miRNAFinder: A Comprehensive Web Resource for Plant Pre-microRNA Classification

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

[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.

June 30, 2021

Abstract

microRNAs (miRNAs) are known as one of the small non-coding RNA molecules, which control the expressions of genes at the RNA level. They typically range 20-24 nucleotides in length and can be found in the plant and animal kingdoms and in some viruses. Computational approaches have overcome the limitations in 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 the pre-miRNAs 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, 94% specificity, and 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 positive training and testing datasets presented in PlantMiRNAPred, a study done by Xuan et al.for the classification of real and pseudo plant pre-miRNAs. The new dataset and the classifier are deployed on a web server which is freely accessible via `http://mirnafinder.shyaman.me/`.

Keywords bioinformatics, novel miRNA, machine learning, microRNA, plant, neural networks, multilayer perceptron

1 Introduction

RNA silencing is an important and essential process that controls the type and amount of proteins in a cell. This form of gene regulation in eukaryotes is mediated by a set of non-coding RNAs, which they do not code proteins [1]. MicroRNA (miRNA), small interfering RNA (siRNA), PIWI-interacting RNA (piRNA, animals only), and transfer RNA-derived small RNAs (tsRNAs) are the four categories of small non-coding RNAs that are divided based on their origin, processing mode, and effector protein association [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. They are identified in plants, animals, and other eukaryotes [7], but have different biogenesis. In the genome, miRNAs can be located in the exons and introns of non-coding genes, introns of coding genes, and the intragenic regions [8]. This study focuses on the 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 polymerase (Pol) II [9, 10, 11, 12]. Primary miRNA (pri-miRNA) is the initial miRNA 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) [13]. This pre-miRNA is a 60-70 nucleotide long stem-loop hairpin secondary structure. DCL1 further processes this pre-miRNA into a mature miRNA/miRNA* duplex [14]. A small RNA methyltransferase, named Hua Enhancer (HEN)1 HEN1, methylates 3'-end of the duplex [15]. Differing from animals, the synthesis of plant miRNA/miRNA* duplex completes within the nucleus and is exported into the cytoplasm by the Hasty (HST), a plant homolog of Exportin-5 [16]. In the cytoplasm, the miRNA/miRNA* duplex is separated, and the discarding mechanism of the miRNA* strand (passenger strand) is unknown [17]. Loading the miRNA strand (guide strand) into cytoplasmic Argonaute (Ago) protein forms a ribonucleoprotein complex, RISC, to silence the target messenger RNAs (mRNAs) [17].

The miRNAs have been identified as critical regulators of the developmental process such as vegetative phase change, flowering time, leaf morphogenesis, and response adaption for environmental stresses linking to most aspects of plant biology [18]. Similarly, many recent studies have reported that plant miRNAs are involved in plant immune, hormone signaling, and signal transduction pathways [19]. Increasing evidence of the importance of miRNAs has led genetic engineering applications into 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 have been identified via experimental approaches such as cloning [20, 21], splinted-ligation mediation [22], and genetic screening [23]. However, these methods are expensive, time-consuming, labor-intensive, subjected to chance, challenging to identify miRNAs expressed at lower levels, and difficult to clone due to some physical properties like sequence composition and eliminate sequences during sample separation (RNA degradation) [24, 25]. 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 [26]. Data generated through these recent technological advances require computational approaches for miRNA identification using them. For instance, computational systems implemented using NGS data have become the most reliable approach for large-scale miRNAs detection [27, 28].

Based on the implementation and performance, we can classify computational methods in miRNA identification as comparative and non-comparative. Comparative methods search for the exact or near-exact counterparts to previously validated miRNAs based on the miRNA sequences' conservative characteristics. Though this approach provides relatively easiness in detecting high-throughput evolutionarily conserved miRNAs in different species, it is difficult to identify novel miRNAs with distant sequence homology compared with the known miRNAs [29, 30]. However, the non-comparative methods have overcome this limitation. They are based on pre-trained machine learning (ML) models that have utilized a set of rules for the miRNA identification according to miRNA properties such as the structural and thermo-dynamical and sequential variations [29, 30, 31]. 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 affect the performance (accuracy, specificity, and sensitivity) of the ML model [32]. Since pre-miRNAs have a hairpin loop structure with structural features and mature miRNAs are short, most of the studies focus on developing computational models for the pre-miRNAs identification [33].

Support Vector Machine [34, 35], Random forest [36], Decision trees [37] and Naïve Bayes [38] 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 [39] and have been able to achieve promising results in pre-miRNA identification. According to the Yousef et al. [40, 41] one-class classifiers can effectively distinguish miRNAs from other small non-coding RNA types. Some of the studies have combined an adaptive boosting algorithm with SVM (AdaBoost-SVM) to transfer weaker classifiers into one stronger classifier [42]. 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 [43] 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 [44, 45, 46]. Most of such models are not trained with only plant pre-miRNAs but both humans and plants pre-miRNAs. However, these studies have concluded the need to collect more data to improve the identification and address the class imbalance problem. Almost all the related studies have used different versions of miRBase [47] 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 [48, 49] for the training process. As the negative dataset, coding sequences (CDSs) are taken since miRNAs are non-coding sequences. The Phytozome database [50] and EnsemblPlants [51] are

commonly used database for this purpose [39, 52]. However, some studies have taken other types such as non-coding small RNA sequences as the negative training and testing dataset.

Most of the studies conducted on pre-miRNA classification have used the data published on the web server PlantMiRNAPred [53] [52, 54, 55, 56]. In this dataset, plant pre-miRNAs were downloaded from miRBase (version 14), which is now outdated. This study has 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 identified that there are 634 common pre-miRNA sequences among the positive training and testing sets. To have a proper evaluation, the model must be tested on unseen data. Thus, if a study uses this dataset, it should use another positive testing dataset. Similarly, the latest version of the miRBase is 22.1. Accordingly, we have developed a new expanded gold-standard dataset derived based on the mentioned latest version to the research community via this study.

For addressing the previously reported issues related to NN-based models, we developed a novel model using 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, results evaluation, and deployment of the classifier.

2 Methodology

2.1 Data preparation

In this study, we present a new dataset for the research community. Thus, this section covers 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* [47] version 22.1 (latest version), which is the primary online repository for miRNA sequences and annotations. This database contains 38,589 entries of hairpin precursor miRNAs which express 48,885 mature miRNAs of 271 species. We extracted 8615 plant pre-miRNAs and removed 261 duplicate pre-miRNA sequences. Then we removed 106 sequences that contain nucleotides other than Adenine(A), Guanine (G), Cytosine (C), and Uracil (U). Finally, we obtained a total of 8248 pre-miRNAs as the positive dataset.

To further improve the confidence of the curated data , we further filtered out the pre-miRNAs not having at least one experimentally validated mature miRNA for the pre-miRNA and not having at least two reference articles. We considered all the pre-miRNA sequences which satisfy both conditions as our (high-confidence) positive training sequences. This resulted in a final positive training dataset composed of 1,211 pre-miRNAs.

We grouped the remaining 7037 pre-miRNAs by their species and sorted them by the number of sequences per species. Then we considered only ten species with the highest pre-miRNA count and randomly selected 100 pre-miRNAs from each species to prepare the positive testing set. The selected species are *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*).

According to the ratios of 1211 real pre-miRNAs in different length ranges represented in Figure 1, it can be stated that lengths of the most known plant pre-miRNAs range from 70 to 260 nt.

2.1.2 Negative dataset

The protein-coding sequence (CDS) regions were adopted as the "pseudo" pre-miRNA hairpin structure sequences to construct the negative dataset. The CDSs of *Arabidopsis thaliana, Triticum aestivum, Oryza sativa, Hordeum vulgare, Zea mays,* and *Physcomitrella patens* were obtained from EnsemblPlants [51] database (Release 47). To perform well, the model should able to distinguish the positive and negative sequences with similar G+C% content. Similarly, the model should have the ability to separate positive sequences from negative sequences even though they have similar minimum free energy (MFE). Thus, we created a dataset including positive and negative sequences with similar G+C% and MFE following the criterion described below.

First, we eliminated the sequences in the positive testing dataset from 8248 pre-miRNAs, which were remained after removing the sequences with non-ACGU nucleotides. Hence, the sequences in the negative training dataset was mutually exclusive of those of the testing dataset. Then, we calculated the MFE and G+C% of remaining known plant

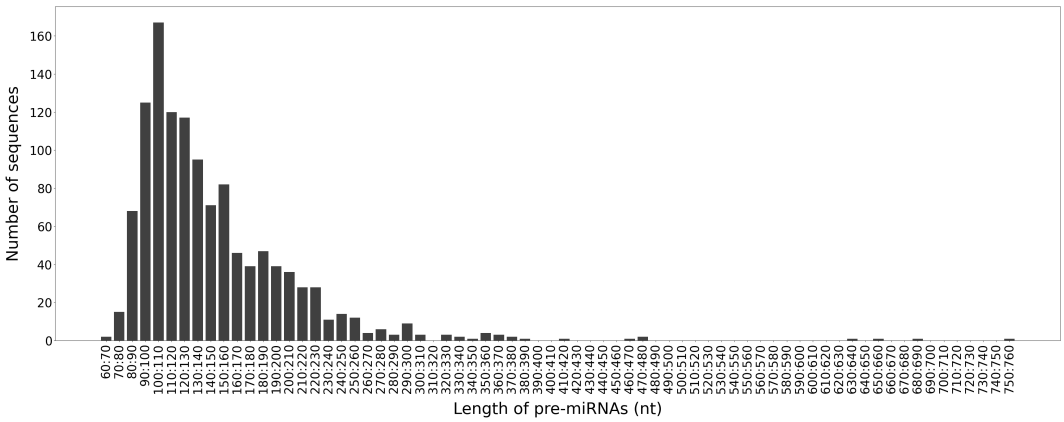

Figure 1: Length distribution of pre-miRNA sequences in training data

pre-miRNAs and CDS segments using RNAfold from the Vienna package [57] with the default parameters. We grouped the pre-miRNAs according to their lengths. Next, we determined the minimum and maximum values of MFE and G+C% of each length group. According to the most frequent length of training sequences shown in Figure 1, we chose 70 nt to 260 nt length ranges in selecting the CDSs to produce negative sequences.

Then, we grouped CDSs based on their lengths and extracted the CDSs from each group which satisfy the following two criteria;

- The MFE value 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% value of this negative sequence should be between the minimum and maximum G+C% values 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 sequences for the negative training dataset were collected considering the proportion of the real pre-miRNAs (positive training sample) of different length groups. For example, assume 0.013, 0.059, 0.107, 0.144, and 0.103 are the ratios for the length groups 70–79, 80–89, 90–99, 100–109, and 110-119 of the positive training sequences. Then the negative training samples also should maintain the same proportion for each length group. Finally, we obtained 1,211 CDSs as the negative training dataset. The ratios of CDSs with different lengths in the 1,211 negative training dataset are listed in Figure 2.

There were no common sequences between our negative training dataset and the negative testing dataset available in the PlantMiRNAPred [53]. Thus, we used the same negative testing dataset from PlantMiRNAPred as our negative testing set. This negative testing dataset comprises of CDSs. Besides this testing dataset, we decided to collect other non-coding RNA sequences to test how the model distinguishes pre-miRNA sequences from other non-coding RNA types. 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 typed sequences from RNAcentral (v15) [58], which is a non-coding RNA sequence database. The downloaded collection includes sequences of Ribosomal ribonucleic acid (rRNA), Small nucleolar RNA (snoRNA), Small nuclear RNA (snRNA), and Transfer ribonucleic acid(tRNA). The search criteria to select these sequences are as follows.

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

Each sequence in the given RNA type is verified by at least three databases and folded into stem-loop structures. In this set, there are 43 sequences from rRNA, 579 sequences from snoRNA, 431 sequences from snRNA, and 1,706 sequences from tRNA. Accordingly, our negative testing dataset includes both CDS and 2,759 non-coding sequences.

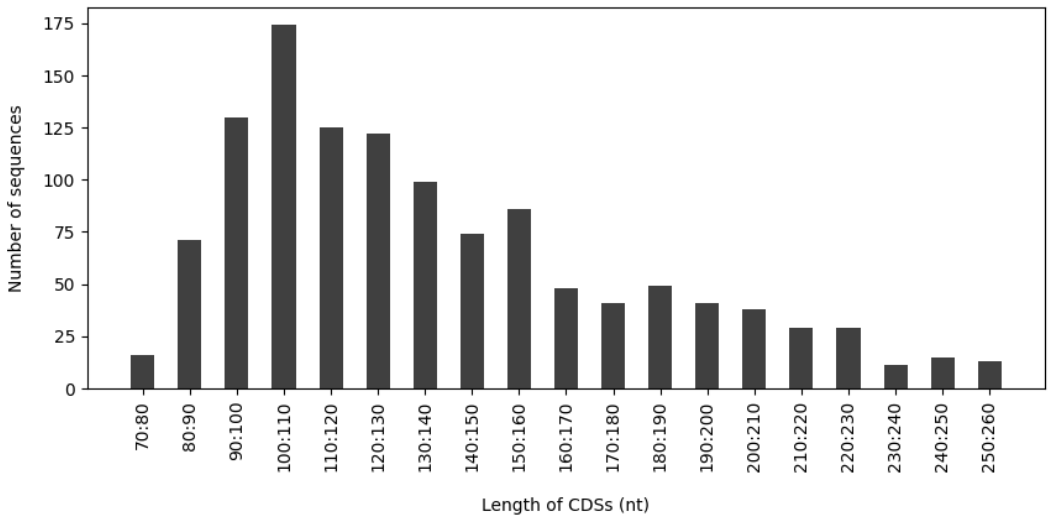

Figure 2: Length distribution of negative (CDS) training sequences

2.2 Features

The recent studies have introduced a set of features in both primary sequence and secondary structure which can be applied to classify real plant pre-miRNAs and pseudo hairpins [53, 59, 36, 54]. We used such 180 features under sequential, structural, and thermodynamic feature categories.

We followed scripts and algorithms described in the study [60] to extract the features. Accordingly, this study includes 48 features introduced in microPred [61], triplet element, and motif features. However, the study [60] has concluded that motif sequences have less impact on pre-miRNA identification than the other features. Similarly, considering the computational time takes for motif discovery we extracted 50 motifs with more than 10 sites from each positive and negative training set. The identified 100 motifs are given as regular expressions by MEME (Multiple Expectation Maximization for Motif Elicitation) suite v5.1.0 [62]. Then the scoring method in the study [60] was applied to generate a motif-based score for every sequence using those regular expressions.

2.3 Feature selection

Feature selection is an important step in creating machine learning models. Selecting most related features helps to reduce the computational cost of feature extraction, model training time and over-fitting. It also helps to improve the model performance.

We have calculated correlation matrix to get an overview of the features and found that there are some highly correlated variables among them. Then we perform dimensionality reduction using *Principal Component Analysis(PCA)* [63]. The idea of *PCA* is to transform large set of variables into a smaller set but still contain enough information to represent the larger set. 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 determined the number of principal components. We selected 123 as the number of components where the cumulative variance reached 97% (see Figure 3).

2.4 Performance evaluation and model training

We adopt three criteria namely, *Accuracy, Specificity* and *Sensitivity* to evaluate the performance of the model. These are most widely used measurements in classification problems to evaluate the prediction system. A prediction of a classifier can be divided into four categories: True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN).If a positive instance is predicted as a positive, then it is a true positive while if it is predicted as negative, it is a false negative. If a negative instance is predicted as a negative, then it is a true negative while if it is predicted as positive, then it is a false positive. The accuracy, specificity, and sensitivity can be defined using these categories as follows,

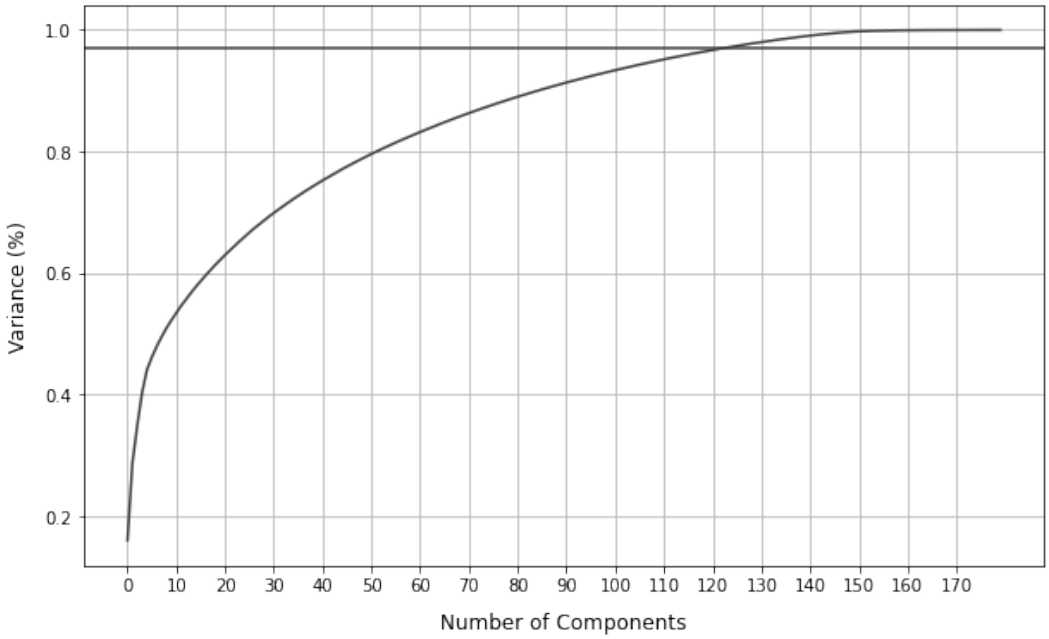

Figure 3: Cumulative variance vs number of principal components

$$Sensitivity = \frac{TP}{TP + FN}, \quad Specificity = \frac{TN}{TN + FP},$$
$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Beside the above evaluation matrices, we used AUC (Area Under The Curve) of ROC (Receiver Operating Characteristics) curve which is an important evaluation metrics for measuring any classification model's performance. ROC and AUC presents a probability curve and degree or measure of separability respectively. Further, AUC-ROC value indicates the capability of the model to distinguish the two classes, i.e., pre-miRNA and non pre-miRNA.

The past literature of miRNA identification have mostly used classification algorithms such as SVM, Decision tree, Random Forests, and Naive Bayes to construct their models. However, no study have focused on deep learning, and neural networks on plant miRNA classification because of the limitation in finding considerable number of positive miRNAs to train the model. Since we have 2,422 sequences for training, we used Multi-Layer Perceptron Classifier (MLPClassifier) which relies on an underlying Neural Network to perform the task of classification. After performing the hyperparameter tuning, the MLP model was trained with 'relu' activation function, 'constant' learning rate, and 3 hidden layers with (100,50,20) units each.

3 Results and Discussion

3.1 Model evaluation

The model evaluation was done using the positive and negative testing datasets as described under the section, Data preparation. The positive testing dataset includes pre-miRNAs of ten plant species, and the negative dataset contains both CDSs and other non-coding RNAs.

First, we checked for the species-wise accuracies (Table 1). All the species have accuracies with more than 75%. However, for some species like *Brachypodium distachyon (bdi), Medicago truncatula(mtr)* and *Oryza sativa (osa)* have fewer accuracies compared to other species.

We tested the model performances for CDSs and other non-coding RNA types separately since we wanted to know how this model distinguishes pre-miRNAs from other non-coding RNA types. We obtained 94% accuracy for the CDSs

Table 1: Accuracy values 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

while 78% 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. Further, 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 obtained the AUC-ROC value of 0.84 for testing the model with non-coding RNA typed sequences and actual pre-miRNA sequences. This result indicates that the developed model has a considerable class separation capacity for distinguishing pre-miRNAs from other non-coding RNA sequences. Testing the model with actual pre-miRNAs and CDSs gives 0.92 for AUC-ROC value indicating the high capability of separating the pre-miRNA sequences from coding sequences.

The overall accuracy, sensitivity, and specificity of the model were calculated considering the whole positive testing dataset and the negative dataset with CDSs. We obtained 92% accuracy for the MLP model. Though no study used a neural network-based classification model for plant pre-miRNA identification, we have gained significant accuracy, specificity, sensitivity, and AUC-ROC for plant pre-miRNA identification.

Comparing our results with past studies is not reasonable since we used a novel training and testing dataset. Thus, we reported only our results in Table 2.

Table 2: Overall evaluation

Accuracy	Specificity	Sensitivity	AUC-ROC
0.92	0.94	0.90	0.90

Further, we analyzed the impact of the different feature types on the ten different plant species and the top 50 features with the highest chi-square scores to find the relationship between the features and the target variables. The results of these analyses are provided in Supplementary Material.

3.2 Comparison of the dataset

For the comparisons, we used the PlantMiRNAPred dataset with miRNAFinder. Here we consider the positive set of both datasets. PlantMiRNAPred consists of 980 sequences in the training set and 1438 sequences testing set. But 634 common sequences belong to both training and testing sets. So altogether there are 1784 sequences in the PlantMiRNAPred dataset. miRNAFinder consists of 1211 sequences in the training dataset and 1000 sequences in the testing dataset. Since there are no overlapping sequences, 2211 sequences are in miRNAFinder.

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

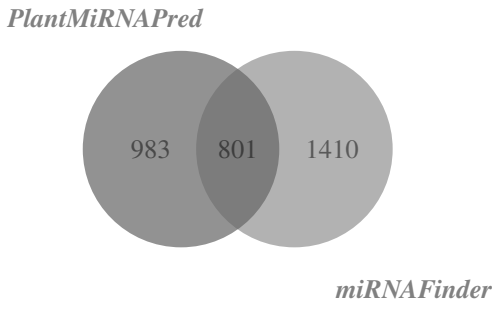

Figure 4: Comparison of PlantMiRNAPred dataset with miRNAFinder dataset

Apart from that, we have compared the training and testing datasets separately. There are 304 common sequences in both training sets, 907 new sequences in miRNAFinder, and 676 sequences in PlantMiRNAPred. But in the testing set, there are only 190 common sequences, 810 sequences in miRNAFinder, and 1248 sequences in PlantMiRNAPred. So in the training set the overlapping is a bit low and the reason could be that miRNAFinder consists of new plant species like *Brachypodium distachyon*, *Gossypium raimondii*, *Lotus japonicus*, *Picea abies*, *Sorghum bicolor* and *Solanum tuberosum*.

4 miRNAFinder web server

We deployed our latest classifier, miRNAFinder, on a web server where anyone could access and use it to predict potential miRNAs.

The input should be in FASTA format. Since feature calculation is computationally intensive and takes considerable time, we limited the maximum number of sequences to 10. There are two options to input FASTA sequences. The user can enter the FASTA sequences in the given text box or use the "Choose file" button to upload (Figure 5). On the results page (Figure 6), the user can view the predicted results by clicking the 'Show Results' button. The output table includes sequence IDs, predictions, and their confidence scores, whereas the both results and extracted features can be downloaded from the same page. Additionally, the training and testing datasets of can be downloaded from the resources page.

Flask, a Python web framework, was used to develop the miRNAFinder web app, and uWSGI was used as the application server. uWSGI can deploy multiple instances of the miRNAFinder flask app, and it can handle requests efficiently. Since feature extraction is a long running task, we used Celery, a background task scheduler, to optimize the availability of the server. Furthermore, Nginx is acting as a reverse proxy and load balancer. The whole application was developed inside docker containers to enables effortless deployments. Figure 7 shows the entire application architecture of miRNAFinder.

5 Web tool and data availability

The source code for the miRNAFinder server is publicly available from Zenodo archive (`https://doi.org/10.5281/zenodo.4719605`) and Github (`https://github.com/shyaman/miRNAFinder-web-server`). Accordingly anyone can deploy their own instance of miRNAFinder server according to the requirements (for example, when a user needs to make predictions for more than ten sequences).

miRNAFinder can be accessible through `http://mirnafinder.shyaman.me/`.

The introduced high quality dataset is deposited on `https://doi.org/10.5281/zenodo.4721396`.

6 Conclusions

Classification of real and pseudo-pre-miRNAs in plant species is challenging since plant pre-miRNAs are more diverse than animal pre-miRNAs. Here, we introduce a new dataset with real and pseudo-pre-miRNAs for the research community. It consists of 2211 real pre-miRNAs from miRBase v22.1, 2353 CDSs from EnsemblPlants database, and 2759 other types of non-coding RNAs from RNAcentral database. We used the scripts in the *microPred* to calculate compositional and thermodynamic features. Although *microPred* focused on human miRNAs, the features have not

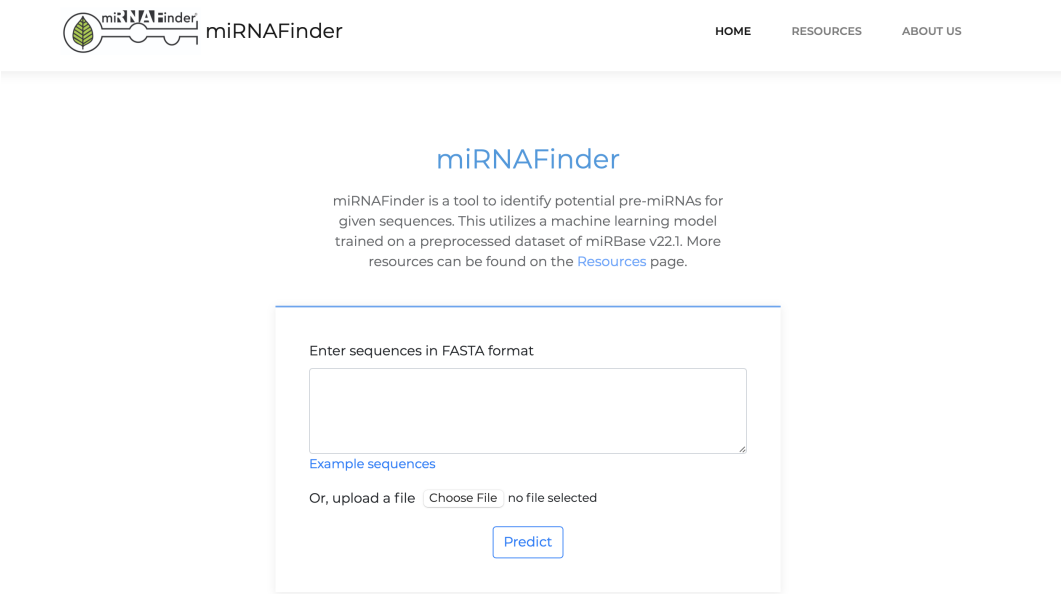

Figure 5: Homepage of the miRNAFinder web server

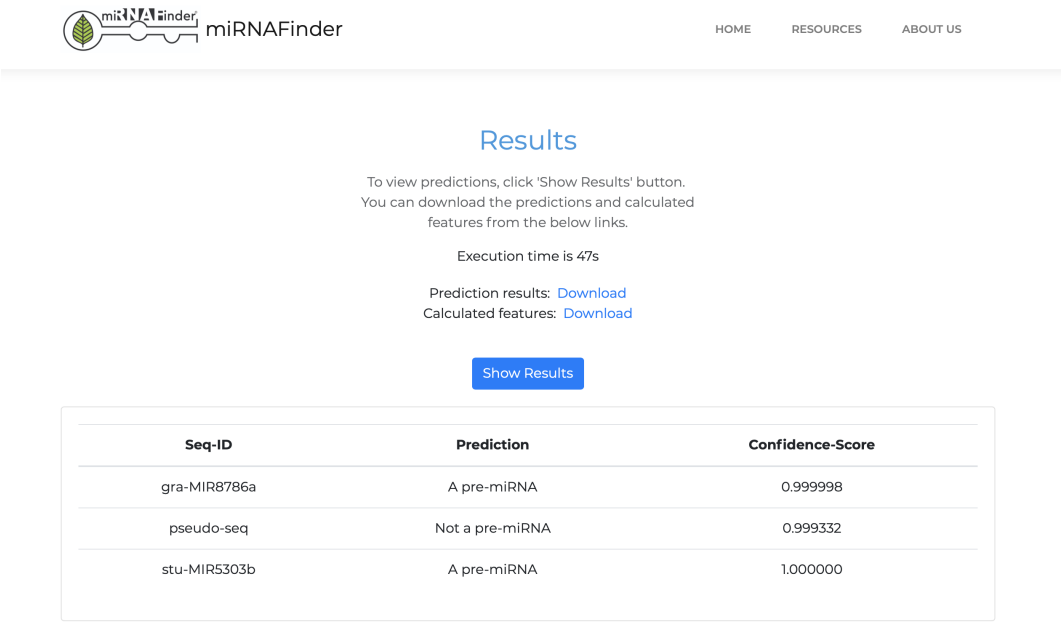

Figure 6: Results view of the miRNAFinder web server

negatively affected the performance. The MLP classifier has classified pre-miRNAs with an accuracy of 92%, a specificity of 94%, and sensitivity of 90%. The results show that the MLP classifier combined with the mentioned features can achieve a significant accuracy and AUC-ROC value.

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

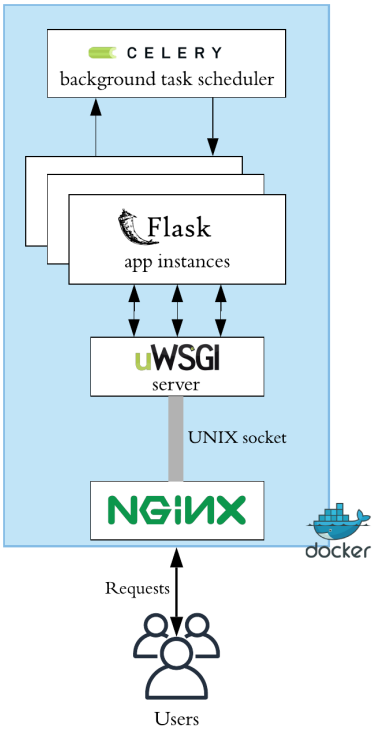

Figure 7: miRNAFinder architecture

time. It will be worthy to explore alternative approaches to reduce the computational time for feature extraction without comprising the amount of data used for training.

The proposed model and the dataset are available via a user-friendly web-server, *"miRNAFinder"*. The mentioned freely accessible dataset is useful in developing, implementing, and evaluating new methods for plant pre-miRNA prediction.

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] Michael Axtell, Jakub Westholm, and Eric Lai. Vive la différence: Biogenesis and evolution of micrornas in plants and animals. *Genome biology*, 12:221, 04 2011.

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

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

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

[19] 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.

[20] 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.

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

[22] 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.

[23] 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.

[24] Jill Koshiol, Ena Wang, Yingdong Zhao, Francesco Marincola, and Maria Landi. Strengths and limitations of laboratory procedures for microrna detection. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*, 19:907–11, 04 2010.

[25] Turgay Unver, Deana Namuth-Covert, and Hikmet Budak. Review of current methodological approaches for characterizing micrornas in plants. *International journal of plant genomics*, 2009:262463, 10 2009.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

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

[40] 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.

[41] 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.

[42] 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.

[43] 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.

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

[45] 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.

[46] 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.

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

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

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

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

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

[59] 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.

[60] 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.

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

[62] Timothy L. Bailey, Mikael Boden, Fabian A. Buske, Martin Frith, Charles E. Grant, Luca Clementi, Jingyuan Ren, Wilfred W. Li, and William S. Noble. MEME Suite: Tools for motif discovery and searching. *Nucleic Acids Research*, 37(SUPPL. 2):202–208, 2009.

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