# MIRNAFINDER: A COMPREHENSIVE WEB RESOURCE FOR PLANT PRE-MICRORNA CLASSIFICATION

**Sandali Lokuge** [1], **Shyaman Jayasundara** [1], **Puwasuru Ihalagedara** [1], **Indika Kahanda** [2], **Damayanthi Herath** [1*],

[1]Department of Computer Engineering, University of Peradeniya, Peradeniya 20400, Sri Lanka

[2]School of Computing , University of North Florida, Jacksonville, FL 32224, United States

[*]Corresponding author: damayanthiherath@eng.pdn.ac.lk

The authors wish it to be known that Sandali Lokuge, Shyaman Jayasundara, and Puwasuru Ihalagedara should be regarded as joint first authors.

March 5, 2022

## ABSTRACT

microRNAs (miRNAs) are known as one of the small non-coding RNA molecules that control the expression of genes at the RNA level, while some operate at the DNA level. They typically range from 20–24 nucleotides in length and can be found in the plant and animal kingdoms as well as in some viruses. Computational approaches have overcome the limitations of the experimental methods and have performed well in identifying miRNAs. Compared to mature miRNAs, precursor miRNAs (pre-miRNAs) are long and have a hairpin loop structure with structural features. Therefore, most in-silico tools are implemented for pre-miRNA identification. This study presents a multilayer perceptron (MLP) based classifier implemented using 180 features under sequential, structural, and thermodynamic feature categories for plant pre-miRNA identification. This classifier has a 92% accuracy, a 94% specificity, and a 90% sensitivity. We have further tested this model with other small non-coding RNA types and obtained 78% accuracy. Furthermore, we introduce a novel dataset to train and test machine learning models, addressing the overlapping data issue in the positive training and testing datasets presented in PlantMiRNAPred for the classification of real and pseudo-plant pre-miRNAs. The new dataset and the classifier that can be used with any plant species are deployed on a web server freely accessible at *http://mirnafinder.shyaman.me/*.

***K*eywords** bioinformatics, microRNA, novel miRNA, machine learning, plant, neural networks, multilayer perceptron

## 1 Introduction

Ribonucleic Acid (RNA) silencing is an important and essential process that controls the class and abundance of proteins in a cell. This form of gene regulation in eukaryotes is mediated by a set of small non-coding RNAs [1]. Small non-coding RNAs are classified into four types depending on their origin, processing style, and effector protein association: microRNA (miRNA), small interfering RNA (siRNA), PIWI-interacting RNA (piRNA, animals only), and transfer RNA-derived small RNAs (tsRNAs) [2, 3, 4]. Among them, miRNAs are the most phylogenetically conserved and function post-transcriptionally to repress target genes [5, 6]. miRNAs typically range from 20–24 nucleotides in length and have been identified in plants and animals with different biogenesis. In the genome, miRNAs can be located in different regions, such as in introns, long non-coding transcripts, and/or in the chromosomal regions between two genes [7]. This study focuses on plant miRNAs and hereinafter refers to plant miRNAs as miRNAs if not otherwise specified.

The biogenesis of a miRNA starts through the transcription of miRNA genes by the RNA polymerase (Pol) II [8, 9, 10, 11]. Primary miRNA (pri-miRNA) is the initial transcript and it forms a stable hairpin-shaped stem-loop structure. Then, the pri-miRNA is processed into a pre-miRNA by a Dicer-like enzyme (DCL1) [12]. This pre-miRNA has a stem-loop hairpin secondary structure and ranges from 53nt to 938nt in length [13]. The DCL1 enzyme further processes this pre-miRNA into a mature miRNA/miRNA* duplex [14]. A small RNA methyltransferase, named Hua Enhancer (HEN1), methylates the 3' end of the duplex [15]. Unlike in animals, the plant miRNA/miRNA* duplex is completed within the nucleus and exported into the cytoplasm by Hasty (HST), a plant homolog of Exportin-5 [16].The miRNA/miRNA* duplex is separated in the cytoplasm [17]. Loading the miRNA strand (guide strand) into the cytoplasmic Argonaute (AGO) protein (AGO1, AGO7, and AGO10 [18]) forms a ribonucleoprotein complex, RISC, to silence the target messenger RNAs (mRNAs) [17].

The miRNAs have been identified as critical regulators of developmental processes such as vegetative phase change, flowering time, leaf morphogenesis, and response adaption to environmental stresses, which are linked to most aspects of plant biology [19]. Similarly, studies have reported that plant miRNAs are involved in plant immune, hormone signaling, and signal transduction pathways [20]. Increasing evidence of the importance of miRNAs has led genetic engineering applications to a new level of designing artificial miRNAs for many potential applications to improve the agronomic properties of crops and make medicines for various human/animal/plant diseases. Therefore, the identification of existing miRNAs and their target genes is essential and a crucial task.

Studying microRNAs is challenging since microRNAs are very short, and the closely related microRNA family members differ by about one nucleotide. In the beginning, miRNAs were identified via experimental approaches such as cloning [21, 22], splinted-ligation strategy [23], and genetic screening [24]. These methods, however, are costly, time-consuming, labor-intensive, subject to chance, difficult to identify miRNAs expressed at lower levels, and difficult to clone due to physical properties such as sequence composition (Koshiol2010, Ver2009).Thus, these methods have been replaced by next-generation sequencing-based (NGS) approaches like small RNA (small RNA-Seq) sequencing in high-throughput miRNA identification [25]. The data generated by these recent technological advances requires the use of computational approaches for miRNA identification. For instance, computational systems implemented using NGS data have become the most reliable approach for large-scale miRNA detection [26, 27].

A recent review highlights the need for plant-specific computational methods for miRNA identification [28]. Based on the implementation and performance, we can classify computational methods in miRNA identification as comparative and non-comparative. Because mature miRNA sequences are short, traditional feature engineering approaches generally fail to extract effective features. As a result, most computational methods focus on identifying pre-miRNAs rather than mature miRNAs [29]. Comparative methods search for the exact or near-exact counterparts to previously validated pre-miRNAs based on the conserved characteristics of pre-miRNA sequences. Though this approach provides relative ease in detecting high-throughput evolutionarily conserved pre-miRNAs in different species, it is difficult to identify novel pre-miRNAs with distant sequence homology compared with the known pre-miRNAs [30, 31]. However, the non-comparative methods have overcome this limitation. They are based on pre-trained machine learning (ML) models that use a set of rules for pre-miRNA identification based on pre-miRNA properties such as structural, thermodynamical, and sequential variations [30, 31, 32]. Still, it requires wet-lab experiments like northern blotting or reverse transcription polymerase chain reaction (PCR) for the output validation. Further, the amount of positive and negative datasets used for the training process directly affects the performance (accuracy, specificity, and sensitivity) of the ML model [33].

Support Vector Machine [13, 34], Random forest [35], Decision trees [36] and Naïve Bayes [37] are commonly used machine learning algorithms for this task. Additionally, some studies have implemented classification models using a Stochastic-based Random Forest classifier and Naïve Bayes Classifier [38] and have been able to achieve promising results in pre-miRNA identification. According to Yousef et al. [39, 40] one-class classifiers can effectively distinguish miRNAs from other small non-coding RNA classes. Some of the studies have combined an adaptive boosting algorithm with SVM (AdaBoost-SVM) to transfer weaker classifiers into one stronger classifier [41]. This combined algorithm has shown a higher degree of efficacy compared with the other studies in mature miRNAs identification. Similarly, a limited number of studies used both homology-based methods and the ab-initio method (Machine Learning) for miRNA prediction. mirHunter [42] which identifies both plant and animal pre-miRNAs, is such an example. Neural network (NN) algorithms such as convolutional neural networks (CNN), long short-term memory neural networks (LSTM) have also been utilized in pre-miRNA identification [43, 44, 45]. Most of these models are not trained with only plant pre-miRNAs but both humans and plants pre-miRNAs. The studies have concluded that more data must be collected in order to improve identification and address the class imbalance problem. Almost all the related studies have used different versions of the miRBase [46] database as a source of a positive dataset. Most of the plant miRNA identification models have used only plant pre-miRNAs while some of them have used human, animal, virus pre-miRNAs along with plant pre-miRNAs [47, 48] for the training process. As the negative dataset, coding sequences (CDSs) are taken since miRNAs are non-coding sequences. The Phytozome database [49] and EnsemblPlants [50] are the commonly used databases for this purpose [38, 51]. However, some studies have taken other types such as non-coding small RNA sequences as the negative training and testing dataset.

The majority of studies on pre-miRNA classification have relied on data published on the web server PlantMiRNAPred [52] [51, 53, 54, 55]. In this dataset, plant pre-miRNAs were downloaded from miRBase (version 14), which is now outdated. This study used their in-house sample selection algorithm to filter 980 sequences from each plant pre-miRNAs and pseudo-pre-miRNAs (coding sequences) to train the model. The testing dataset contains 1438 real pre-miRNAs and 1142 pseudo-pre-miRNAs. However, we discovered that there are 634 common pre-miRNA sequences among the positive training and testing sets. The model must be tested on unseen data in order to be properly evaluated. Therefore, if a study employs this dataset, it should use another positive testing dataset for the above purpose. Accordingly, we have created a new expanded gold-standard dataset derived from miRBase 22.1 for the research community.

To address the previously reported issues related to NN-based models, we developed a novel model using a multilayer perceptron (MLP) classifier, incorporating sequence-based, structural, and thermodynamic features. This model was trained and tested with our newly constructed dataset. In order to make it accessible to the research community, we deployed our classifier as a public web-server, *http://mirnafinder.shyaman.me/*. Similarly, the developed dataset is also available on the same server. This article reports the dataset generation, feature calculation, model training, result evaluation, and deployment of the classifier.

# 2 Methodology

## 2.1 Data preparation

In this study, we present a new dataset to the research community. This section discusses the criteria we followed to construct the complete dataset to train and test the pre-miRNA classification models.

### 2.1.1 Positive dataset

We downloaded plant pre-miRNAs from *miRBase* [46] version 22.1, which serves as the primary online repository for miRNA sequences and annotations. This database comprises 38,589 hairpin precursor miRNA entries, which express 48,885 mature miRNAs from 271 species. We retrieved 8615 plant pre-miRNA sequences and eliminated 261 duplicates. Then we discarded 106 sequences that contained nucleotides other than Adenine (A), Guanine (G), Cytosine (C), and Uracil (U). After the initial preprocessing stage, we obtained a total of 8,248 pre-miRNAs.

To generate a generalised test set, we filtered out pre-miRNA records that did not have at least one experimentally validated mature miRNA for the pre-miRNA and at least two referenced articles. The final positive training dataset consisted of 1,211 pre-miRNAs after meeting the above criteria. Figure 1 illustrates the length distribution of the 1,211 real pre-miRNAs. This set includes pre-miRNAs from *Arabidopsis lyrata*, *Arabidopsis thaliana*, *Brachypodium distachyon*, *Brassica napus*, *Brassica rapa*, *Chlamydomonas reinhardtii*, *Citrus sinensis*, *Gossypium hirsutum*, *Glycine max*, *Hordeum vulgare*, *Malus domestica*, *Manihot esculenta*, *Medicago truncatula*, *Nicotiana tabacum*, *Oryza sativa*, *Picea abies*, *Prunus persica*, *Physcomitrella patens*, *Pinus taeda*, *Populus trichocarpa*, *Sorghum bicolor*, *Solanum lycopersicum*, *Solanum tuberosum*, *Vitis vinifera*, and *Zea mays*.

The remaining 7037 pre-miRNAs were classified into species and ranked according to sequence count. The testing dataset then included 100 sequences from each of the top ten ranked species. *Brachypodium distachyon* (bdi), *Glycine max* (gma), *Gossypium raimondii* (gra), *Lotus japonicus* (lja), *Medicago truncatula* (mtr), *Oryza sativa* (osa), *Picea abies* (pab), *Populus trichocarpa* (ptc), *Sorghum bicolor* (sbi), and *Solanum tuberosum* (stu) were among the selected species.

### 2.1.2 Negative dataset

To create the negative dataset, coding DNA sequences (CDS) were used as the pseudo pre-miRNA hairpin structure sequences. The EnsemblPlants (Release 47) database was used to obtain CDS of *Arabidopsis thaliana*, *Triticum aestivum*, *Oryza sativa*, *Hordeum vulgare*, *Zea mays*, and *Physcomitrella patens*. We chose CDS only in the range of 70nt to 260nt to be comparable with the positive training set (Figure 1). To be effective, the model should be able to differentiate between positive and negative sequences with similar GC content, as well as sequences with similar Minimum Free Energy (MFE). Accordingly, we created a negative training dataset that was similar to the positive training dataset by employing the G+C% and MFE as described below.

Since testing and training datasets have to be mutually exclusive, we first removed 1000 pre-miRNA sequences in the testing dataset from the preprocessed 8248 pre-miRNAs. The MFE and G+C% of the resulting 7248 plant pre-miRNAs

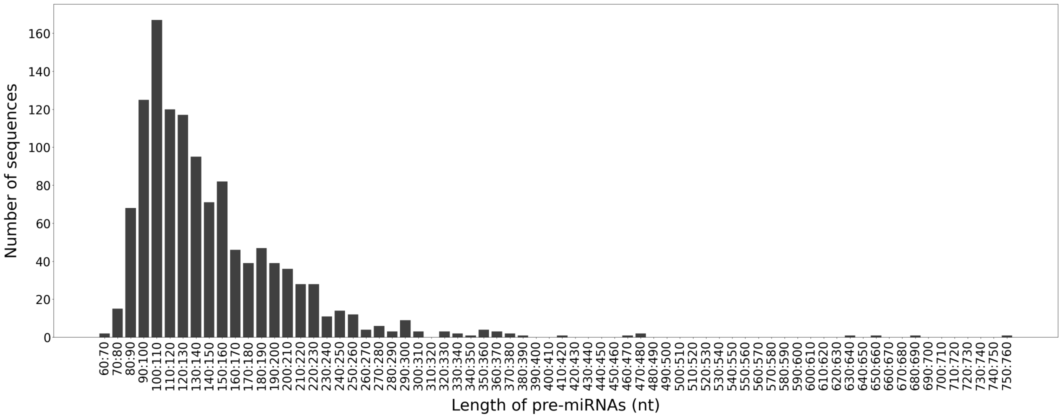


Figure 1: Length distribution of pre-miRNA sequences in training data

and CDSs were then calculated using RNAfold from the Vienna package [56] with the default parameters and grouped according to their lengths. Then, for each length group, we calculated the minimum and maximum values of MFE and G+C percent. The CDSs that meet the following two criteria were then extracted from each group.

- The MFE of a CDS in a particular length range should be between the minimum and maximum MFE values of the positive sequences in the same length range **AND**
- The G+C% of the same CDS should be between the minimum and maximum G+C% of the positive sequences in the same length range

For instance, a coding sequence in length from 70 to 79 nt should satisfy 9.6 <= MFE <= 70.9 and 17.16 <= G+C % <= 75.71.

The CDSs filtered using the aforementioned criteria were then randomly sampled in the same proportion as in the length intervals of the positive training set. Assume that the ratios for the length groups 70–79, 80–89, 90–99, 100–109, and 110–119 of the positive training sequences are 0.013, 0.059, 0.107, 0.144, and 0.103. The negative training samples should then have the same proportion for each length group as the positive training samples. Finally, we obtained 1,211 CDSs as the negative training dataset. The length distribution of the 1,211 CDSs in the negative training dataset is presented in the Figure 2.

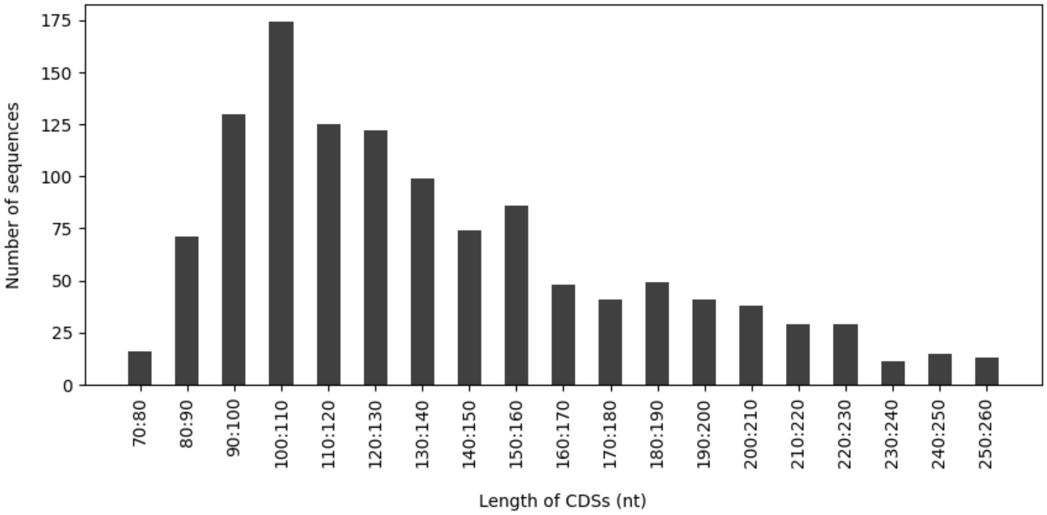


Figure 2: Length distribution of negative (CDS) training sequences

We used the same negative testing dataset in PlantMiRNAPred [52] because there were no shared sequences with our negative training dataset. This negative testing dataset was comprised of 1,142 CDSs. In addition to this testing dataset, we decided to collect additional non-coding RNA sequences to see how the model differentiates pre-miRNA sequences from other non-coding RNAs. A robust predictive model should be able to distinguish pre-miRNA not only from CDSs but also from other non-coding sequences. We downloaded other non-coding RNA sequences from RNAcentral (v15) [57], a non-coding RNA sequence database. Using the following search query, we included only sequence records that had been verified by at least three databases and had a secondary structure. The downloaded collection included 43 sequences of ribosomal RNA (rRNA), 579 sequences of small nucleolar RNA (snoRNA), 431 sequences of small nuclear RNA (snRNA), and 1,706 sequences of transfer RNA (tRNA).

- plant* AND expert_db: "Ensembl Plants" AND expert_db: "Rfam" AND expert_db: "ENA" AND has_secondary_structure: "True" AND rna_type: "rRNA"

Accordingly, our negative testing dataset included 1,142 CDSs and 2,759 other non-coding RNA sequences.

## 2.2 Features

Recent studies have introduced a set of features in both primary sequence and secondary structure that can be used to distinguish real plant pre-miRNAs and pseudo hairpins [52, 58, 35, 53]. We used 180 features under sequential, structural, and thermodynamic feature categories. We used scripts and algorithms described in our previous study, [59] to extract the features. Accordingly, this study includes 48 features introduced in microPred [60], triplet element features, and motif features. Our previous research [59], however, concluded that motif sequences have less of an impact on pre-miRNA identification than the other features. Furthermore, taking the computational time required for motif discovery into account, we extracted only 50 motifs with more than 10 sites from each positive and negative training set.

We generated the correlation matrix to get an overview of the features and discovered that there are certain highly correlated variables among them. Then we perform dimensionality reduction using *Principal Component Analysis (PCA)* [61]. Before performing the *PCA*, we normalized all the variables. A cumulative variance graph was plotted against the number of components, using the explained variances to determine the number of principal components. We selected 123 as the number of components where the cumulative variance reached 97% (Figure 3).

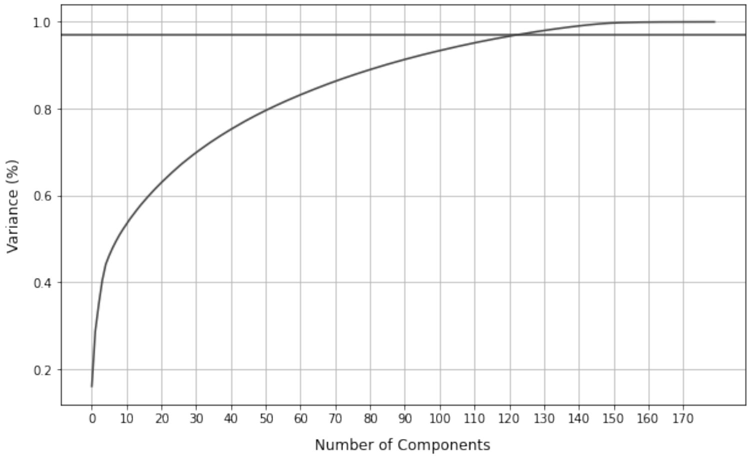


Figure 3: Cumulative variance vs number of principal components

## 2.3 Model training and performance evaluation

The past literature on miRNA identification has largely employed classification methods such as SVM, Decision tree, Random Forests, and Naive Bayes to develop their models. However, relatively few studies have focused on deep learning and neural network-based techniques for plant miRNA classification because of the limitation on obtaining a large number of positive miRNAs to train the model. Since we have 2,422 sequences for training, we employed a Multi-Layer Perceptron Classifier (MLP), which relies on an underlying neural network to execute the task of classification. After executing the hyperparameter tuning, the MLP model was trained with the Relu activation function, a constant learning rate, and 3 hidden layers with (100, 50, 20) units each.

To evaluate the model's performance, we use four criteria: accuracy, specificity, sensitivity, and area under the receiver operating characteristics curve (AUC-ROC).

## 3 Results and Discussion

### 3.1 Model evaluation

The model was developed using 5-fold cross-validation on 2422 sequences. The training accuracy at 5-fold cross-validation was 98.8% with a standard deviation of 0.005. The model was evaluated for classification accuracy on a test set of 1000 plant pre-miRNA sequences and 1142 CDSs. Accuracy, specificity, and sensitivity were 92%, 94%, and 90% respectively. Testing the model with actual pre-miRNAs and CDSs gives a 0.92 AUC-ROC value, indicating the high capability of differentiating the pre-miRNA sequences from coding sequences.

We further tested the model with actual pre-miRNAs and other non-coding RNAs and obtained an AUC-ROC value of 0.84. This result indicates that the developed model has a considerable class separation capacity for distinguishing pre-miRNAs from other non-coding RNA sequences. We obtained 78% testing accuracy for other non-coding RNA types. Though the negative training dataset doesn't include sequences from these non-coding RNA types, the model has performed well for such sequences as well. Furthermore, we tested our model for each non-coding RNA type individually. We obtained 70%, 72%, 90%, and 76% accuracy values for rRNA, snRNA, snoRNA, and tRNA, respectively.

We checked for the species-wise accuracies (Table 1) on the positive test set. All the species have an accuracy of greater than 75%. However, some species like *Brachypodium distachyon* (bdi), *Medicago truncatula* (mtr) and *Oryza sativa* (osa) have fewer accuracies compared to other species.

Table 1: Testing accuracy for different species

| Dataset | MLP model accuracy |
|---|---|
| Brachypodium distachyon | 0.82 |
| Glycine max | 0.90 |
| Gossypium raimondii | 0.94 |
| Lotus japonicus | 0.96 |
| Medicago truncatula | 0.79 |
| Oryza sativa | 0.82 |
| Picea abies | 0.91 |
| Populus trichocarpa | 0.93 |
| Sorghum bicolor | 0.95 |
| Solanum tuberosum | 0.95 |

Comparing our results with past studies is not reasonable as we employed a novel training and testing dataset. However, a few research have utilized a neural network-based classification model for plant pre-miRNA identification. We have acquired significant accuracy, specificity, sensitivity, and AUC-ROC for plant pre-miRNA identification.

We also examined the overall (Supplementary table S1) and species-specific (Supplementary table S2) classification performance of the models trained on each feature category (i.e., Micropred features, triplet element features, and motif features). Models were trained on both PCA and non-PCA features. In addition, we used chi-square scores to rank the original features (without PCA) to find the relationship between the features and the target variable (Supplementary table S3).

### 3.2 Comparison of the dataset

We compared our positive training and testing datasets to that of PlantMiRNAPred. PlantMiRNAPred consists of 980 sequences in the training set and 1438 sequences in the testing set. However, 634 common sequences belong to both the training and testing sets. Thus, the PlantMiRNAPred dataset had 1784 sequences in total. The dataset in this study comprises of 1211 sequences in the training dataset and 1000 sequences in the testing dataset, with no overlapping sequences.

We used sequence ID as the sequence compactor and found that there were 801 common sequences in both datasets (Figure 4). But miRNAFinder has 1410 new sequences since it is derived from miRBase v22.1.

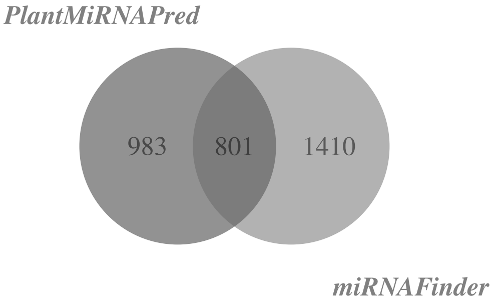


Figure 4: Comparison of PlantMiRNAPred dataset with miRNAFinder dataset

In addition, we compared the pre-miRNA sequences in the training datasets of the two studies. There were 304 sequences that were shared by both training sets, 907 novel sequences in miRNAFinder, and 676 sequences that were unique to the PlantMiRNAPred dataset. When the two testing datasets were compared, there were 190 similar sequences. Similarly, 810 and 1248 distinct sequences were found solely in miRNAFinder and PlantMiRNAPred, respectively. The low number of overlapping sequences in the testing datasets could be due to the fact that miRNAFinder contains pre-miRNAs of plant species such as *Brachypodium distachyon*, *Gossypium raimondii*, *Lotus japonicus*, *Picea abies*, *Sorghum bicolor* and *Solanum tuberosum* that have not been considered in previous studies.

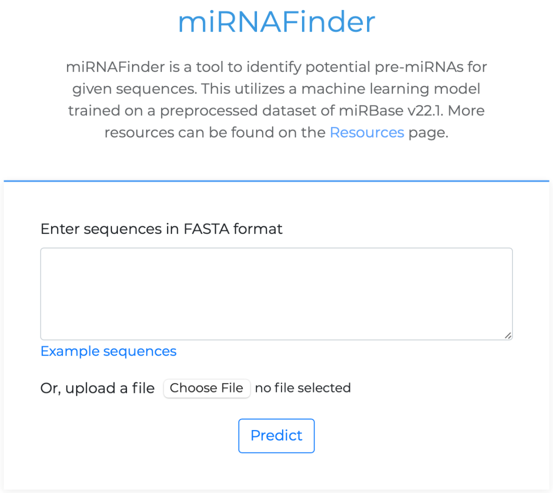
## 4 miRNAFinder web server

We published our classifier, miRNAFinder, on a web server that anyone could use to predict potential miRNAs. The input sequences should be in FASTA format, and there are two ways to enter them. The user can either enter the FASTA sequences into the provided text box or upload them using the "Choose file" button (Figure 5). We set the maximum number of sequences to 10 because feature calculation is computationally intensive and takes a long time. For each sequence, the output table comprises the sequence ID, predictions, and confidence scores (Figure 6). The results and extracted features are available for download from the same page. The training and testing datasets can also be downloaded from the resources page.

Flask, a Python web framework, was used to create the miRNAFinder web app. Multiple instances of the miRNAFinder flask program were run on the uWSGI application server, and user requests were handled efficiently. We used Celery, a background job scheduler, to improve server availability because feature extraction is a time-consuming task. Nginx functions as a reverse proxy as well as a load balancer. To allow for easier deployments, the entire program was built within Docker containers. Figure fig:architecture depicts the whole application architecture of miRNAFinder.

## 5 Web tool and data availability

miRNAFinder can be found at *http://mirnafinder.shyaman.me/*. The source code for the miRNAFinder server is publicly available on Github (*https://github.com/shyaman/miRNAFinder-web-server*) and in the Zenodo archive (*https://doi.org/10.5281/zenodo.4719605*). Anyone can set up their own instance of the miRNAFinder server based on their needs (for example, when a user needs to make predictions for more than ten sequences). The high-quality dataset presented here is available at *https://doi.org/10.5281/zenodo.4721396*.



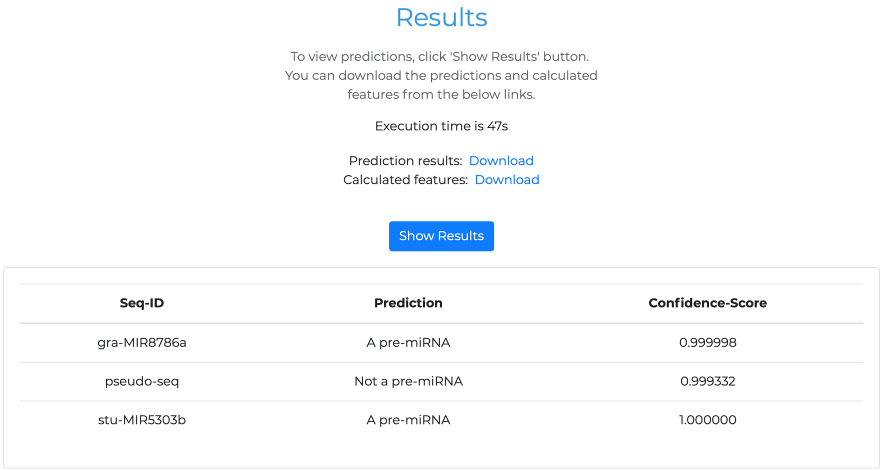
miRNAFinder

miRNAFinder is a tool to identify potential pre-miRNAs for given sequences. This utilizes a machine learning model trained on a preprocessed dataset of miRBase v22.1. More resources can be found on the Resources page.

Enter sequences in FASTA format

Or, upload a file Choose File no file selected

Figure 5: Homepage of the miRNAFinder web server



Results

To view predictions, click 'Show Results' button. You can download the predictions and calculated features from the below links.

Execution time is 47s

Prediction results: Download<br>
Calculated features: Download

| Seq-ID | Prediction | Confidence-Score |
|---|---|---|
| gra-MIR8786a | A pre-miRNA | 0.999998 |
| pseudo-seq | Not a pre-miRNA | 0.999332 |
| stu-MIR5303b | A pre-miRNA | 1.000000 |

Figure 6: Results view of the miRNAFinder web server

## 6 Conclusion and future work

Classification of real and pseudo-pre-miRNAs in plant species is challenging since plant pre-miRNAs have a diverse range of lengths compared to the conserved length of animal pre-miRNAs. For the scientific community, we present a new dataset of real and pseudo-pre-miRNAs. It includes 2211 real pre-miRNAs from miRBase (v22.1), 2353 CDSs from the EnsemblPlants database, and 2759 additional non-coding RNAs from the RNAcentral database. The scripts in the *microPred* were used to calculate compositional and thermodynamic features. Despite the fact that *microPred* concentrated on human miRNAs, the features had no detrimental impact on performance. The MLP classifier categorized pre-miRNAs with 92% accuracy, 94% specificity, and 90% sensitivity. The results suggest that the MLP classifier combined with the mentioned features can achieve significant accuracy and an AUC-ROC value. In future work, we suggest assessing the performance of MLP against classifiers such as RF, NB, and SVM using the new dataset.

In terms of the computational time, calculating the features introduced in the *microPred* took the longest time. In this work, the number of sequences used in training was reduced based on multiple criteria to reduce the computational time. It will be worthwhile to investigate alternate ways of reducing computing time for feature extraction without compromising the amount of data utilized for training.

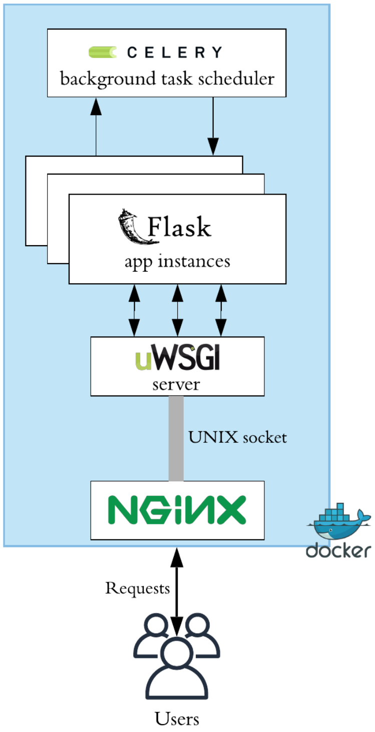


Figure 7: miRNAFinder architecture

The proposed model and dataset are accessible through a user-friendly web-server, *"miRNAFinder"*. The freely available dataset indicated above can be used to create, implement, and test new methods for plant pre-miRNA prediction and can be used with any species.

## References

[1] Akiko Hata. Functions of micrornas in cardiovascular biology and disease. *Annual Review of Physiology*, 75(1):69–93, 2013. PMID: 23157557.

[2] Diane Bortolamiol-Becet, Maghsoud Pazhouhandeh, Katia Marrocco, Pascal Genschik, and Véronique Ziegler-Graff. The polerovirus f box protein p0 targets argonaute1 to suppress rna silencing. *Current biology : CB*, 17:1615–21, 10 2007.

[3] Thomas Treiber, Nora Treiber, and Gunter Meister. Regulation of microrna biogenesis and its crosstalk with other cellular pathways. *Nature Reviews Molecular Cell Biology*, 20, 09 2018.

[4] Lei Zhu, David Ow, and Zhicheng Dong. Transfer rna-derived small rnas in plants. *Science China Life Sciences*, 61, 11 2017.

[5] Filipe Borges and Robert A Martienssen. The expanding world of small RNAs in plants. *Physiology & behavior*, 176(1):100–106, 2015.

[6] David Bartel. Metazoan micrornas. *Cell*, 173:20–51, 03 2018.

[7] Girish Shukla, Jagjit Singh, and Sailen Barik. Micrornas: Processing, maturation, target recognition and regulatory functions. *Molecular and cellular pharmacology*, 3:83–92, 01 2011.

[8] Yoontae Lee, Minju Kim, Jinju Han, Kyu-Hyun Yeom, Sanghyuk Lee, Sung Baek, and V Kim. microrna genes are transcribed by rna polymerase ii. *The EMBO journal*, 23:4051–60, 11 2004.

[9] Xuezhong Cai, Curt Hagedorn, and Bryan Cullen. Human micrornas are processed from capped, polyadenylated transcripts that can also function as mrnas. *RNA (New York, N.Y.)*, 10:1957–66, 01 2005.

[10] K. Rogers and X. M. Chen. Biogenesis, turnover, and mode of action of plant MicroRNAs. *Plant Cell*, 25:2383–2399, 2013.

[11] Z. X. Xie, E. Allen, N. Fahlgren, A. Calamar, S. A. Givan, and J. C. Carrington. Expression of Arabidopsis MIRNA genes. *Plant Physiology*, 138:2145–2154, 2005.

[12] Y. Kurihara and Yuichiro Watanabe. Arabidopsis micro-rna biogenesis through dicer-like 1 protein functions. *Proc. Natl. Acad. Sci. USA*, 1011:2753–12758, 01 2004.

[13] Jun Meng, Dong Liu, Chao Sun, and Yushi Luan. Prediction of plant pre-microRNAs and their microRNAs in genome-scale sequences using structure-sequence features and support vector machine. *BMC Bioinformatics*, 15(1):1–14, 2014.

[14] Olivier Voinnet. Origin, biogenesis, and activity of plant micrornas. *Cell*, 136:669–87, 03 2009.

[15] Zhiyong Yang, Yon Ebright, Bin Yu, and Xuemei Chen. Hen1 recognizes 21-24 nt small rna duplexes and deposits a methyl group onto the 2' oh of the 3' terminal nucleotide. *Nucleic acids research*, 34:667–75, 02 2006.

[16] Mee Yeon Park, Gang Wu, Alfredo Gonzalez-Sulser, Hervé Vaucheret, and R Poethig. Nuclear processing and export of micrornas in arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America*, 102:3691–6, 04 2005.

[17] Zhiqing Li, Ruodan Xu, and Ning Li. Micrornas from plants to animals, do they define a new messenger for communication? *Nutrition & Metabolism*, 15, 10 2018.

[18] Junli Wang, Jun Mei, and Guodong Ren. Plant microRNAs: Biogenesis, homeostasis, and degradation. *Frontiers in Plant Science*, 10(March):1–12, 2019.

[19] Anthony A Millar. The function of mirnas in plants. *Plants*, 9:198, 02 2020.

[20] Catherine Kidner and Robert Martienssen. The developmental role of microrna in plants. curr opin plant biol 8:38-44. *Current opinion in plant biology*, 8:38–44, 03 2005.

[21] Ramanjulu Sunkar, Thomas Girke, Pradeep Jain, and Jian-Kang Zhu. Cloning and characterization of micrornas from rice. *The Plant cell*, 17:1397–411, 06 2005.

[22] C. Lu. Elucidation of the small rna component of the transcriptome. *Science*, 309(5740):1567–1569, Feb 2005.

[23] Sangpen Chamnongpol, Patricia Maroney, and Timothy Nilsen. A rapid, quantitative assay for direct detection of micrornas and other small rnas using splinted ligation. *Methods in molecular biology (Clifton, N.J.)*, 667:3–17, 08 2010.

[24] Milo Aukerman and Hajime Sakai. Regulation of flowering time and floral organ identity by a microrna and its apetala2-like target genes. *The Plant cell*, 15:2730–41, 12 2003.

[25] Miya Howell, Noah Fahlgren, Elisabeth Chapman, J.s Cumbie, Christopher Sullivan, Scott Givan, Kristin Kasschau, and James Carrington. Genome-wide analysis of rna dependent rna polymerase6/dicerlike4 pathway in arabidopsis reveals dependency on mirna and tasirna directed targeting. *The Plant cell*, 19:926–42, 03 2007.

[26] Hikmet Budak, Zaeema Khan, and Melda Kantar. History and current status of wheat mirnas using next-generation sequencing and their roles in development and stress. *Briefings in functional genomics*, 14, 06 2014.

[27] Hikmet Budak and Melda Kantar. Harnessing ngs and big data optimally: Comparison of mirna prediction from assembled versus non-assembled sequencing data—the case of the grass aegilops tauschii complex genome. *Omics : a journal of integrative biology*, 19, 06 2015.

[28] Shyaman Jayasundara, Sandali Lokuge, Puwasuru Ihalagedara, and Damayanthi Herath. Machine learning for plant microrna prediction: A systematic review. *arXiv preprint arXiv:2106.15159*, 2021.

[29] Yuanlin Ma, Zuguo Yu, Guo-Sheng Han, Jinyan Li, and Vo Anh. Identification of pre-micrornas by characterizing their sequence order evolution information and secondary structure graphs. *BMC Bioinformatics*, 19, 12 2018.

[30] Malik Yousef, Louise Showe, and M.K. Showe. A study of micrornas in silico and in vivo: Bioinformatics approaches to microrna discovery and target identification. *The FEBS journal*, 276:2150–6, 03 2009.

[31] Kleftogiannis Dimitrios, Aigli Korfiati, Konstantinos Theofilatos, Likothanassis Spiros, Athanasios Tsakalidis, and Seferina Mavroudi. Where we stand, where we are moving: Surveying computational techniques for identifying mirna genes and uncovering their regulatory role. *Journal of biomedical informatics*, 46, 03 2013.

[32] Philip H. Williams, Rod Eyles, and Georg Weiller. Plant MicroRNA Prediction by Supervised Machine Learning Using C5.0 Decision Trees. *Journal of Nucleic Acids*, 2012:1–10, nov 2012.

[33] Nuno Mendes, Ana Teresa Freitas, and Marie-France Sagot. Survey and summary current tools for the identification of mirna genes and their targets. *Nucleic acids research*, 37:2419–33, 04 2009.

[34] Kuan Chieh Tseng, Yi Fan Chiang-Hsieh, Hsuan Pai, Chi Nga Chow, Shu Chuan Lee, Han Qin Zheng, Po Li Kuo, Guan Zhen Li, Yu Cheng Hung, Na Sheng Lin, and Wen Chi Chang. MicroRPM: A microRNA prediction model based only on plant small RNA sequencing data. *Bioinformatics*, 34(7):1108–1115, apr 2018.

[35] Peng Jiang, Haonan Wu, Wenkai Wang, Wei Ma, Xiao Sun, and Zuhong Lu. MiPred: Classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids Research*, 35(SUPPL.2):339–344, 2007.

[36] Philip H Williams, Rod Eyles, and Georg Weiller. Plant microrna prediction by supervised machine learning using c5. 0 decision trees. *Journal of nucleic acids*, 2012, 2012.

[37] Stephen Douglass, Ssu-Wei Hsu, Shawn Cokus, Robert B Goldberg, John J Harada, and Matteo Pellegrini. A naïve bayesian classifier for identifying plant micro rna s. *The Plant Journal*, 86(6):481–492, 2016.

[38] Worrawat Engchuan and Jonathan H Chan. Classification of real and pseudo pre-mirnas in plant species. *Procedia Computer Science*, 11:17–23, 2012.

[39] Malik Yousef, Jens Allmer, Waleed Khalifa, et al. Sequence motif-based one-class classifiers can achieve comparable accuracy to two-class learners for plant microrna detection. *Journal of Biomedical Science and Engineering*, 8(10):684, 2015.

[40] Malik Yousef, Naim Najami, Waleed Khalifav, et al. A comparison study between one-class and two-class machine learning for microrna target detection. *Journal of Biomedical Science and Engineering*, 3(03):247, 2010.

[41] Ying Wang, Jidong Ru, Yueqiu Jiang, and Jian Zhang. Adaboost-SVM-based probability algorithm for the prediction of all mature miRNA sites based on structured-sequence features. *Scientific Reports*, 9(1):1–14, 2019.

[42] Insong Koh and Ki-Bong Kim. miRHunter: A tool for predicting microRNA precursors based on combined computational method. *BioChip Journal*, 11(2):164–171, jun 2017.

[43] Binh Do, Vladimir Golkov, Goktug Gurel, and Daniel Cremers. Precursor microrna identification using deep convolutional neural networks. 09 2018.

[44] Md Rahman, Rashedul Islam, Shahidul Islam, Shakhinur Islam Mondal, and Md Amin. Mirann: A reliable approach for improved classification of precursor microrna using artificial neural network model. *Genomics*, 99:189–94, 02 2012.

[45] Huiqing Wang, Yue Ma, Chunlin Dong, Chun Li, Jingjing Wang, and Dan Liu. Cl-pmi: A precursor microrna identification method based on convolutional and long short-term memory networks. *Frontiers in Genetics*, 10, 10 2019.

[46] Ana Kozomara, Maria Birgaoanu, and Sam Griffiths-Jones. miRBase: from microRNA sequences to function. *Nucleic Acids Research*, 47(D1):D155–D162, 11 2018.

[47] Malik Yousef, Waleed Khalifa, İlhan Erkin Acar, and Jens Allmer. Microrna categorization using sequence motifs and k-mers. *BMC bioinformatics*, 18(1):170, 2017.

[48] Dimitrios M. Vitsios, Elissavet Kentepozidou, Leonor Quintais, Elia Benito-Gutiérrez, Stijn Van Dongen, Matthew P. Davis, and Anton J. Enright. Mirnovo: Genome-free prediction of microRNAs from small RNA sequencing data and single-cells using decision forests. *Nucleic Acids Research*, 45(21), 2017.

[49] David M Goodstein, Shengqiang Shu, Russell Howson, Rochak Neupane, Richard D Hayes, Joni Fazo, Therese Mitros, William Dirks, Uffe Hellsten, Nicholas Putnam, et al. Phytozome: a comparative platform for green plant genomics. *Nucleic acids research*, 40(D1):D1178–D1186, 2012.

[50] Dan Bolser, Daniel M Staines, Emily Pritchard, and Paul Kersey. Ensembl plants: integrating tools for visualizing, mining, and analyzing plant genomics data. In *Plant bioinformatics*, pages 115–140. Springer, 2016.

[51] Yuangen Yao, Chengzhang Ma, Haiyou Deng, Quan Liu, Jiying Zhang, and Ming Yi. PlantMirP: An efficient computational program for the prediction of plant pre-miRNA by incorporating knowledge-based energy features. *Molecular BioSystems*, 12(10):3124–3131, 2016.

[52] Ping Xuan, Maozu Guo, Xiaoyan Liu, Yangchao Huang, Wenbin Li, and Yufei Huang. PlantMiRNAPred: efficient classification of real and pseudo plant pre-miRNAs. *Bioinformatics*, 27(10):1368–1376, may 2011.

[53] Malik Yousef, Jens Allmer, and Waleed Khalifa. Sequence Motif-Based One-Class Classifiers Can Achieve Comparable Accuracy to Two-Class Learners for Plant microRNA Detection. *Journal of Biomedical Science and Engineering*, 08(10):684–694, 2015.

[54] Yingli Zhong, Ping Xuan, Ke Han, Weiping Zhang, and Jianzhong Li. Improved Pre-miRNA classification by reducing the effect of class imbalance. *BioMed Research International*, 2015, 2015.

[55] Malik Yousef, Jens Allmer, and Waleed Khalifa. Accurate plant microrna prediction can be achieved using sequence motif features. *Journal of Intelligent Learning Systems and Applications*, 8(1):9–22, 2016.

[56] L. Hofacker Ivo. Vienna rna secondary structure server. *Nucleic Acids Research*, 31(13):3429–3431, 2003.

[57] The RNAcentral Constortium. Rnacentral:a hub of information for non-coding rna sequences. *Nucleic Acids Research*, 01 2019.

[58] Chenghai Xue, Fei Li, Tao He, Guo Ping Liu, Yanda Li, and Xuegong Zhang. Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics*, 6:1–7, 2005.

[59] P. Ihalagedara, S. Lokuge, S. Jayasundara, D. Herath, and I. Kahanda. mirnafinder: A pre-microrna classifier for plants and analysis of feature impact. In *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pages 1–7, 2020.

[60] Rukshan Batuwita and Vasile Palade. microPred: Effective classification of pre-miRNAs for human miRNA gene prediction. *Bioinformatics*, 25(8):989–995, 2009.

[61] Jon Shlens. A tutorial on principal component analysis: derivation, discussion and singular value decomposition. *Mar*, 25(1):16, 2003.