# Automatic learning of pre-miRNAs from different species

Ivani de O. N. Lopes*[1,2,3] and Alexander Schliep[2]and André P. de L. F. de Carvalho[3]

[1]Empresa Brasileira de Pesquisa Agropecuária, Embrapa Soja,
Caixa Postal 231, Londrina-PR, CEP 86001-970, Brasil
[2]Department of Computer Science and BioMaPs Institute for Quantitative Biology
Rutgers University
110 Frelinghuysen Road, Piscataway, NJ, 08854, USA
[3]Instituto de Ciências Matemáticas e de Computação
Avenida Trabalhador são-carlense, 400 - Centro, São Carlos - SP, Brasil

Email: Ivani de O. N. Lopes*- ivani.negrao@embrapa.br; Alexander Schliep - schliep@cs.rutgers.edu; André C. P. de L. F. de Carvalho - andre@icmc.usp.br;

*Corresponding author

## Abstract

**Background**: As microRNAs (miRNAs) genes are short sequences, and, by themselves, uninformative, miRNAs discovery tools usually have an embedded predictive model that uses characteristic features from miRNA precursors (pre-miRNAs) for the identification of miRNAs. To accommodate the peculiarities of plant and animal miRNAs systems, tools for both systems have evolved differently. However, these tools are biased towards the species for which they were primarily developed and, consequently, their predictive performance on data sets from other species of the same kingdom may not hold. While these biases are intrinsic to the species, the characterization of their occurrence can lead to computational approaches able to diminish its negative effect in the accuracy of pre-miRNAs predictive models. For such, we investigate in this study how 45 predictive models induced for data sets from 45 species, distributed in eight subphyla/classes, perform when applied to a species different from the species used in its induction.

**Results**: Our experimental analyzes show that the classification complexity is species-dependent and no feature set can uniquely represent instances from all species, even those from the same subphylum/class. To further support this statement, we show that an ensemble of classifiers reduced the classification errors for all 45 species. As the ensemble members were obtained using meaningful, and yet computationally viable feature sets, the ensembles also have a lower computational cost than single classifiers that relied on energy stability parameters,

which are distinct pre-miRNAs features, but of prohibitive computational cost in large scale applications.

**Conclusion**: In this study, the combination of multiple pre-miRNAs feature sets and multiple learning biases enhanced the predictive accuracy of pre-miRNAs classifiers of 45 species. This is certainly a promising approach to be incorporated in miRNA discovery tools towards more accurate and less species-dependent tools.

## Background

**MicroRNAs** (miRNAs) constitute one of the most widely-studied class of endogenous small (approx. 22 nucleotides) non-coding RNAs genes, due to their regulatory role in pos-transcription gene regulation in animals, plants and fungi [1,2]. The miRNAs biogenesis involves the participation of several enzymes, which depend on the origin (e.g. intergenic or intronic miRNAs) and on the kingdom of the species. However, all miRNAs are processed from long primary miRNA transcripts (pri-miRNAs), which are processed to hairpin-shaped intermediates (pre-miRNAs) and, subsequently, to a double strand RNA (miRNA:miRNA*) and a terminal loop. In the cytoplasm of animal and plant cells, the mature miRNA enters in the RNA-induced silencing complex (RISC) to silence target messenger RNAs (tmRNAs) by partial or near-perfect antisense complementarity. Partial antisense complementarity inhibits the translation of tmRNAs, whereas the later causes the degradation of tmRNAs. Reviews on biogenesis, diversification and evolution of miRNAs can be obtained at [2–4].

**RNAseq methods, followed by computational analysis**, became the *de facto* approach for miRNA discovery [4]. These methods, also called deep sequencing of transcriptome, can reveal the identities of most RNA species inside a cell, providing tens to hundreds of millions of sequence "reads" [5]. These reads provide both the sequence and the frequency of RNA molecules present in a cell. When applied to detect miRNAs, the RNA material is isolated through a procedure of size selection, such that only small reads (approx. 25 nt long) are sequenced [5]. The computational challenge consists in distinguishing miRNAs from other small RNA (sRNA) types and degradation products [4,6].

**The challenge of building a multi-species miRNA prediction tool** can be inferred from the sensitivity ranges estimated from eight deep sequencing miRNA prediction tools, when they were applied to data sets from *H. sapiens*, *G. Gallus* and *C. elegans* by [7]. Taking apart the average performance of each tool, the sensitivity ranges varied between 24% and 38%. For example, the sensitivity of the tool with the highest average sensitivity (68%) varied between 55% (*H. sapiens*) and 78% (*G. Gallus*) and the sensitivity of the tool with lowest average sensitivity (15%) varied between between 0% (*H. sapiens*) 25%

(*C. elegans*). The species bias is also present in the analysis performed with miRDeep2 [8], a newer version of miRDeep [6], which incorporated new features to increase the detection of known and novel miRNAs in all animal major clades. Even though the average sensitivity of miRDeep2 (80%) has clearly increased, compared to its first version, it still ranged from 71% (Sea squirt) to 90% (Anemone). In order to identify the source of these variabilities, it is imperative to explore how the main factors involved in the development of such computational tools vary throughout species.

**As miRNAs are processed from hairpin regions**, computational tools developed to predict miRNAs from RNA-seq libraries include at least four steps: pre-processing; read mapping to a reference genome; detection of energetically stable hairpins in the genomic region surrounding the mapped read and; detection of miRNAs biogenesis 'signature'. The latter is derived from the abundance and from the distribution of the reads across the hairpin and is fundamental to reduce false detections, since the hairpin shape structure is a necessary but not sufficient condition to process miRNA. Three criteria have been used as evidence of miRNAs biogenesis: a) the frequency of the mature strand is higher than the frequencies of the corresponding star and loop strands; b) the positions of the Drosha and Dicer cleavage sites in the 5' ends of the putative miRNA and miRNA* are nearly uniform and; c) the putative miRNA and miRNA* sequences align in the hairpin keeping approx. 2 nt overhang in the 3' end [4]. Nevertheless, the hairpin analysis is possibly the most critic step affecting negatively the sensitivity of the tools, since the biogenesis signature analysis is performed either after the selection of the energetically most favorable hairpin containing the mapped read stack (e.g. as in miRanalyzer [9]) or concomitantly with hairpin features (as in miRDeep2 [8]).

**The hairpin analysis** has been performed mostly through machine learning based predictive models. To obtain these models, a feature set (feature vectors) describing sequence and/or structural aspects of pre-miRNAs sequences (+) and hairpin like (-) sequences is extracted to create a training data set, which is subsequently fed to a machine learning algorithm. An investigation on human pre-miRNAs classifiers ndicated that the feature set, instead of the learning algorithm, had the major effect in the classification accuracy of the induced models [10]. However, it remained to be investigated the relevance of those features for the correct classification of pre-miRNAs from other species.

**Since miRNA systems in plants and animals differ substantially** [4], computational tools for plant and animal miRNAs discovery have been developed separately (eg. [9,11]). However, in practice, even instances of species from the same kingdom apparently diverge substantially regarding their intrinsic and extrinsic features. Therefore, in order to develop miRNA discovery tools less sensitive to species-specific characteristics, one first step is to determine if a unique feature set can capture the diversification of pre-

miRNAs throughout species. Moreover, it is important to establish the boundaries of the applicability of cross-species miRNAs predictive models, since the relevance of any tool depends on its ability to detect the miRNAs present in the data set under analysis. Another important aspect to be considered is the computational cost of extracting a feature set, since this cost can be prohibitive for some distinct pre-miRNAs features (e.g. energy stability parameters) it they are to be computed from millions of hairpins. These issues were addressed in this study, considering eight feature sets investigated in [10], three learning algorithms and 45 species representing eight subphyla/classes.

**Our experimental results showed that the classification complexity of pre-miRNAs is species-dependent**, albeit some feature sets and learning algorithms were more likely to maximize the predictive accuracy of pre-miRNAs classifiers for most species (first subsection of the Results and Discussion section). To interpret this dependency, in the following subsections, we analyzed how relevant are the features extracted from instances of one species for the classification of instances of other species. These analyzes indicated that pre-miRNAs classifiers restricted to predict instances of species from the same subphylum of the species used on its induction (training species), instead of the same kingdom, are more likely to achieve higher accuracies. Nevertheless, our results also showed that ensembles of classifiers obtained with computationally less costly feature sets reduced significantly the relevance of the subphylum membership of training species. Therefore, the relevance of the ensemble approach can be accounted by its potential to enhance the applicability of a pre-miRNA predictive model to a broader number of species, while keeping the computational cost close to that of single classifiers that also do not include computationally costly features.

## Material and Methods

### Experimental design

The analyzes carried out in this study were based on the accuracy of classifiers obtained in two steps: (1) create pre-miRNA data sets and (2) induce and test classifiers for classification of pre-miRNAs. In the step (1), for each species, 30 sequences from each class were randomly sampled from the pre-processed positive and negative sets to compose the test sets. From the remaining sequences, 60 sequences from each class were randomly sampled to construct the training set. Afterwards, all features were extracted from each sequence. This first step was repeated 10 times. As these data sets were built by species, they are also referred as training and test species. In the step (2), instances from all test sets were classified by the classifiers obtained with the training data built in the step (1). The accuracy of these classifiers were analyzed under the two-way analysis of variance (anova) equations 1 and 2.

The sizes of the training and test sets were, respectively, 2/3 and 1/3 of the smallest number of positive non-redundant sequences, shown in the Additional file 1. By fixing the sizes of training and test sets, we reduced the sources of random variations, i.e., variations that cannot be assigned to a main factor. However, since our main goal was to study the effect of the training species ($S$) in the predictive accuracy of pre-miRNAs classifiers, we considered the effects of the classification algorithm and the feature set in a unique factor, represented here by $M$. Therefore, considering three algorithms and eight feature sets, the number of levels of the factor $M$ is 24 (or 3× 8).

***Anova 1: $M \times S$***

The first analysis was performed to study the relationship $M \times S$ in order to identify the levels of $M$ that led to higher predictive accuracies for each species. For such, we considered the Equation 1, where the accuracies were estimated considering the same training and test species.

$$A_{ijk} = \mu + M_l + S_i + MS_{li} + e_{lik}, \tag{1}$$

such that:

$l = 1..24$ indexes the classifiers,

$i = 1..45$ indexes the species,

$k = 1..10$ indexes the repetition,

$A_{lik}$ = accuracy of the classifier $l$, obtained with the training species $i$ in the repetition $k$,

$\mu$ = overall mean accuracy,

$M_l$ = effect of the classifier $l$,

$S_i$ = effect of the species $i$,

$MS_{li}$ = interaction between the effects of the classifier $l$ and the species $i$, and

$e_{lik}$ = random error, or part of $A_{lik}$ that could not be assigned to the classifier $l$, the species $i$ and the repetition $k$; $e \sim N(0, \sigma^2)$.

***Anova 2: cross-species classifiers***

To investigate the suitability of instances from one species to build pre-miRNAs predictive models for other species, we fixed a classifier $l$, $l = 1..24$, and varied the training and test species. The accuracies were analyzed according to Equation 2:

$$A_{lijk} = \mu + M_{li} + T_j + MT_{lij} + e_{lijk}, \tag{2}$$

such that:

$l$ indexes one out the 24 classifiers,

$i, j = 1..45$ indexes training and test species,

$k = 1..10$ indexes the repetition,

$A_{lijk}$ = accuracy of the classifier $l$, obtained with data from the species $i$, in predicting the classes of instances from the species $j$ in repetition $k$,

$\mu$ = overall mean accuracy,

$M_{li}$ = effect of a species $i$,

$T_j$ = effect of the species $j$,

$MT_{lij}$ = effect of the interactions model species $l$ and test species $j$, and

$e_{lijk}$ = random error, or part of $A_{lijk}$ that could not be assigned to the species $i$, the test species $j$ and the repetition $k$; $e \sim N(0, \sigma^2)$.

### *Clustering algorithm*

The equations 1 and 2 are particularly useful to estimate the variance of random errors ($\sigma^2$). Once this variance is known, we can decide how typical are the variances estimated from the controlled factors (e.g. $M$, $S$ and $MS$), compared to $\sigma^2$, using the $p$-value obtained from the $F$-test. In this work, significant $p$-values were lower or equal to 0.05 ($p$<0.05). Since significant $p$-values of $F$-test on a factor only supports the inference that the at least two levels of that factor had different average effects, we applied the clustering algorithm Scott and Knott [12] to identify the levels of each factor in equations 1 and 2 that led to non-significantly different accuracies, using the R package ScottKnott [13].

## Data sets

### *Positive sequences*

To construct positive data sets, we downloaded all pre-miRNAs from miRBase release 20. This release contains 24,521 miRNA loci from 206 species, processed to produce 30,424 mature miRNA products [14]. However, only 65 species had at least 100 pre-miRNAs. From these 65 species, 48 had at least 90 non-redundant sequences (see criterion in the pre-processing subsection). Based on the availability of sequences that could be used to generate negative examples, positive sequences from only 45 species were considered. The identification of these species per phylum/division, subphylum/class, the acronyms used in their identification, the amount of available and non-redundant pre-miRNAs, the mean and the standard deviation of

their sequence length are shown in Table 6, Additional file 1.

***Negative sequences***

Negative data sets were constructed from a pool of 1,000 pseudo hairpins per species. These pseudo hairpins were excised from Protein Coding Sequences (CDS) or pseudo gene sequences, downloaded from the repositories: Metazome v3.0, Phytozome v9.0 or ncbi. The excision points were randomly chosen in the interval $[0, L - l_{pse} - 100]$, where $L$ was the sequence length of the CDS or pseudo gene and $l_{pse}$ was the length of the excised sequence. The number of pseudo hairpins of length $l_{pse}$ were determined in accordance with the length distribution of the available pre-miRNAs from each species. Afterwards, the excised sequence was evaluated for the resemblance with real pre-miNAs. Sequences that passed the criteria described in the items 1 to 4 below were stored as pseudo hairpins, and those that failed any of these criteria were discarded. These criteria were:

1. fold-back structure (FB);
2. $bp \geq 18$, $bp$ = base pairing;
3. $Q_{seq} \geq 0.9$, $Q_{seq}$ = sequence entropy;
4. Minimum Free Energy of folding ($MFE$) rules:
   $MFE_{l_{pse}} \leq -10.0$, if $l_{pse} < 70$
   $MFE_{l_{pse}} \leq -18.0$, if $70 < l_{pse} \leq 100$
   $MFE_{l_{pse}} \leq -25.0$, if $l_{pse} > 100$.

$Q_{seq}$ was used to filter out meaningless sequences, since genomic sequences are usually contiguously padded with "N" characters and the $MFE$ rules were derived to accommodated the correlation between $MFE$ and $L$.

***Pre-processing***

Genes in a miRNA family can have about 65% or more sequence identity [15]. Since the number of miRNA families is relatively small compared to the number of positive examples available, redundancy removal is an important pre-processing procedure to avoid overfitted predictive models. We used dnaclust [16] to remove redundant sequences, prior to the sampling of examples to compose training and test sets. With dnaclust, sequences in positive sets of each species were clustered such that the similarities between sequences within

a cluster were at least 80%. Afterwards, one sequence from each cluster was randomly sampled to construct the positive non-redundant sets. The same pre-processing procedure was applied to the sets of negative sequences, but very few sequences were removed from each set, since the excision from random positions reduces the chances of obtaining similar pseudo hairpins.

### Feature sets

The eight features sets primarily studied in this investigation were extensively evaluated on human sets by Lopes et al. [10]. Here, these feature sets are referred by the same notation ($FS_i$, $i \in \{1, .., 7\}$ and SELECT). These feature sets contain most of the features used in computational pipelines for pre-miRNA discovery. References of computational pipelines that used these feature sets and their composition can be seen in Table 1. This table also shows two important aspects of these feature sets: feature diversity and feature sets overlapping. For example, $FS_1$, $FS_2$, $FS_7$ and SELECT have 13 overlapping features, from which five are also in $FS_3$. Features in these sets are measures of different characteristics of the sequences, whereas the features in $FS_4$, $FS_5$ and $FS_6$ are mostly sequence-structure patterns.

### Learning algorithms

The learning algorithms used in this work were Support Vector Machines (SVMs), Random Forest (RF) and J48. These algorithms have different learning biases, which is important for the present work, since learning biases may favor a feature set over others. SVMs and RFs are the algorithms most used for pre-miRNA classification and J48 was chosen because of its simplicity and interpretability.

J48 implements the well known C4.5 algorithm [17]. As one of the most popular algorithm based on the divide-and-conquer paradigm, C4.5 recursively divides the training set into two or more smaller subsets, in order to maximize the information entropy. The J48 implementation builds pruned or unpruned decision trees from a set of labeled training data. We used RWeka [18], an R interface of Weka [19], with the default parameter values. RWeka induces pruned decision trees from a data set.

To train SVMs, we used a Python interface for the library libsvm 3.12 [20]. This interface implements the C-SVM algorithm using the RBF kernel. The kernel parameters $\gamma$ and $C$ were tuned by 5-fold cross validation (CV) over the grid $(C; \gamma) = (2^{-5}, 2^{-3}, ..., 2^{15}; 2^{-15}, 2^{-13}, ..., 2^{3})$. The pair $(C; \gamma)$ that led to the highest CV predictive accuracy in the training subsets was used to train the SVMs using the whole training set. The resulting classifier was applied to classify the instances from the corresponding test set.

RF ensembles were induced over the grid (30, 40, 50, 60, 70, 80, 90, 100, 150, 250, 350, 450)×[ (0.5, 0.75,

1, 1.25, 1.5)*$\sqrt{d}$ ], representing respectively the number of trees and the number of features. The value $\sqrt{d}$ is the default number of features tried in each node split, where $d$ is the dimension of the feature space or the number of features in the feature set. We chose the ensemble with the lowest generalization error over the grid, according to the training set, and applied it to classify the instances of the corresponding test set. The ensembles were obtained using the *randomForest* R package [21] in *in house* R pipeline.

***Ensembles and other feature sets***

Besides the predictive accuracy, the applicability of any pre-miRNA classifier to larger data sets may be limited by the computational time necessary to compute the feature set representation of each pre-miRNA candidate. To increase the predictive accuracy while keeping the computational cost under feasible limits, subsets of the existing features sets, removing features computed from shuffled sequences, were employed to construct ensemble of classifiers. These subsets were named Ss1 and Ss7, such that: Ss1 = $\mathrm{FS}_1 - \{zG, zP, zQ, zD, zF\}$ and Ss7 = $\{orf, \%LCRs, loops, A_{(((}, C_{(((}, G_{(((}, U_{(((}\}$. Ss1 features measure the largest variety of pre-miRNA characteristics, whereas Ss7 combine features widely used in pre-miRNA classification ($A_{(((}, C_{(((}, G_{(((}, U_{(((}$) with three features introduced in pre-miRNA classification in [22]. The first subset was evaluated individually, and combined with the latter ($\mathrm{Hyb}_{17} = Ss7 \cup Ss1$). The subset Ss7 was also combined with the feature sets $\mathrm{FS}_3$ ($\mathrm{Hyb}_{37} = \mathrm{FS}_3 \cup \mathrm{Ss7}$) and SELECT ($\mathrm{Hyb}_{S7} = \mathrm{Ss7} \cup \mathrm{SELECT}$). The prefix Hyb is used to represent these 'hybrid' feature sets.

An ensemble of classifiers combine the prediction of a set of individual classifiers. The ensembles used in this study are described in Table 2, along with all other classifiers investigated. The computational time for the extraction of the feature sets used in the ensembles are close to the time spent to extract the feature set SELECT and presented in [10]. As shown in this table, the final prediction of the ensembles were defined by the majority vote (ensemble Emv) and weighted vote (ensemble Ewv). Each ensemble, therefore, combines the class prediction, class vote, from each one of its classifiers. In the first approach, the class predicted by the majority of the classfiers is the ensemble class prediction. In the weighted approach, the vote of each classifiers was weighted by its predictive accuracy in the training set. Ties were resolved by random choice.

## Results and discussions

### Predictive accuracy of pre-miRNA classifiers by species

As the $F$-test on the effect of $MS$ in Equation 1 was highly significant ($p < 0.001$), the effect of the simple factor $M$ was studied within fixed levels of $S$ ($\mathrm{M/S}_j$, $j = 1..45$), and vise-verse ($\mathrm{S/M}_l$, $l = 1..24$). The

analyzes of M/S$_j$, $j = 1..45$, are summarized in Figure 1 and Table 3. The green bars in Figure 1 indicate the pre-miRNA classifiers whose obtained accuracy is within the cluster of maximal accuracies $C_1$. As indicated in Figure 1, SVMs and RFs obtained using the feature sets FS$_3$, FS$_6$, FS$_7$ and SELECT achieved accuracies within $C_1$ for most species. These results agree with the results reported in [10], which used larger training and test sets of human instances.

Figure 1 indicates only the algorithms and feature set combinations more likely to produce pre-miRNAs classifiers of maximal accuracy, but the maximal depends on the species, as it can be observed in Table 3. According to this table, the mean accuracy in $C_1$ varied from 86% (cin) to 96% (ssc). As the clusters were obtained for each species using the estimated accuracies of the same 24 classifiers and the number of clustes varied from two (bfl, dme, hsa, ath, lus, mdm, ptc, osa, zma) to five (gga), Table 3 indicates that either the instances from some species are easier to classify than instances from other species, or that pre-miRNAs of different species carry specific features that identify related characteristics. In both cases, these results indicate that the incorporation of intrinsic characteristics of the species could improve the accuracy of pre-miRNAs predictive models in the classification of sequences from different species.

Table 4 presents the results of the analyzes of S/M$_l$, $l = 1..24$. Similar to what was observed in the analyzes of M/S$_j$, $j = 1..45$, the number of clusters and the corresponding centers depended on the levels of $M$. However, the number of clusters and the accuracy intervals (Range columns) in both tables show that the effect of $S$ in the accuracy of pre-miRNA classifiers is broaden than the effect of $M$. For example, the number of clusters in Table 4 varied from two to six and the ranges varied from 14% (FS$_7$-RFs) to 41% (FS$_1$-J48). Moreover, although the average accuracies estimated from 17 out of 24 pre-miRNA classifiers were above 95% for some species (column $c_1$), the average accuracies of the same level $M_i$ for other species were as low as 57%. In fact, no $M_l$, $l = 1..24$ led to classifiers of accuracies within $c_1$ for all species, supporting again the conjecture that the learning complexity of pre-miRNAs is species-dependent.

In the next subsection, we discuss the representativeness of the instances from the 45 species considered in this work for the induction of classifiers able to predict the classes of each other's instances, given a classification algorithm and a feature set. In addition, we discuss the occurrence of species-specific features and their effect in the predictive accuracy of cross-species pre-miRNAs classifiers.

#### Cross-species pre-miRNAs classifiers: M$_l$ × T

Given a learning algorithm and a feature set, the relevance of the instances of a species $i$ (training species) in the prediction of instances from a species $j$ (test species), $i \neq j$, can be inferred from the effects of the

factors in Equation 2. Since the $F$-test on the interaction $M_lT$ was significant ($p < 0.05$), the factor $M_l$ was analyzed within each level of the factor $T$ ($M_l/T_j, j = 1..45$), and vise-verse ($T/M_{li}, i = 1..45$). The results of the analyzes of $M_l/T_j, j = 1..45$ indicate the training species that resulted in pre-miRNA classifiers of higher accuracies ($c_1$) for each test species. From the results of the analyzes of $T/M_{li}, i = 1..45$, we discussed the learning complexity of pre-miRNAs from the 45 species.

***Choosing the training species - $M_l/T$***

By clustering the average accuracies $\bar{A}_{lij.}$, within $j$, $i, j = 1..45$, we identified the training species $i$ that led to accuracies within $c_1$ for each test species $j$. Figure 2 shows these cases in green ($c_1$) and red ($c_2$,...,$c_6$), where $i$ is shown in the $Y$-axis and $j$ in the $X$-axis. The results for the other 20 models were similar. As the black frames enclose species from the same subphylum/class and within each frame the green pixels are more numerous than the red ones, we conclude that a pre-miRNAs classifiers was more likely to achieve predictive accuracies within $c_1$ when the species $i$ and $j$ were from the same subphylum/class. In particular, $\bar{A}_{lij.}$ were always assigned $c_1$ when $i = j$ (diagonal), indicating that species-specific classifiers is a good approach to improve the predictive accuracy of pre-miRNAs predictive models.

Figure 2 also shows that instances from some species were systematically harder to classify than instances from other species, which can be inferred through the number of red pixels by column. Among them, instances from bmo were typically harder to classify than instances from other species. The columns showing the clusters associated with different training species in the classification of instances from B. mori (bmo) and L. usitatissimum (lus) illustrate these cases. Particularly, the average of the clusters obtained from SVMs_SELECT classifiers generated with instances of all species in predicting the classes of bmo instances were 80% ($c_1$), 70% ($c_2$) and 65% ($c_3$), whereas the corresponding measures for lus were 98% ($c_1$), 93% ($c_2$), 89% ($c_3$), 80% ($c_4$) and 65% ($c_5$).

Although the phylogenetic proximity of training and test species is fundamental to obtain pre-miRNAs classifiers of higher accuracies, the learning biases of the classification algorithm may increase or decrease the relevance of the subphylum/class membership, as Figure 2 shows. In this figure, SVMs were more sensitive to the phylogenetic proximity of training and test species. An interpretation for this pattern is provided in the subsection Feature importance.

***Inferring learning complexity - $T/M_l$***

In these comparisons, we clustered the accuracies estimated from all test sets, fixing the training species and a level of $M$. These clusters are displayed in Figure 3, for four levels of $M$. In this figure, a row shows the test species ($X$-axis) assigned to the cluster $c_1$ (green) or to another cluster (orange), when its instances were classified using a training species $i$ ($Y$-axis). The highest quantities of green pixels clearly associated with the Angiosperm test species suggest that instances from Angiosperm test species were easier to classify than instances from other test species, particularly vertebrates.

Although this pattern was consistent in all 24 level of $M$, we also looked into the learning complexity by analyzing the importance of the 85 unique features in the classification of instances from all species. The idea was to indirectly compare the similarities between the instances from different species, using a feature importance measure obtained during the induction of RF classifiers. These results are discussed next.

## Feature importance

Given a feature set, the importance of each feature for the correct classification of the test set instances can be estimated by a feature importance measure, which in this work was taken from the RF results. The rationale of investigating the relevance of the RNA features used in this work for the correct classification of pre-miRNAs of different species is to infer, at least indirectly, if the phylogenetic proximity of these species is a valid criterion to choose a feature set.

The feature importance measure ($FI$) used in this study estimates the increase of misclassified OOB (Out-Of-Bag) instances when that feature is permuted in the training vectors. Since that measure is an absolute value, to allow its comparison for different classifiers induced with instances of different species, its values were re-scaled to the interval [0, 1] by the formula $RFI = (FI - FI_{min})/(FI_{max} - FI_{min})$. The maximum ($FI_{max}$) and minimum ($FI_{min}$) $FI$ values were obtained from the subset of features used in the induction of each pre-miRNA classifier. We estimated the $RFI$ values for each of the 85 unique features considered in this work feature, when they were simultaneously fed to the RF algorithm to induce pre-miRNA classifiers for each of the 45 species. These estimates were discussed based on two criteria: the pairwise Pearson correlation coefficient between species and the distributions of the $RFI$ for the 45 species.

***Pearson correlation coefficients of RFIs between species***

Figure 7 shows the pairwise Pearson correlation coefficients of $RFI$ for all pairs of species. These correlations are in the interval [0, 1], where the black pixels indicate zero correlation and the white pixels indicate

correlation one. Therefore, white or light gray pixels represent the cases where the pre-miRNAs of the two corresponding species shared most of the features. As the red frames indicate, these cases are more likely if the two species are from the same subphylum/class. However, there are many exceptions within and outside the subphylum/class umbrella. For example, with few exceptions (e.g. ame, bmo and bta), the features that are important for the correct classification of instances from the species bfl, cin, cbr, cel and aae, were also important for the correct classification of instances from other species. Differently, the difficulty in establishing a general rule on the association between phylogenetic proximity and feature conservation using the $RFI$ criteria can be observed by the majority of dark pixels associated with Hexapoda species. This exceptions and the features with the highest $RFI$ are presented next.

***RFI distributions***

The $FRI$ distributions are shown in Table 5, omitting the cases where $RFI \leq 0.1$ for all species. According to this table, only 40% of the features met this criterion. Among them, $p$ and $MFEI_1$ obtained $RFI$ larger than 0.6 for 89% ($p$) and 94% ($MFEI_1$) of the species, whereas the $FRI$ distributions of the other features were closer to a right-tailed distribution. In fact, the $RFI$ estimates of 75 out of 85 of features were lower than 0.3 for 80% of the species. These small amount of highly relevant features helps to interpret the tendency of SVMs to reduce the predictive accuracy when the training and the test species were more distantly related, as those from Chordate and Angiosperm (Figure 2). Since SVMs use the full feature space and RFs use only subspaces of it, the classification by RFs may have been dominated by features that are more conserved throughout species. The interactions between the learning biases and the species is also analyzed through the classification errors of the three learning algorithms in the next subsection.

**Classification error**

The classification errors of a particular instance by different classifiers can provide information on how typical that instance is, assuming that atypical instances or outliers are more likely to be misclassified by most classifiers. Moreover, the classification errors estimated from test sets of instances from different species by multiple classifiers is also informative of the separability of classes, in the instance space of each species. To facilitate the notation, the errors $e_1, e_2, .., e_7$ are defined as exclusive classification errors of SVM ($e_1$), RF ($e_2$), J48 ($e_3$), SVM and RF ($e_4$), SVM and J48 ($e_5$), RF and J48 ($e_6$) and SVM and RF and J48 ($e_7$). Since $e_1, .., e_7$ are exclusive errors, they sum one or 100%, symbolically: $\sum_{i=1}^{7} e_i = 1$ or $\sum_{i=1}^{7} e_i = 100\%$. These errors are shown in Figure 4, for $FS_1$, $FS_6$ and SELECT.

As can be observed in Figure 4, the error distributions were strongly dependent on the species, which shows in another way the classification biases associated with species sequence data. For example, Figure 4 (a) shows that $e_1$ was zero for 15 species (cbr, tca, aca, gga, eca, ggo, ptr, cgr, ppt, aly, ath, mdm, ptc, osa, zma). Nevertheless, this same figure also shows $e_1$ of up to 80% for other species (e.g., bfl, cin, ame, mtr, stu, sbi). In these cases, and others where the exclusive error of a classifier induced by one of the three algorithms is higher than the errors achieved simultaneously by at least two classifiers induced by different algorithms, the separability of the classes is a matter of choosing an algorithm with the appropriate learning bias. On the other hand, the cases where $e7 > 50\%$ (e.g. mdm) could be better described by other feature spaces or by a combination of subspaces.

To summarize, the classification errors in each feature space, the errors $e_1, .., e_7$, were summed up for the 45 species and represented in Venn diagrams. Figure 6 shows the cases $FS_1$, $FS_6$, $FS_7$ and SELECT. The interaction between learning algorithm and feature set, indicated by the significant variation of the areas of the circles, between feature sets, is the most noticeable pattern in this figure. For example, classification models induced by J48 tended to achieve higher exclusive error rates ($e_3$) in higher dimensional feature spaces. Moreover, the proportion of instances misclassified simultaneously by classifiers induced by the three algorithms varied varied between 3.2% to 6.7% ($3.7\% \leq e_7 \leq 6.7\%$), which is a 3.5% interval. These two factor alone are sufficient to conjecture that the combination of multiple hypotheses may lead to pre-miRNA classifiers of higher accuracies than a single hypothesis, for a larger number of species. To provide a preliminary insight on this conjecture, we carried out additional computational experiments, using ensemble approaches to combine multiple hypothesis to improve the predictive accuracy of pre-miRNA classifiers. These results from these experiments are presented and discussed in the next subsection.

### Ensembles

Figure 5 shows the comparisons between the 44 classifiers, as defined in Table 2. According to Figure 5, the ensembles Emv24, Ewv24, Emv8-RF, Emv8-SVMs, Ewv8-RF, Ewv8-SVMs, Ewv24 and the classifiers obtained with the new feature sets presented better predictive accuracies than the 24 previously discussed, for many species, although none of the them achieved predictive accuracies within $C_1$ for all 45 species. Moreover, it is important to remind that these ensembles and the new feature sets do not include features extracted from shuffled sequences. Figure 5 also shows that the simple combination of different hypotheses can increase the predictive accuracy, even using the algorithm J48, which typically led to equal or lower classification accuracies than RFs and SVMs.

Based on the results shown in figures 5 and 7, and in Table 5, we can state that it is unlike that a unique learning algorithm and a unique set of features is able to produce the best pre-miRNA predictive model for all species. In fact, the experimental results obtained in this study suggested that the learning of good predictive models for pre-miRNAs classification depends on the learning complexity inherited of the problem and the peculiarities of the instances from different species. Since ensembles apparently provide an alternative and efficient approach to accommodate these peculiarities, an appropriate construction of hypothesis diversity (e.g. [23]) may enhance the performance of miRNA discovery tools in the classification of pre-miRNAs of different species.

## 1 Conclusion

The computational analysis of large amounts of sequencing data to detect miRNAs has increased the sequence analysis capacity and supported the recent advances in the discovery of novel miRNAs from over a hundred species. Albeit miRNA systems vary throughout species, miRNA discovery tools from the literature have not addressed the impact of these differences in the performance. As a consequence, the performance of these tools is usually reduced when data sets from species not used in their development are analyzed. For different reasons, to build species-specific miRNA discovery tools may not be viable. Since the detection of putative pre-miRNAs is an important step in the development of miRNA discovery tools, it is important to investigate how the peculiarities naturally occurring in pre-miRNAs throughout species relates with the learning bias of machine learning approaches. In this study, we presented the results of a systematic investigation on the automatic learning of pre-miRNAs of 45 species, using techniques traditionally employed by miRNA discovery tools from the literature. The results presented in this study not only showed the need to develop new approaches to handle the intrinsic characteristics of pre-miRNAs from different species, but we also indicated the way to go.

## 2 Competing interests

The authors declare that they have no competing interests.

## Author's contributions

AS and AC conceived and supervised the study. IL assembled the data, implemented the scripts, ran the experiments and summarized the results. The three authors wrote and approved the final manuscript.

## Acknowledgements

We thank Empresa Brasileira de Pesquisa Agropecuária (Embrapa Soybean) for the continuum financial support to the first author.

## References


1. Lim LP, Lau NC, Weinstein EG, Abdelhakim A, Yekta S, Rhoades MW, Burge CB, Bartel DP: **The microRNAs of Caenorhabditis elegans**. *Genes & Development* 2003, **17**(8):991–1008.
2. Westholm JO, Lai EC: **Mirtrons: microRNA biogenesis via splicing.** *Biochimie* 2011, **93**(11):1897–904, [http://dx.doi.org/10.1016/j.biochi.2011.06.017].
3. Bartel DP: **MicroRNAs: genomics, biogenesis, mechanism, and function.** *Cell* 2004, **116**:281–297.
4. Berezikov E: **Evolution of microRNA diversity and regulation in animals.** *Nature reviews. Genetics* 2011, **12**(12):846–60, [http://dx.doi.org/10.1038/nrg3079].
5. Chu Y, Corey DR: **RNA sequencing: platform selection, experimental design, and data interpretation.** *Nucleic Acid Therapeutics* 2012, **22**(4):271–4, [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3426205&tool=pmcentrez&rendertype=abstract].
6. Friedländer MR, Chen W, Adamidi C, Maaskola J, Einspanier R, Knespel S, Rajewsky N: **Discovering microRNAs from deep sequencing data using miRDeep.** *Nature Biotechnology* 2008, **26**(4):407–15, [http://dx.doi.org/10.1038/nbt1394].
7. Li Y, Zhang Z, Liu F, Vongsangnak W, Jing Q, Shen B: **Performance comparison and evaluation of software tools for microRNA deep-sequencing data analysis.** *Nucleic Acids Research* 2012, **40**(10):4298–305, [http://nar.oxfordjournals.org/content/early/2012/01/28/nar.gks043.full].
8. Friedländer MR, Mackowiak SD, Li N, Chen W, Rajewsky N: **miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades.** *Nucleic Acids Research* 2012, **40**:37–52, [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3245920&tool=pmcentrez&rendertype=abstract].
9. Hackenberg M, Rodríguez-Ezpeleta N, Aransay AM: **miRanalyzer: an update on the detection and analysis of microRNAs in high-throughput sequencing experiments.** *Nucleic Acids Research* 2011, **39**(Web Server issue):W132–8, [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3125730&tool=pmcentrez&rendertype=abstract].
10. Lopes IDON, Schliep A, Carvalho APDLFD: **The discriminant power of RNA features for pre-miRNA recognition**. *BMC Bioinformatics* 2014, **15**:124, [http://www.biomedcentral.com/1471-2105/15/124].
11. Yang X, Li L: **miRDeep-P: a computational tool for analyzing the microRNA transcriptome in plants.** *Bioinformatics* 2011, **27**(18):2614–5, [http://www.ncbi.nlm.nih.gov/pubmed/21775303].
12. Scott AJ, Knott M: **A cluster analysis method for grouping means in the analysis of variance**. *Biometrics* 1974, **30**(3):507–512, [http://dx.doi.org/10.2307/2529204].
13. Jelihovschi EG, Faria JC, Allaman IB: *The ScottKnott Clustering Algorithm*. Universidade Estadual de Santa Cruz - UESC, Ilheus, Bahia, Brasil 2013.
14. Kozomara A, Griffiths-Jones S: **miRBase: annotating high confidence microRNAs using deep sequencing data.** *Nucleic Acids Research* 2014, **42**(Database issue):D68–73, [http://nar.oxfordjournals.org/content/42/D1/D68].
15. Kamanu TKK, Radovanovic A, Archer JAC, Bajic VB: **Exploration of miRNA families for hypotheses generation.** *Scientific Reports* 2013, **3**:2940, [http://www.nature.com/srep/2013/131015/srep02940/full/srep02940.html].
16. Ghodsi M, Liu B, Pop M: **DNACLUST: accurate and efficient clustering of phylogenetic marker genes.** *BMC Bioinformatics* 2011, **12**:271+, [http://dx.doi.org/10.1186/1471-2105-12-271].
17. Quinlan JR: *C4.5: programs for machine learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. 1993.
18. Hornik K, Buchta C, Zeileis A: **Open-Source Machine Learning: R Meets Weka**. *Computational Statistics* 2009, **24**(2):225–232.
19. Witten IH, Frank E: *Data Mining: Practical Machine Learning Tools and Techniques*. San Francisco: Morgan Kaufmann, 2nd edition 2005.
20. Chang CC, Lin CJ: **LIBSVM: A library for support vector machines**. *ACM Transactions on Intelligent Systems and Technology* 2011, **2**:27:1–27:27. [Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm].

21. Liaw A, Wiener M: **Classification and Regression by randomForest**. *R News* 2002, **2**(3):18–22, [http://CRAN.R-project.org/doc/Rnews/].

22. Gudyś A, Szcześniak MW, Sikora M, Makalowska I: **HuntMi: an efficient and taxon-specific approach in pre-miRNA identification**. *BMC Bioinformatics* 2013, **14**:83, [http://www.biomedcentral.com/1471-2105/14/83].

23. Ding J, Zhou S, Guan J: **MiRenSVM: towards better prediction of microRNA precursors using an ensemble SVM classifier with multi-loop features.** *BMC Bioinformatics* 2010, **11 Suppl 1**:S11, [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3024864&tool=pmcentrez&rendertype=abstract].

24. Batuwita R, Palade V: **microPred: effective classification of pre-miRNAs for human miRNA gene prediction**. *Bioinformatics* 2009, **25**(8):989–995.

25. Hsieh CH, Chang DTH, Hsueh CH, Wu CY, Oyang YJ: **Predicting microRNA precursors with a generalized Gaussian components based density estimation algorithm.** *BMC Bioinformatics* 2010, **11 Suppl 1**:S52, [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3009525&tool=pmcentrez&rendertype=abstract].

26. Xue C, Li F, He T, Liu GP, Li Y, Zhang X: **Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine.** *BMC Bioinformatics* 2005, **6**.

27. Liu X, He S, SkogerbøG, Gong F, Chen R: **Integrated sequence-structure motifs suffice to identify microRNA precursors.** *Plos One* 2012, **7**(3):e32797, [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3305290&tool=pmcentrez&rendertype=abstract].

28. Jiang P, Wu H, Wang W, Ma W, Sun X, Lu Z: **MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features**. *Nuc. Ac. Res.* 2007, **35**(suppl 2):W339–W344.

## Figures

**Figure 1 - .**

**Figure 2 - A pixel ($i$,$j$) represent the outcome of a classifier induced with instances of a model species $i$ ($y$-axis) in predicting the classes of instances of a $j$ ($x$-axis); black frames encloses species from the same subphylum/class.**

**Figure 3 - A pixel ($i$,$j$) represent the outcome of a classifier induced with instances of a model species $i$ ($y$-axis) in predicting the classes of instances of a $j$ ($x$-axis); black frames encloses species from the same subphylum/class.**

**Figure 4 - Distribution of classification errors per species. Exclusive errors by SVMs ($e_1$), RF ($e_2$), J48 ($e_3$), SVMs and RF ($e_4$), SVMs and J48 ($e_5$), RF and J48 ($e_6$) and SVMs and RF and J48 ($e_7$).**

**Figure 5. Distribution of the accuracies of 44 classifiers within the accuracy clusters. Mean$_{C_1}$ > ... > Mean$_{C_5}$.**

**Figure 6 - Venn diagram of the classification errors of the classification algorithms, by feature set. Results were obtained from the classification of 27000 = 45 (test species) ×10 (repetitions) ×60 (30+,30-).**

**Figure 7 - Pairwise Pearson correlation coefficient of $FRI$ throughout species.**

## Tables

**Table 1 - Feature set composition, dimension, literature reference and associated literature tool.**

**Table 2 - Centers of accuracy clusters from 24 classification models, per species.**

**Table 3 - Centers of accuracy clusters obtained from classification models induced with examples from different species, per combination of feature set and learning algorithm.**

## Additional Files

**Additional file 1 — Species description**

Table 1: Feature set composition, dimension, literature reference and associated literature tool.

| FEATURE | FEATURE SET | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | $FS_1$ | $FS_2$ | $FS_3$ | $FS_4$ | $FS_5$ | $FS_6$ | $FS_7$ | SELECT |
| Di-nucleotide frequencies ($XY, X, Y \in \{A, C, U, G\}$) | x | | | | | | | |
| $\%G + C$ | x | x | | | | | x | |
| Maximal length of the amino acid string without stop codons ($orf$) | | | | | | | x | |
| Percentage of low complexity regions ($\%LCRs$) | | | | | | | x | |
| Triplets | | | | x | | x | | |
| Stacking triplets ($X_{(((}$, $X \in \{A, C, G, U\}$) | | | | | | | x | |
| Motifs ($ss-$substrings) | | | | | x | | | |
| Minimum free energy of folding ($MFE$) | | | | | | x | | |
| Randfold ($p$) | | | | | | x | | |
| Normalized MFE ($dG$) | x | x | x | | | | x | x |
| MFE index 1 ($MFEI_1$) | x | x | x | | | | x | x |
| MFE index 2 ($MFEI_2$) | x | x | x | | | | x | x |
| MFE index 3 ($MFEI_3$) | x | x | | | | | x | x |
| MFE index 4 ($MFEI_4$) | x | x | | | | | x | |
| Normalized Ensemble Free Energy ($NEFE$) | x | x | | | | | x | x |
| Normalized difference ($MFE - EFE$) ($Diff$) | x | x | | | | | x | x |
| Frequency of the MFE structure ($Freq$) | x | | | | | | | |
| Normalized base-pairing propensity ($dP$) | x | | x | | | | | |
| Normalized Shannon entropy ($dQ$) | x | x | x | | | | x | x |
| Structural diversity ($Diversity$) | x | x | | | | | x | |
| Normalized base-pair distance ($dD$) | x | | x | | | | | |
| Average base pairs per stem (Avg_Bp_Stem) | x | x | | | | | x | |
| Normalized A-U pairs counts ($\|A - U\|/L$) | x | x | | | | | x | |
| Normalized G-C pairs counts ($\|G - C\|/L$) | x | x | | | | | x | x |
| Normalized G-U pairs counts ($\|G - U\|/L$) | x | x | | | | | x | x |
| Content of A-U pairs per stem ($\%(A - U)/stems$) | x | x | | | | | x | |
| Content of G-C pairs per stem ($\%(G - C)/stems$) | x | x | | | | | x | |
| Content of G-U pairs per stem ($\%(G - U)/stems$) | x | x | | | | | x | x |
| Cumulative size of internal loops ($loops$) | | | | | | | x | |
| Structure entropy ($dS$) | x | x | | | | | x | x |
| Normalized structure entropy ($dS/L$) | x | x | | | | | x | x |
| Structure enthalpy ($dH$) | x | | | | | | | |
| Normalized structure enthalpy ($dH/L$) | x | | | | | | | |
| Melting energy of the structure | x | | | | | | | |
| Normalized melting energy of the structure | x | | | | | | | |
| Topological descriptor (dF) | x | x | x | | | | x | x |
| Normalized variants ($zG$, $zP$ and $zQ$) | x | | | | | | | |
| Normalized variants ($zD$) | x | x | | | | | x | |
| Normalized variants ($zF$) | x | | | | | | | |
| DIMENSION | 48 | 21 | 7 | 32 | 1300 | 34 | 28 | 13 |
| REFERENCE | [24] | [24] | [25] | [26] | [27] | [28] | [22] | [10] |

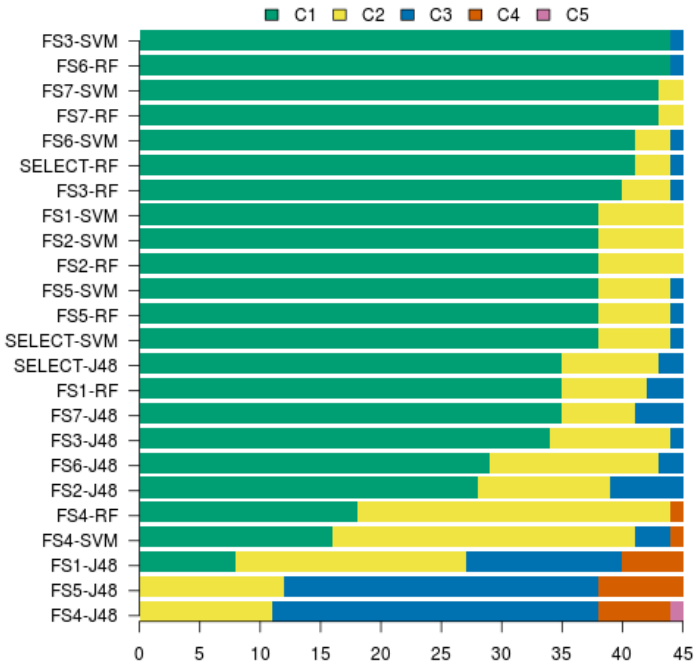



Figure 1: Frequencies of species for who each classification model achieved accuracies in the clusters $C_1$-$C_5$. $\text{Mean}_{C_1} \geq .. \geq \text{Mean}_{C_5}$.

Table 2: Definition of all 44 classification models compared in this work, according to feature sets and learning algorithms. $M_{ij}$ is the classifier induced with the feature set $i$ and algorithm $j$, $i = 1..12$ and $j = 1, 2, 3$, and $w_{ij}$ is the accuracies of the classifier $M_{ij}$. $\hat{M}_{ij}$ is the predicted class by $M_{ij}$, $\hat{M}_{ij} \in \{-1, 1\}$. Emv=Ensemble majority votes, Ewv=Ensemble weighted votes.

| | 1. SVMs | 2. RF | 3. J48 |
|---|---|---|---|
| 1. $FS_1$ | $M_{11}$ | $M_{12}$ | $M_{13}$ |
| 2. $FS_2$ | $M_{21}$ | $M_{22}$ | $M_{23}$ |
| 3. $FS_6$ | $M_{31}$ | $M_{32}$ | $M_{33}$ |
| 4. $FS_7$ | $M_{41}$ | $M_{42}$ | $M_{43}$ |
| 5. $FS_3$ | $M_{51}$ | $M_{52}$ | $M_{53}$ |
| 6. $FS_4$ | $M_{61}$ | $M_{62}$ | $M_{63}$ |
| 7. $FS_5$ | $M_{71}$ | $M_{72}$ | $M_{72}$ |
| 8. SELECT | $M_{81}$ | $M_{82}$ | $M_{83}$ |
| 9. $Hyb_{37}$ | $M_{91}$ | $M_{92}$ | $M_{93}$ |
| 10. $Hyb_S7$ | $M_{101}$ | $M_{102}$ | $M_{103}$ |
| 11. $Hyb_17$ | $M_{111}$ | $M_{112}$ | $M_{113}$ |
| 12. $Ss_1$ | $M_{121}$ | $M_{122}$ | $M_{123}$ |
| Emv8 | $\sum \hat{M}_{i1}$, $i = 5..12$ | $\sum \hat{M}_{i2}$, $i = 5..12$ | $\sum \hat{M}_{i3}$, $i = 5..12$ |
| Ewv8 | $\sum w_{i1}\hat{M}_{i1}$, $i = 5..12$ | $\sum w_{i2}\hat{M}_{i2}$, $i = 5..12$ | $\sum w_{i3}\hat{M}_{i3}$, $i = 5..12$ |
| Emv24 | | $\sum \hat{M}_{ij}$, $i = 5..12$ and $j = 1, 2, 3$ | |
| Ewv24 | | $\sum w_{ij}\hat{M}_{ij}$, $i = 5..12$ and $j = 1, 2, 3$ | |

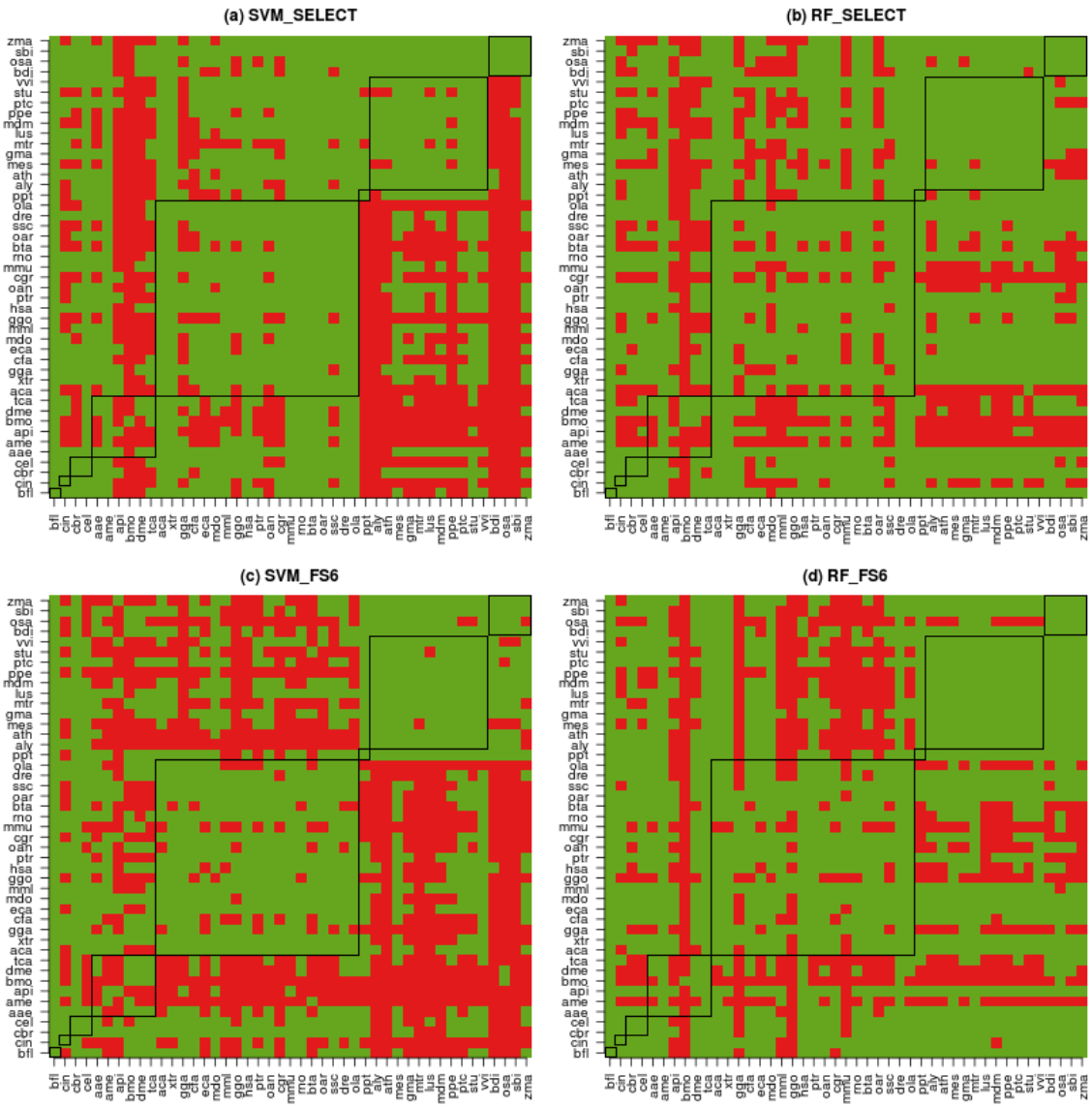



Figure 2: Accuracy cluster membership for cross-species pre-miRNAs classifiers. Green= $c_1$; red=other; $y$-axis=model species; $x$-axis=test species; black frames encloses species from the same subphylum/class.

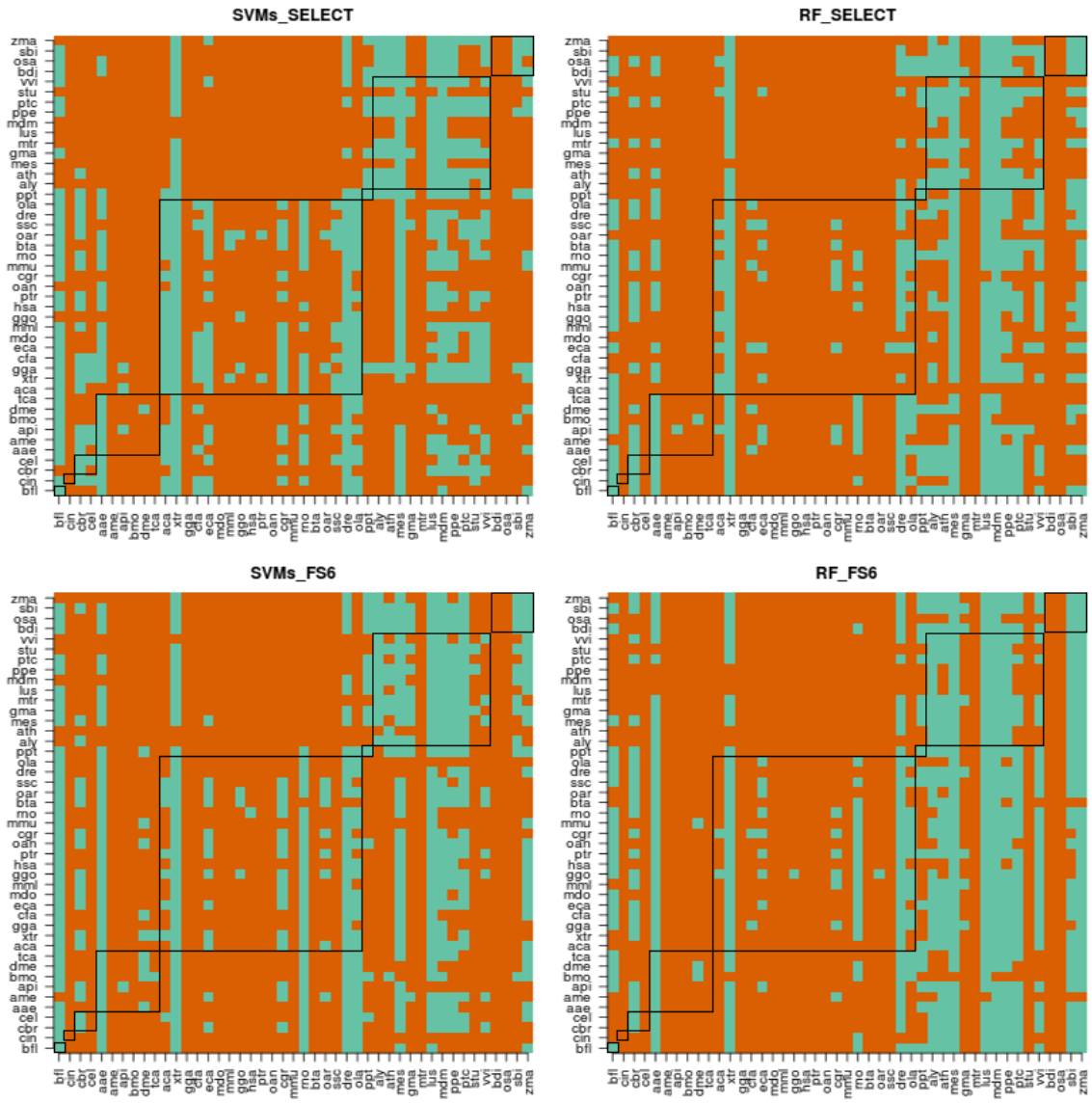
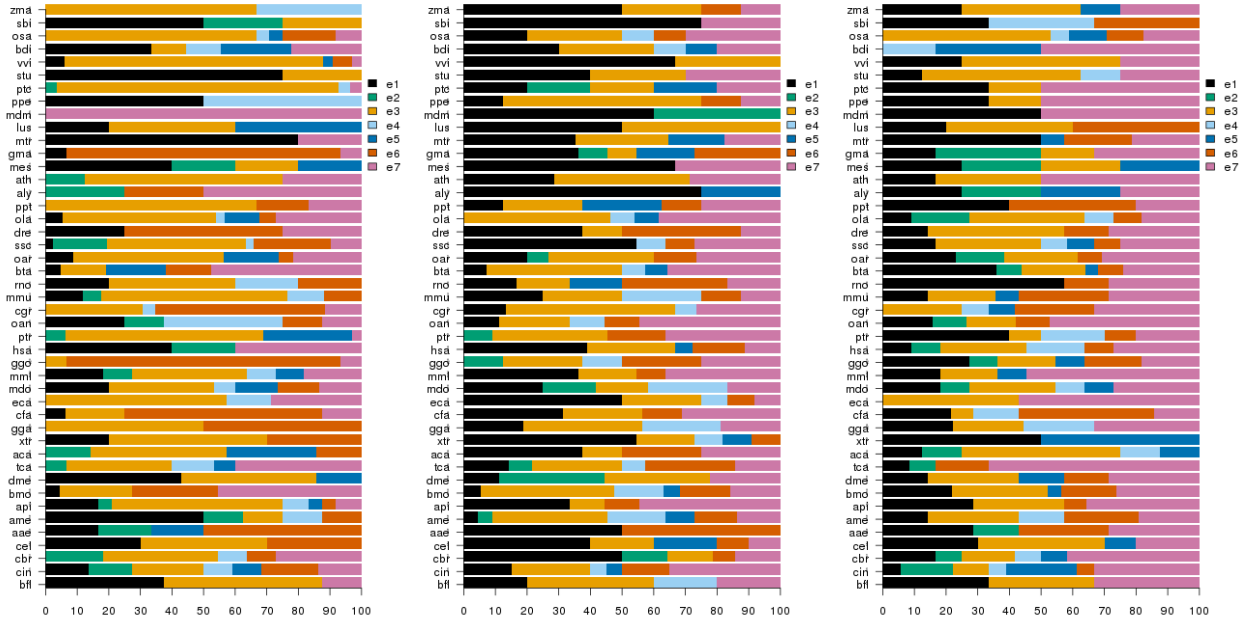



Figure 3: Accuracy cluster membership for cross-species pre-miRNAs classifiers. Green= $c_1$; red=other; $y$-axis=model species; $x$-axis=test species; black frames encloses species from the same subphylum/class.




(a) $FS_1$ (b) $FS_6$ (c) SELECT

Figure 4: Distribution of classification errors per species. Exclusive errors by SVMs ($e_1$), RF ($e_2$), J48 ($e_3$), SVMs and RF ($e_4$), SVMs and J48 ($e_5$), RF and J48 ($e_6$) and SVMs and RF and J48 ($e_7$).

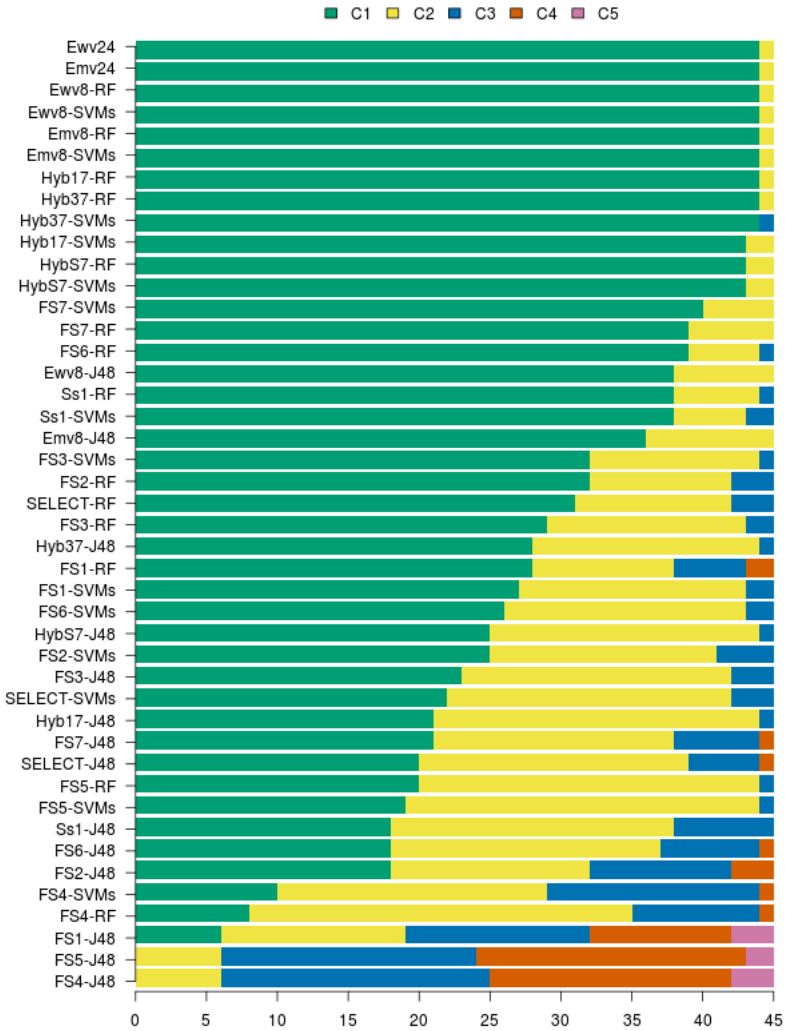
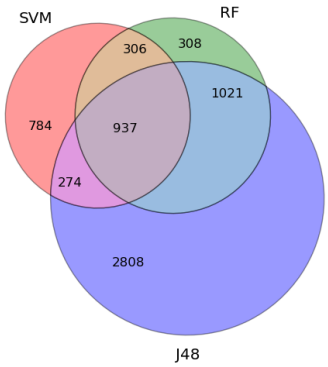
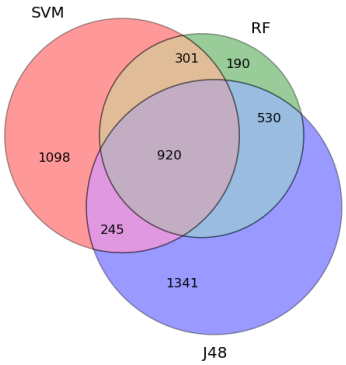



Figure 5: Distribution of the accuracies of 44 classifiers within the accuracy clusters.$\text{Mean}_{C_1} > ... > \text{Mean}_{C_5}$.

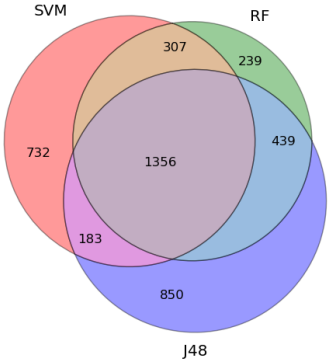



(a) $FS_1$




(b) $FS_6$


(c) $FS_7$




(d) SELECT

Figure 6: Venn diagram of the classification errors of the classification algorithms, by feature set. Results were obtained from the classification of 27000 = 45 (test species) ×10 (repetitions) ×60 (30+,30-).

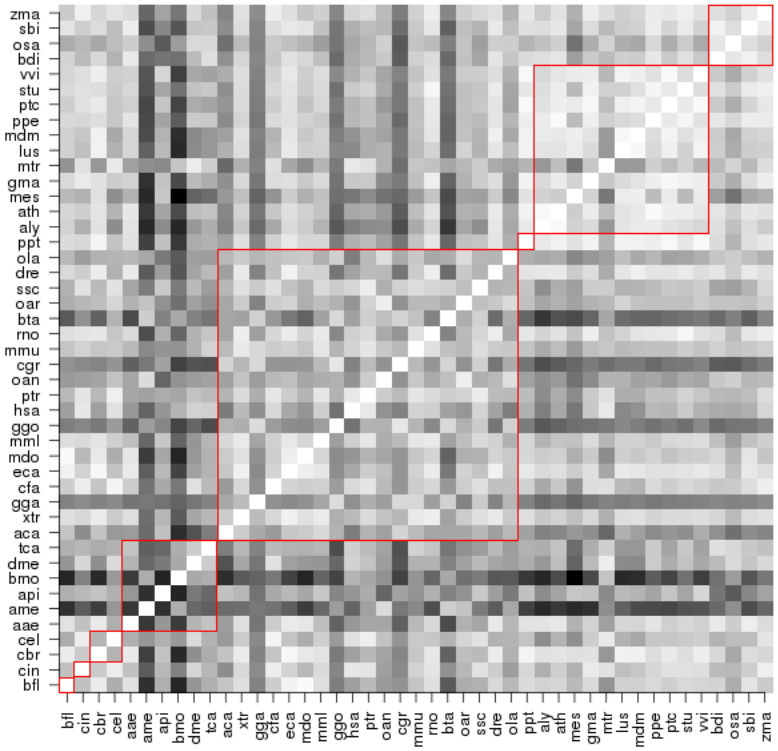



Figure 7: Pairwise Pearson correlation coefficient of $FRI$ throughout species.

Table 3: Centers of accuracy clusters from 24 classification models, per species. Range = Maximum - minimum.

| Acronym for species | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_5$ | Range |
|---|---|---|---|---|---|---|
| bfl | 94 | 83 | - | - | - | 15.0 |
| cin | 83 | 79 | 75 | 68 | - | 19.0 |
| cbr | 93 | 85 | 79 | - | - | 17.0 |
| cel | 92 | 87 | 81 | 75 | - | 20.0 |
| aae | 95 | 90 | 80 | - | - | 18.0 |
| ame | 85 | 78 | 72 | - | - | 20.0 |
| api | 92 | 88 | 82 | 73 | - | 22.0 |
| bmo | 84 | 79 | 71 | 57 | - | 31.0 |
| dme | 91 | 78 | - | - | - | 22.0 |
| tca | 89 | 82 | 76 | - | - | 18.0 |
| aca | 93 | 86 | 80 | - | - | 16.0 |
| xtr | 97 | 87 | 82 | - | - | 18.0 |
| gga | 95 | 90 | 85 | 76 | 68 | 27.0 |
| cfa | 91 | 83 | 75 | - | - | 22.0 |
| eca | 93 | 86 | 77 | - | - | 20.0 |
| mdo | 87 | 79 | 71 | - | - | 21.0 |
| mml | 89 | 82 | 75 | - | - | 17.0 |
| ggo | 89 | 77 | 66 | - | - | 27.0 |
| hsa | 88 | 77 | - | - | - | 16.0 |
| ptr | 89 | 82 | 73 | - | - | 23.0 |
| oan | 88 | 83 | 77 | 70 | - | 23.0 |
| cgr | 92 | 88 | 84 | 78 | - | 16.0 |
| mmu | 85 | 79 | 72 | - | - | 17.0 |
| rno | 93 | 88 | 81 | - | - | 17.0 |
| bta | 84 | 80 | 75 | 68 | - | 18.0 |
| oar | 91 | 86 | 77 | - | - | 18.0 |
| ssc | 90 | 85 | 79 | 64 | - | 29.0 |
| dre | 93 | 86 | 80 | - | - | 17.0 |
| ola | 92 | 88 | 80 | 68 | - | 26.0 |
| ppt | 93 | 84 | 76 | - | - | 20.0 |
| aly | 95 | 88 | 81 | - | - | 17.0 |
| ath | 94 | 83 | - | - | - | 15.0 |
| mes | 98 | 91 | 85 | - | - | 14.0 |
| gma | 91 | 86 | 79 | - | - | 18.0 |
| mtr | 86 | 82 | 72 | - | - | 21.0 |
| lus | 97 | 84 | - | - | - | 18.0 |
| mdm | 98 | 85 | - | - | - | 15.0 |
| ppe | 95 | 87 | 80 | - | - | 18.0 |
| ptc | 94 | 83 | - | - | - | 16.0 |
| stu | 93 | 87 | 82 | - | - | 16.0 |
| vvi | 93 | 86 | 78 | - | - | 20.0 |
| bdi | 91 | 87 | 75 | - | - | 22.0 |
| osa | 87 | 77 | - | - | - | 16.0 |
| sbi | 96 | 89 | 81 | - | - | 20.0 |
| zma | 96 | 82 | - | - | - | 17.0 |

Table 4: Centers of accuracy clusters obtained from classification models induced with examples from different species, per combination of feature set and learning algorithm. Range = Maximum - minimum.

| FEATURE SET | ALGORITHM | $c_1$ | $c_2$ | $c_3$ | $c_4$ | $c_5$ | $c_6$ | Range |
|---|---|---|---|---|---|---|---|---|
| FS1 | SVM | 95 | 88 | 78 | - | - | - | 21 |
| FS2 | | 96 | 92 | 87 | 80 | - | - | 20 |
| FS3 | | 95 | 90 | 85 | - | - | - | 15 |
| FS4 | | 92 | 86 | 81 | 77 | - | - | 22 |
| FS5 | | 94 | 90 | 86 | 80 | - | - | 20 |
| FS6 | | 93 | 88 | 83 | - | - | - | 17 |
| FS7 | | 95 | 88 | - | - | - | - | 16 |
| SELECT | | 96 | 92 | 86 | 80 | - | - | 20 |
| FS1 | RF | 97 | 92 | 87 | 82 | 72 | - | 30 |
| FS2 | | 97 | 93 | 89 | 83 | - | - | 20 |
| FS3 | | 95 | 88 | 84 | - | - | - | 18 |
| FS4 | | 91 | 87 | 84 | 79 | - | - | 18 |
| FS5 | | 92 | 85 | 77 | - | - | - | 19 |
| FS6 | | 95 | 88 | - | - | - | - | 16 |
| FS7 | | 96 | 89 | - | - | - | - | 14 |
| SELECT | | 96 | 92 | 86 | 78 | - | - | 21 |
| FS1 | J48 | 98 | 91 | 85 | 75 | 67 | 57 | 41 |
| FS2 | | 96 | 90 | 84 | 77 | - | - | 24 |
| FS3 | | 97 | 92 | 87 | 81 | - | - | 21 |
| FS4 | | 84 | 79 | 75 | 69 | - | - | 21 |
| FS5 | | 83 | 78 | 75 | 71 | - | - | 17 |
| FS6 | | 97 | 93 | 89 | 83 | 78 | 72 | 27 |
| FS7 | | 96 | 91 | 87 | 81 | - | - | 21 |
| SELECT | | 97 | 92 | 86 | 80 | 74 | - | 26 |

Table 5: Relative feature importance ($FRI$) distributions. Omitting those of $RFI \leq 0.1$ for all species.

| Nº | FEATURE | $RFI$ intervals | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | ≤0.3 | (0.3,0.4 | (0.4,0.5] | (0.5,0.6] | (0.6,0.7] | (0.7,0.8] | (0.8,0.9] | ≥1.0 |
| 1 | $p$ | 4 | 0 | 4 | 2 | 9 | 13 | 16 | 51 |
| 2 | $MFEI_1$ | 4 | 0 | 4 | 0 | 7 | 4 | 22 | 58 |
| 3 | $dP$ | 33 | 29 | 20 | 11 | 7 | 0 | 0 | 0 |
| 4 | $ZG$ | 36 | 9 | 9 | 16 | 7 | 2 | 4 | 18 |
| 5 | $orf$ | 36 | 31 | 4 | 11 | 9 | 2 | 4 | 2 |
| 6 | $dG$ | 47 | 20 | 13 | 11 | 7 | 2 | 0 | 0 |
| 7 | $ZP$ | 53 | 13 | 9 | 2 | 7 | 2 | 0 | 13 |
| 8 | $Avg_$Bp_Stem | 53 | 33 | 9 | 4 | 0 | 0 | 0 | 0 |
| 9 | $EAFE$ | 64 | 20 | 4 | 7 | 2 | 2 | 0 | 0 |
| 10 | $MFEI_3$ | 67 | 13 | 9 | 4 | 2 | 0 | 2 | 2 |
| 11 | $MFEI_4$ | 80 | 9 | 7 | 0 | 0 | 2 | 2 | 0 |
| 12 | $\|A-U\|/L$ | 80 | 16 | 4 | 0 | 0 | 0 | 0 | 0 |
| 13 | $ZQ$ | 82 | 2 | 4 | 2 | 0 | 2 | 2 | 4 |
| 14 | $ZD$ | 84 | 0 | 2 | 4 | 2 | 2 | 2 | 2 |
| 15 | $A_{(((}$ | 89 | 9 | 0 | 2 | 0 | 0 | 0 | 0 |
| 16 | $Tm$ | 93 | 7 | 0 | 0 | 0 | 0 | 0 | 0 |
| 17 | $dQ$ | 93 | 7 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18 | $Diversity$ | 98 | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| 19 | $U_{(((}$ | 98 | 0 | 2 | 0 | 0 | 0 | 0 | 0 |
| 20 | $MFE$ | 98 | 0 | 2 | 0 | 0 | 0 | 0 | 0 |
| 21 | $\%(A-U)/stems$ | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 22 | $dD$ | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 23 | $Diff$ | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 24 | $Tm/L$ | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 25 | $\%UU$ | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 26 | $dm$ | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 27 | $\%G+C$ | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 28 | $MFEI_4$ | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 29 | $A_{...}$ | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 30 | $\%GA$ | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 31 | $C_{...}$ | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 32 | $dH/L$ | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 33 | $G_{...}$ | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 34 | $\%UA$ | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Table 6: Phylum/division, subphylum/class, species, acronyms, number of positive examples available at miRBase 20, mean and standard deviation of the length distributions. NR=Non-Redundant.

| Phylum/Division | Subphylum/Class | Species genus | Acronym | #pre-miRNA | | Length |
|---|---|---|---|---|---|---|
| | | | | All | NR | (Mean ± SD) |
| Chordate | Cephalochordata | Branchiostoma floridae | bfl | 156 | 143 | 87 ± 13 |
| | Urochordata | Ciona intestinalis | cin | 346 | 331 | 63 ± 16 |
| | Nematoda | Caenorhabditis briggsae | cbr | 177 | 148 | 92 ± 19 |
| | | Caenorhabditis elegans | cel | 233 | 214 | 89 ± 17 |
| | Hexapoda | Aedes egypti | aae | 101 | 90 | 94 ± 21 |
| | | Apis mellifera | ame | 218 | 215 | 100 ± 20 |
| | | Acyrthosiphon pisum | api | 117 | 101 | 66 ± 9 |
| | | Bombyx mori | bmo | 489 | 432 | 100 ± 22 |
| | | Drosophila melanogaster | dme | 238 | 236 | 95 ± 23 |
| | | Tribolium castaneum | tca | 220 | 210 | 95 ± 22 |
| | Vertebrate | Anolis carolinensis | aca | 282 | 272 | 89 ± 9 |
| | | Xenopus tropicalis | xtr | 189 | 163 | 83 ± 11 |
| | | Gallus gallus | gga | 734 | 695 | 92 ± 17 |
| | | Canis familiaris | cfa | 324 | 280 | 69 ± 14 |
| | | Equus caballus | eca | 341 | 298 | 78 ± 15 |
| | | Monodelphis domestica | mdo | 460 | 370 | 67 ± 12 |
| | | Macaca mulatta | mml | 615 | 524 | 86 ± 17 |
| | | Gorilla gorilla | ggo | 332 | 313 | 105 ± 12 |
| | | Homo sapiens | hsa | 1,872 | 1,640 | 82 ± 17 |
| | | Pan troglodytes | ptr | 659 | 542 | 90 ± 17 |
| | | Ornithorhynchus anatinus | oan | 396 | 327 | 100 ± 24 |
| | | Cricetulus griseus | cgr | 200 | 199 | 82 ± 12 |
| | | Mus musculus | mmu | 1,186 | 1,078 | 83 ± 19 |
| | | Rattus norvegicus | rno | 449 | 428 | 92 ± 17 |
| | | Bos taurus | bta | 798 | 710 | 80 ± 13 |
| | | Ovis aries | oar | 105 | 96 | 97 ± 18 |
| | | Sus scrofa | ssc | 280 | 247 | 81 ± 10 |
| | | Danio rerio | dre | 346 | 240 | 93 ± 18 |
| | | Oryzias latipes | ola | 168 | 146 | 95 ± 9 |
| Bryophyta | Musci | Physcomitrella patens | ppt | 229 | 204 | 161 ± 56 |
| Angiospermae | Eudicotyledons | Arabidopsis lyrata | aly | 298 | 177 | 183 ± 100 |
| | | Arabidopsis thaliana | ath | 298 | 257 | 183 ± 103 |
| | | Manihot esculenta | mes | 153 | 109 | 117 ± 38 |
| | | Glycine max | gma | 505 | 361 | 131 ± 47 |
| | | Medicago truncatula | mtr | 672 | 373 | 165 ± 91 |
| | | Linum usitatissimum | lus | 124 | 100 | 144 ± 34 |
| | | Malus domestica | mdm | 206 | 90 | 130 ± 66 |
| | | Prunus persica | ppe | 180 | 147 | 136 ± 51 |
| | | Populus trichocarpa | ptc | 352 | 246 | 128 ± 46 |
| | | Solanum tuberosum | stu | 224 | 163 | 95 ± 43 |
| | | Vitis vinifera | vvi | 163 | 131 | 127 ± 56 |
| | | Brachypodium distachyon | bdi | 258 | 228 | 178 ± 101 |
| | Monocotyledons | Oryza sativa | osa | 592 | 482 | 153 ± 77 |
| | | Sorghum bicolor | sbi | 205 | 174 | 142 ± 54 |
| | | Zea mays | zma | 172 | 133 | 132 ± 45 |