation of plant MicroRNAs. The Plant Cell, 20, 3186–3190. DOI: https://doi.org/10.1105/tpc.108.064311.

Owusu Adjei, M., Zhou, X., Mao, M., Rafique, F., & Ma, J. (2021). MicroRNAs roles in plants secondary me-

ISSN: EN TRÁMITE
https://doi.org/10.59741/p05c6793

tabolism. Plant Signaling & Behavior, 16, 1915590. DOI: https://doi.org/10.1080/15592324.2021.1915590.

Pertea, G., & Pertea, M. (2020). GFF utilities: GffRead and GffCompare. F1000Research, 9. DOI: https://doi.org/10.12688/f1000research.23297.2.

Rojo-Arias, J. E., & Busskamp, V. (2019). Challenges in microRNAs' targetome prediction and validation. Neural Regeneration Research, 14, 1672–1677. DOI: https://doi.org/10.4103%2F1673-5374.257514.

Shavanov, M. V. (2021). The role of food crops within the Poaceae and Fabaceae families as nutritional plants. IOP Conference Series: Earth and Environmental Science, 624, p. 012111. DOI: https://doi.org/10.1088/1755-1315/624/1/012111.

Song, X., Li, Y., Cao, X., & Qi, Y. (2019). MicroRNAs and their regulatory roles in plant–environment interactions. Annual Review of Plant Biology, 70, 489–525. DOI: https://doi.org/10.1146/annurev-arplant-050718-100334.

Tiwari, R., & Rajam, M. V. (2022). RNA-and miRNA-interference to enhance abiotic stress tolerance in plants. Journal of Plant Biochemistry and Biotechnology, 31, 689–704. DOI: https://doi.org/10.1007/s13562-022-00770-9.

Trevethan, R. (2017). Sensitivity, specificity, and predictive values: foundations, pliabilities, and pitfalls in research and practice. Frontiers in Public Health, 5, 307. DOI: https://doi.org/10.3389/fpubh.2017.00307.

Wani, S. H., Kumar, V., Khare, T., Tripathi, P., Shah, T., Ramakrishna, C., . . . Mangrauthia, S. K. (2020). miRNA applications for engineering abiotic stress tolerance in plants. Biologia, 75, 1063–1081. DOI: http://dx.doi.org/10.2478/s11756-019-00397-7.

Yang, X., Zhang, L., Yang, Y., Schmid, M., & Wang, Y. (2021). miRNA mediated regulation and interaction between plants and pathogens. International Journal of Molecular Sciences, 22, 2913. DOI: https://doi.org/10.3390%2Fijms22062913.

Zhang, F., Yang, J., Zhang, N., Wu, J., & Si, H. (2022). Roles of microRNAs in abiotic stress response and characteristics regulation of plant. Frontiers in Plant Science, 13, 919243. DOI: https://doi.org/10.3389%2Ffpls.2022.919243.

Zhang, L., Xiang, Y., Chen, S., Shi, M., Jiang, X., He, Z., & Gao, S. (2022). Mechanisms of microRNA biogenesis and stability control in Plants. Frontiers in Plant Science. 13, 844149. DOI: https://doi.org/10.3389/fpls.2022.844149.

Zhang, Y., Huang, J., Xie, F., Huang, Q., Jiao, H., & Cheng, W. (2024). Identification of plant microRNAs using convolutional neural network. Frontiers in Plant Science, 15, 1330854. DOI: https://doi.org/10.3389/fpls.2024.1330854.

Zhang, Z., Teotia, S., Tang, J., & Tang, G. (2019). Perspectives on microRNAs and phased small interfering RNAs in maize (Zea mays L.): functions and big impact on agronomic traits enhancement. Plants, 8, 170. DOI: https://doi.org/10.3390/plants8060170.

Zhao, Y., Wang, G., Tang, C., Luo, C., Zeng, W., & Zha, Z.-J. (2021). A battle of network structures: An empirical study of CNN, Transformer, and MLP. arXiv preprint arXiv:2108.13002. DOI: https://doi.org/10.48550/arXiv.2108.13002.

Zheng, X., Xu, S., Zhang, Y., & Huang, X. (2019). Nucleotide-level convolutional neural networks for pre-miRNA classification. Scientific Reports, 9, 628. DOI: https://doi.org/10.1038/s41598-018-36946-4

Zielezinski, A., Vinga, S., Almeida, J. and Karlowski, W.M., 2017. Alignment-free sequence comparison: benefits, applications, and tools. Genome biology, 18, pp.1-17. DOI: https://doi.org/10.1186/s13059-017-1319-7.