Lopes *et al. BMC Bioinformatics* (2016) 17:224
DOI 10.1186/s12859-016-1036-3

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Automatic learning of pre-miRNAs from different species

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

Abstract

Background: Discovery of microRNAs (miRNAs) relies on predictive models for characteristic features from miRNA precursors (pre-miRNAs). The short length of miRNA genes and the lack of pronounced sequence features complicate this task. 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 might be lower. While these biases are intrinsic to the species, their characterization can lead to computational approaches capable of diminishing their negative effect on the accuracy of pre-miRNAs predictive models. 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 computational experiments show that the separability of pre-miRNAs and pseudo pre-miRNAs instances is species-dependent and no feature set performs well for all species, even within the same subphylum/class. Mitigating this species dependency, 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 individual classifiers that rely on energy stability parameters, which are 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.

The material to reproduce the results from this paper can be downloaded from http://dx.doi.org/10.5281/zenodo.49754.

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 post-transcriptional 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 the double strand RNA miRNA:miRNA* and a terminal loop. The miRNA* strand is the reverse complement of the functional miRNA, which usually degrades after being unwound by the action of specific enzymes. 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].

*Correspondence: ivani.negrao@embrapa.br
[1]Empresa Brasileira de Pesquisa Agropecuária, Embrapa Soja, Caixa Postal 231, Londrina-PR, CEP 86001-970, Brasil
Full list of author information is available at the end of the article

BioMed Central

© 2016 Lopes et al. **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

These methods, also called deep sequencing of the 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 is reflected in the wide range of sensitivities estimated for eight deep sequencing miRNA prediction tools, when they were applied to data sets from *H. sapiens*, *G. Gallus* and *C. elegans* [7]. 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 0 % (*H. sapiens*) and 25 % (*C. elegans*). The species bias is also present in the analysis performed with miRDeep2 [8], a newer version of miRDeep [6], which incorporated additional 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 varies depending on the species 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 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 approximately 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 simultaneously, where the distribution of the reads in the putative hairpin and hairpin features are considered (as in miRDeep2 [8]). Variants of those approaches have also been proposed in the literature. NoraDesk [10] was the first method to incorporate structural energy based features and read coverage information to increase the detection accuracy of small ncRNAs, including miRNAs, whereas miReader [11] relies only on hybridization patterns of the reads to detect miRNAs.

The hairpin analysis has been performed mostly through machine learning based predictive models. To obtain these models, a feature set (feature vector) 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 indicated that the feature set, instead of the learning algorithm, had the major effect in the classification accuracy of the induced models [12]. However, the relevance of those features for the correct classification of pre-miRNAs from other species remained an open question.

Since miRNA systems in plants and animals differ substantially [4], computational tools for plant and animal miRNAs discovery have been developed separately (e.g. [9, 13]). 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 robust to species-specific differences, a first step is to determine if a unique feature set can capture the diversity of pre-miRNAs throughout species. Moreover, it is important to establish 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 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) if they are to be computed for millions of hairpins. These issues were addressed in this study, considering eight feature sets investigated in [12], 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 discussions section). To interpret this dependency, we analyzed how relevant the features extracted from instances of one species are for the classification of instances of other species (in the following subsections). This analysis 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 using computationally inexpensive feature sets performed well even if the subphylum of the training species disagrees. The ensemble approach has the potential to extend the applicability of pre-miRNA predictive models to a broader number of species, while keeping the computational cost close to that of single classifiers.

Methods

Experimental design

The analysis carried out in this study was 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) Eqs. 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. Once training and test sequence sets had been randomly sampled, all features were extracted. Therefore, the data sets of feature vectors diverged only by the feature composition, which can be geometrically seen as different subspaces of the unknown space of the pre-miRNAs features. 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. Moreover, 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 × S

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

$$A_{ilk} = \mu + M_l + S_i + MS_{li} + R_k + e_{ilk}, \quad (1)$$

such that:
$l = 1, \ldots, 24$ indexes the classifiers,
$i = 1, \ldots, 45$ indexes the species,
$k = 1, \ldots, 10$ indexes the repetition,
A_{ilk} = accuracy of the classifier l, obtained with the training species i in the repetition k,
μ = 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 R_k=effect of repetition k, blocking factor;
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, $l = 1, \ldots, 24$, and varied the training and test species. The accuracies were analyzed according to Eq. 2:

$$A_{lijk} = \mu + M_{li} + T_j + MT_{lij} + R_k + e_{lijk}, \quad (2)$$

such that:
l indexes one out the 24 classifiers,
$i, j = 1, \ldots, 45$ indexes training and test species,
$k = 1, \ldots, 10$ 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,
μ = 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 i and test species j, and R_k=effect of repetition k, blocking factor;
e_{lijk} = random error, or part of A_{lijk} that could not be assigned to the species i, the test species j in the repetition k; $e \sim N(0, \sigma^2)$.

Clustering algorithm

The Eqs. 1 and 2 are particularly useful to estimate the variance of random errors (σ^2). Once this variance is known, we can decide how typical the variances estimated from the controlled factors (e.g. M, S and MS) are, compared to σ^2, using the p-value obtained from the F-test. In this work, significant p-values were lower or equal to 0.05 ($p \leq 0.05$). Since significant p-values of F-test on a factor only supports the inference that at least two levels of that factor had different average effects, we applied a clustering algorithm due to Scott and Knott [14] to

identify the levels of each factor in Eqs. 1 and 2 that led to non-significantly different accuracies using the R package ScottKnott [15].

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 [16]. 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 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, as detailed in the Additional file 2. 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-miRNAs. 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;
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 three MFE rules were applied to accommodated the correlation between MFE and L that occurs in pre-miRNAs.

Pre-processing

Genes in a miRNA family can have sequence identity of 65 % or higher [17]. 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 [18] 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 similarity 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. As detailed in Additional file 2, 15 or less sequences were removed from 35 out of 45 negative sequence sets. The relatively lower number of redundant pseudo hairpins in those sets, compared to pre-miRNAs sequence sets, is due to the random choice of the starting position of the pseudo hairpin excision. However, at least 35 redundant pseudo hairpins were removed from the other 10 sequence sets.

Feature sets

The eight features sets primarily studied in this investigation were extensively evaluated on human sets by Lopes et al. [12]. Here, these feature sets are referred by the same notation (FS_i, $i \in \{1, .., 7\}$ and SELECT). Table 1 presents the features that compose each feature set, along with references of computational pipelines where they have been used. Although detailed descriptions of these features can be found in the references cited in Table 1 or elsewhere, we provide next a short description of these features regarding four major categories: sequence composition features, structure based features, sequence-structure based features, thermodynamic features and probabilistic properties. Their representation in Table 1 are within parentheses in the text below.

Sequence composition features

This category includes the dinucleotides contents ($\%XY, X, Y \in \{A, C, U, G\}$) and the $G + C$ or $A + G$ contents ($\%G + C$ or $\%A + G$), the maximal length of the amino acid string without stop codons (orf) and the percentage of low complexity regions (dm) in the sequence [19]. The first two groups of features are more intuitive, whereas the last two features may help to distinguish protein coding sequences from ncRNAs, since $\%LCRs$ are defined as short amino acid motifs or regions that contain repeats of single amino acids [20].

Structure based features

The predicted secondary structure is the intra-molecular accommodation demanding the Minimum Free Energy of

Table 1 Feature set composition, dimension, literature reference

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 (*dm*)							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	[21]	[21]	[33]	[23]	[34]	[35]	[19]	[12]

folding (*MFE* or *G*). Some *MFE* variants have been proposed to correct the bias towards sequence length (*dG*), %*G*+*C* ($MFEI_1$) and structural complexity ($MFEI_2$, $MFEI_3$ and $MFEI_4$). In vivo, an RNA molecule commonly exists in an assembly of structures. The distribution of these structures can be modeled by a Boltzmann distribution of free energy and the probabilities of these structures are used to compute the Normalized Ensemble Free Energy (*NEFE*), the normalized difference ($Diff = (MFE - EFE)/L$) and the frequency of the *MFE* structure (*Freq*) [21]. Since pre-miRNAs are typically energetically more stable [22], the *MFE* and its variants are important feature for pre-miRNA prediction.

The minimum number of base-pairings (bp) in the secondary structure of pre-miRNAs is approximately 18 bp [23]. The normalized base-pairing propensity (*dP*) and the average base pairs per stem (*Avg_Bp_Stem*) indicates the occurrence of this pattern in a stem-loop structure, whereas the base-pair distance and its normalized version ($Diversity$ and $dD = Diversity/L$) inform the structural diversity.

More complex patterns of the secondary structure are captured by the topological descriptor (*dF*), the normalized Shannon entropy (*dQ*) and the cumulative size of internal loops found in the secondary structure (*loops*). *dF* is the second eigenvalue of the Laplacian matrix of the graph representation of the secondary structure where, bulges, loops, and other related measures are the vertices and the stems are the edges. The parameter *dQ* characterizes the base-pairing probability distribution per base in a sequence, represented as a chaotic dynamical system [24]. Since the local dominance of a single structure within the Boltzmann distribution of alternative secondary structures is strongly correlated with the reliability of the *MFE* structure, *dQ* is a measure of well-definedness for the structure [25]. In addition, well-defined structures have lower Shannon entropy as compared to structures with many alternative competing base pairs. The last parameter, *loops*, has been proposed as a distinguishing feature for pre-miRNA prediction [19], since pre-miRNAs usually have smaller internal loops.

Sequence-structure based features

Features in this category include the sequence nucleotide information and its state in the predicted secondary structure. The abundance of Watson-Crick base-pairings A-U, C-G and G-U in the secondary structure was considered according to two normalization criteria: the sequence length *L* ($|A-U|/L$, $|C-G|/L$, $|G-U|/L$) and the number of stems on the secondary structure ($|A-U|/n\_stems$, $|C-G|/n\_stems$, $|G-U|/n\_stems$). Another group of features accounts for the occurrence of certain patterns in the sequence and structure level. The triplets $s_l s_x s_r$ represent the normalized frequency of three contiguous states ($\{\text{paired} = `(', \text{unpaired} = `.'\}$) in the secondary structure, where s_x is the state of a fixed middle nucleotide ($\{A, C, G, U\}$) and s_l and s_r are the states of x's left and right neighbors. An extension of the triplets are the sequence-structure motifs. The relative occurrence of a motif is computed from the string obtained by padding the original sequence with the corresponding state of each nucleotide in the secondary structure, distinguishing left and right pairings.

Thermodynamic features

Thermodynamic features are structure entropy (*dS* and dS/L), structure enthalpy (*dH* and dH/L) and melting temperature (T_m). The latter is estimated assuming that the sequence either folds or do not fold. Thus, the estimated T_m is the temperature where the fold-back or hairpin-like structure disrupts. T_m is related with the enthalpy and the entropy by the equation $T_m = \Delta H/\Delta S$, where Δ represents variation.

Probabilistic properties

Probabilistic properties refer to parameters that measure the stability (or variation) of certain features when computed from a sequence and from its randomized (shuffled) versions [22]. Since real pre-miRNAs are energetically more stable than pseudo-hairpins, the differences between the *MFE* of a real pre-miRNAs and the mean *MFE* of its shuffled sequences (MFE_{shuf}) are expected to be neglectable. Two different formulas to capture the energy stability have been proposed. The first, *zG* [24], is the difference between *MFE* and MFE_{shuf}, in unities of standard deviation (SD_{shuf}). The second, *p* (randfold) [22, 26], is the relative frequency by which MFE_{shuf} was lower than *MFE*. Similarly to the computation of *zG*, the *z*-variants of *dP*, *dQ*, *dD* and *dF* were also assessed and represented as *zP*, *zQ*, *zD* and *zF* [24].

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 frequently used for pre-miRNA classification and J48 was chosen because of its simplicity and interpretability.

J48 implements the well known C4.5 algorithm [27]. 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 [28], an R interface of Weka [29], 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 [30]. This interface implements the C-SVM algorithm using the RBF kernel. The kernel parameters γ and C were tuned by 5-fold cross validation (CV) over the grid $(C;\gamma) = (\ 2^{-5}, 2^{-3}, \ldots, 2^{15}; 2^{-15}, 2^{-13}, \ldots, 2^{3})$. 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 [31] in an *in house* R script.

Ensembles and other feature sets

In addition to 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 [19]. 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 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 predictions from a set of single 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 [12]. As shown in this table, the final prediction of the ensembles were defined by majority vote (ensemble Emv) and by weighted vote (ensemble Ewv). In the first approach, the class predicted by the majority of the classifiers is the ensemble class prediction. In the weighted approach, the vote from each classifier was weighted by its predictive accuracy in the training sets. 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 Eq. 1 was highly significant ($p < 0.001$), the effect of the simple factor M was studied within fixed levels of S ($\mathrm{M/S}_j$, $j = 1, \ldots, 45$), and vice-verse ($\mathrm{S/M}_l$, $l = 1, \ldots, 24$). The analysis of $\mathrm{M/S}_j$, $j = 1, \ldots, 45$, is summarized in Fig. 1 and Table 3. The green bars in Fig. 1 indicate the pre-miRNA classifiers whose accuracy is within the cluster of maximal accuracies C_1. As indicated in Fig. 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 [12], 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 clusters 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 $\mathrm{S/M}_l$, $l = 1, \ldots, 24$. Similar to what was observed in the analyzes of $\mathrm{M/S}_j$, $j = 1, \ldots, 45$, 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, $l = 1, \ldots, 24$ 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.

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, $i = 1, \ldots, 12$ and $j = 1, 2, 3$, and w_{ij} is the cross-validation accuracy 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	$\sum \hat{M}_{i1}, i = 5, \ldots, 12$	$\sum \hat{M}_{i2}, i = 5, \ldots, 12$	$\sum \hat{M}_{i3}, i = 5, \ldots, 12$
Ewv8	$\sum w_{i1}\hat{M}_{i1}, i = 5, \ldots, 12$	$\sum w_{i2}\hat{M}_{i2}, i = 5, \ldots, 12$	$\sum w_{i3}\hat{M}_{i3}, i = 5, \ldots, 12$
Emv24		$\sum \hat{M}_{ij}, i = 5, \ldots, 12$ and $j = 1, 2, 3$	
Ewv24		$\sum w_{ij}\hat{M}_{ij}, i = 5, \ldots, 12$ and $j = 1, 2, 3$	

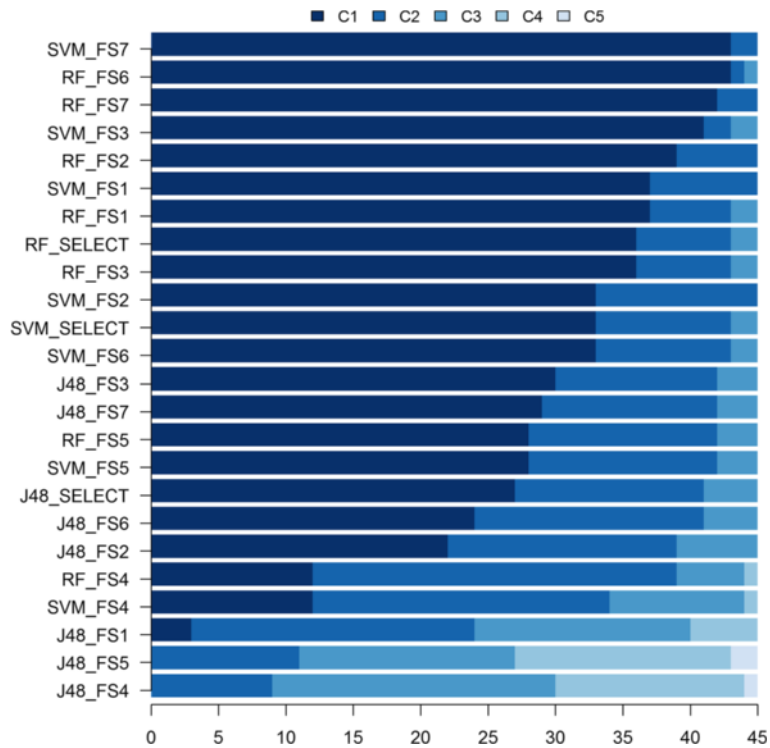

Fig. 1 Frequencies of species for who each classification model achieved accuracies in the clusters C_1-C_5. $Mean_{C_1} \geq .. \geq Mean_{C_5}$

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

In the next subsection, we discuss how representative the instances from the 45 species considered in this work are 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: $M_l \times T$

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 Eq. 2. Since the F-test on the interaction M_lT was significant ($p \leq 0.05$), the factor M_l was analyzed within each level of the factor T ($M_l/T_j, j = 1, \ldots, 45$), and viseverse ($T/M_{li}, i = 1, \ldots, 45$). The results of the analyzes of $M_l/T_j, j = 1, \ldots, 45$ 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 $T/M_{li}, i = 1, \ldots, 45$, we discussed the learning complexity of pre-miRNAs from the 45 species.

Choosing the training species - M_l/T

By clustering the average accuracies $\overline{A}_{lij\cdot}$, within j, $i, j = 1, \ldots, 45$, 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 classifier was more likely to achieve predictive accuracies within c_1 when the species i and j were from the same subphylum/class. In particular, all means $\overline{A}_{lij\cdot}$ were in c_1 when $i = j$

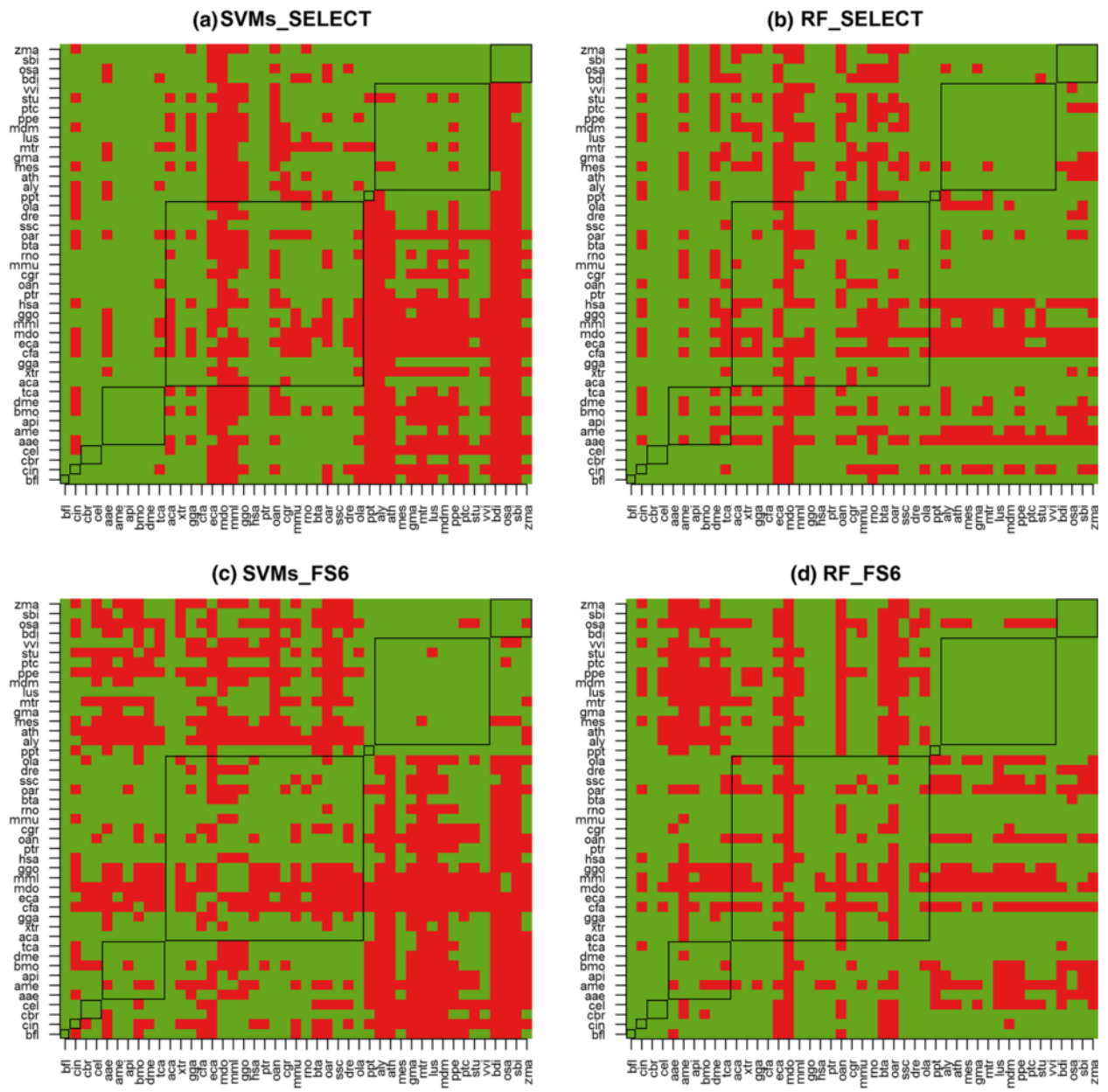

Fig. 2 Accuracy cluster membership (columns) for cross-species pre-miRNAs classifiers. *Green*= c_1; *red*=other; *y*-axis=model species; *x*-axis=test species; black frames encloses species from the same subphylum/class. Figures (**a**), (**b**), (**c**) and (**d**) show interactions between the learning algorithms SVMs and RF and the feature sets SELECT and FS_6

(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 per column. Among them, instances from mdo 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 M. domestica (mdo) 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 mdo 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 Fig. 2 shows. In this figure, SVMs were more sensitive to the phylogenetic proximity of training and test species. While this pattern can be seen as an SVMs drawback for this problem, the phylogeny of 26 metazoan species in Additional file 3: Figure S1, shows that the distances between these species vary widely. As it can be observed in this figure, lighter areas (higher accuracies) were more frequent when training species (dendogram) and test species (rows of the matrix) were within the two groups defined by the last level of the hierarchy; one group has Hexapoda and Nematoda species and the other has Urochordata (cin) and Vertebrata species. However, the figure does not suggest a strong correlation between phylogenetic proximity and predictive accuracies when SVMs were used.

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 Fig. 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, 100] by the formula $RFI = 100 * (FI - FI_{min})/(FI_{max} - FI_{min})$. The maximum ($FI_{max}$) and minimum ($FI_{min}$) FI values were obtained from the set of FI of the features used in the training step. We estimated the RFI values for each of the 85 unique features considered in this work feature, eliminating the 1,300 sequence structure motifs, when they were simultaneously fed to the RF algorithm to induce pre-miRNA classifiers for each of the 45 species. The pairwise Pearson correlation coefficients between species and the *RFIs* for each species are discussed next.

RFI Pearson's correlation coefficients throughout species

Figure 4 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

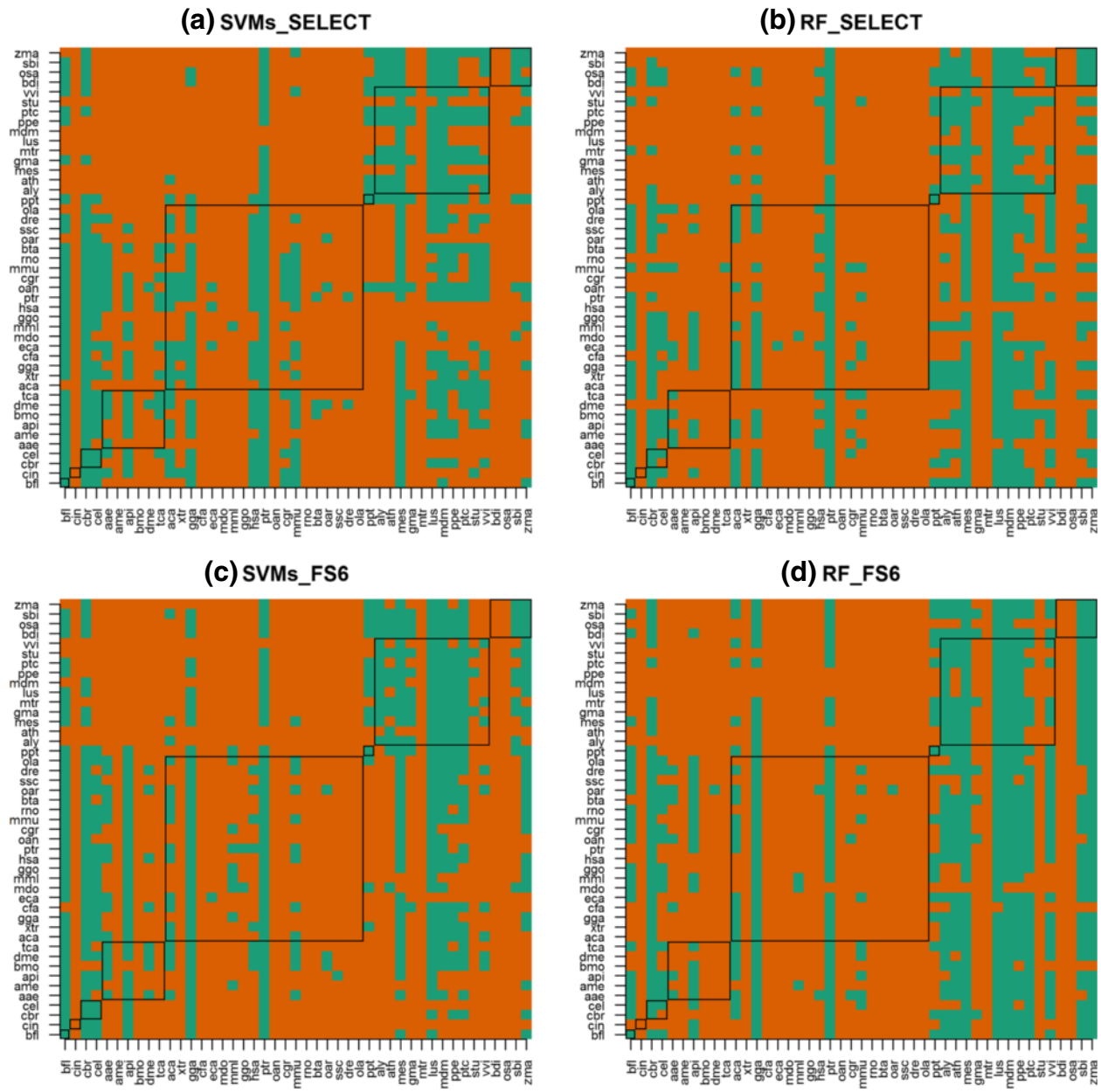

Fig. 3 Accuracy cluster membership (*rows*) for cross-species pre-miRNAs classifiers. *Green*= c_1; *red*=other; *y*-axis=model species; *x*-axis=test species; black frames encloses species from the same subphylum/class. Figures (**a**), (**b**), (**c**) and (**d**) show interactions between the learning algorithms SVMs and RF and the feature sets SELECT and FS_6

dark pixels associated with Hexapoda species. This exceptions and the features with the highest *RFI* are presented next.

Inferring feature relevance from RFI

The dominance of a small number of features in the classification of human pre-miRNAs was pointed out in [12]. In that work, which used larger training sets (875+, 875-), the features that obtained the highest *FI* values were energy related ($MFEI_1$, *zG*, *p*, *NFE*) or structural pairing patterns (*dP*, *zP*). As Fig. 5 shows, similar results were obtained in this work. The features $MFEI_1$ and *p* were the main causes of misclassification, when permuted during the RF trainings, for most species, particularly when extracted from instances of plant species. However, those features were of lower relevance ($RFI \leq 20$) for the species *B. mori* (bmo), *A. mellifera* (ame) and *B. taurus* (bta). For *H. sapiens* (hsa), *D. melanogaster* (dme) and *C. elegans*

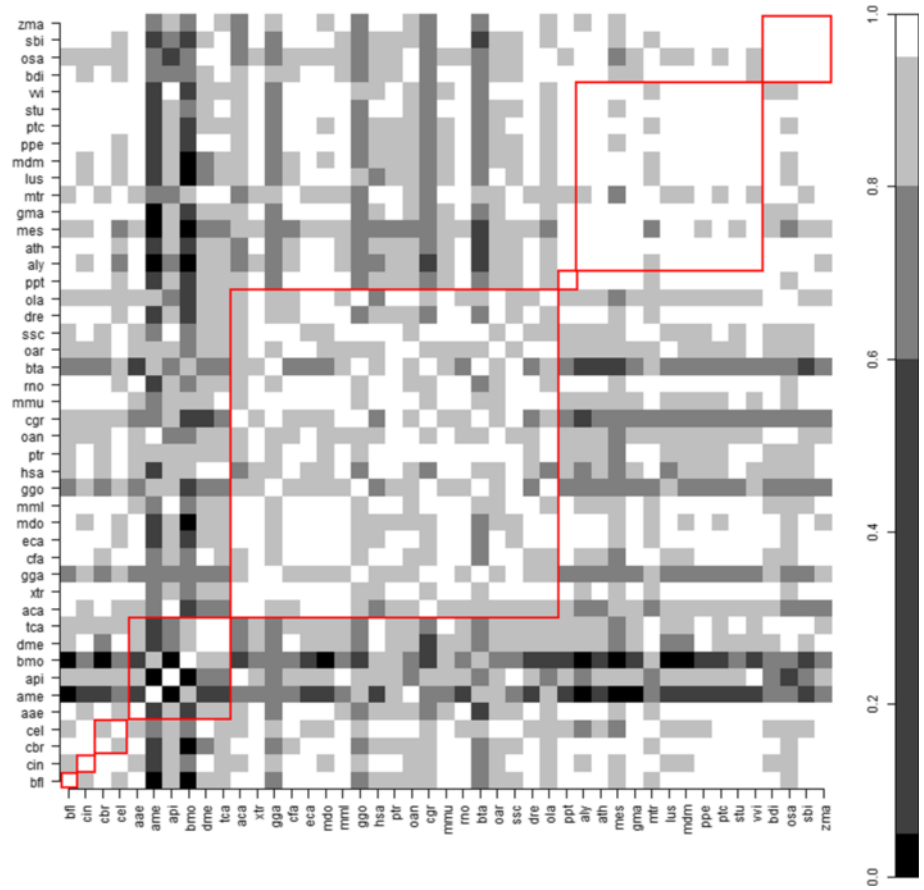

Fig. 4 Pairwise Pearson correlation coefficient of *RFI* throughout species. Correlations breaks: 0, 0.05, 0.6, 0.8, 0.95 and 1

(cel) instances, the feature *zG* obtained the highest *RFI*, whereas the feature of highest *RFI* for ame was *zP*.

The most interesting characteristic in Fig. 5 is the variation of *RFI* throughout species, particularly for Vertebrate and Hexapoda species. This finding suggests that no feature is prevalent in the pre-miRNAs of all species. If on one hand it can be seen as a drawback in the development of pre-miRNAs for multiple species, on the other hand, it suggests that it is possible to combine these features to obtain tools for miRNAs predictions less sensitive to characteristics inherent to different species.

Learning biases

The small amount of highly relevant features (Fig. 5) 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 (Fig. 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 $e_1, \ldots, e_7$ 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 Fig. 6, for FS_1, FS_6 and SELECT.

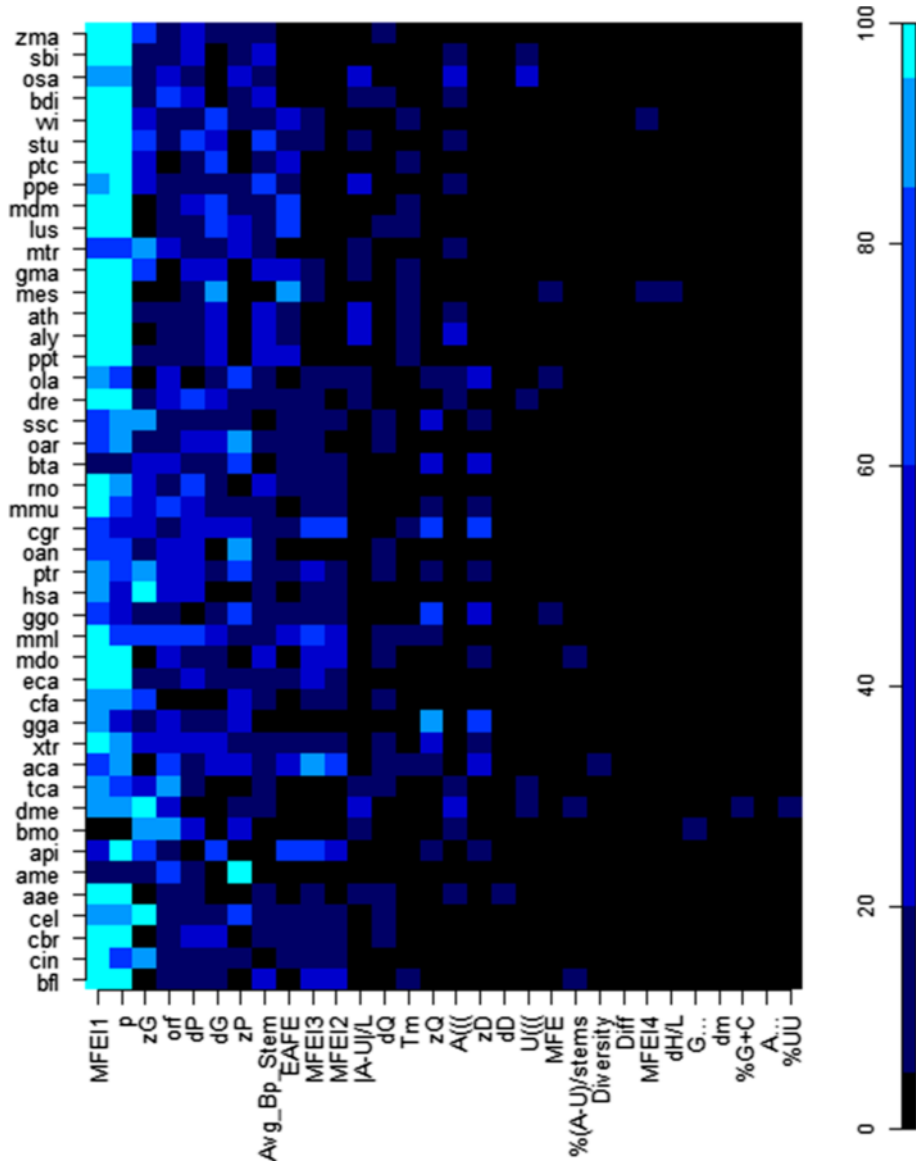

Fig. 5 Thirty most relevant features for pre-miRNAs classification, according to *RFI* values, per species. *RFI* breaks: 0, 5, 20, 60, 85, 95 and 100

As can be observed in Fig. 6, the error distributions were strongly dependent on the species, which shows in another way the classification biases associated with species sequence data. For example, Fig. 6(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 $e_7 > 50\%$ (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 $e_1, \ldots, e_7$, were summed up for the 45 species and represented in Venn diagrams. Figure 7 shows the cases FS_1, FS_6, FS_7 and SELECT. The interaction between learning algorithm and feature set, is evidenced by the large variation in the numbers of misclassified instances in the $e_1,\ldots,e_7$ regions. For example, classification models induced by J48 tended to achieve higher exclusive error rates (e_3) in higher dimensional feature

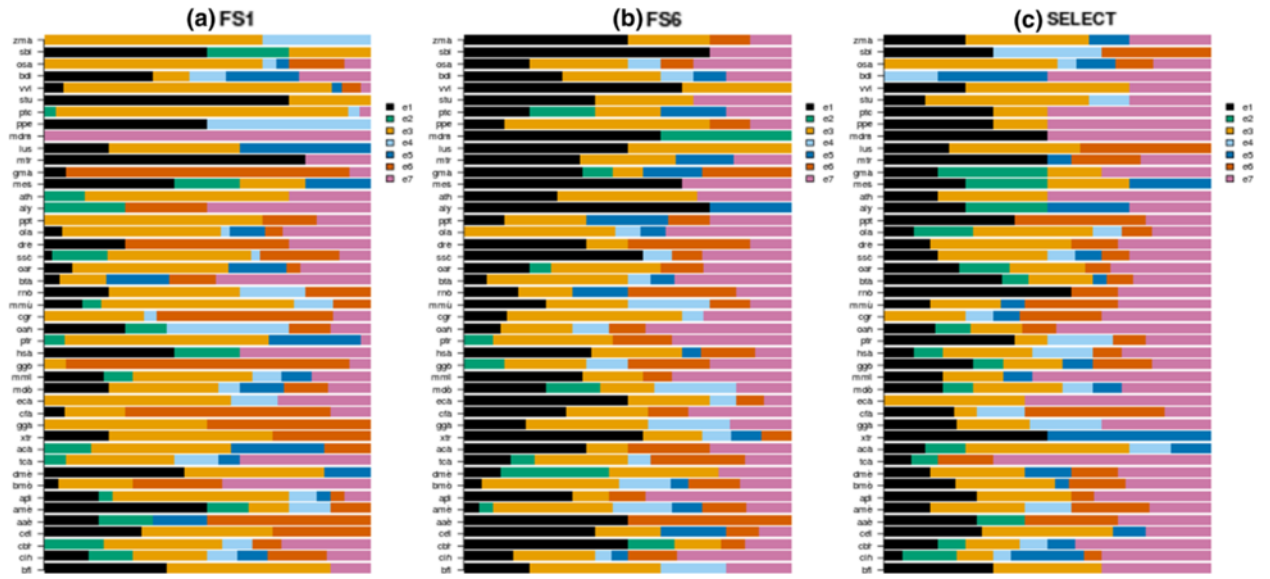

Fig. 6 Distribution of classification errors per species. Exclusive errors by SVMs (e_1), RF (e_2), J48 (e_3), SVMs and RF (e_4), SVMs and J48 (e_5), RF and J48 (e_6) and SVMs and RF and J48 (e_7). Figures (**a**), (**b**) and (**c**) illustrate that, fixing the feature set, the number of instances simultaneously misclassified by two or three learning algorithms depends on species

spaces. Moreover, e_7, the proportion of instances misclassified simultaneously by classifiers induced by the three algorithms varied by 25 % between 3.2 % and 6.7 % ($3.7\% \leq e_7 \leq 6.7\%$). These two facts 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 7 shows the comparisons between the 44 classifiers, as defined in Table 2. According to Fig. 7, the ensembles Ewv8_SVMs, Emv8_SVMs, Emv24, Ewv24, Emv8_RF, Ewv8_RF and the classifiers obtained with the new feature sets presented better predictive accuracies than the 24 previously discussed (Fig. 1), 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 7 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 Figs. 5, 7 and 8, it is possible to 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. [32]) may enhance the performance of miRNA discovery tools in the classification of pre-miRNAs of different species.

Benchmarking

The main focus of this research was to investigate the importance of different feature sets throughout species, and not to develop a new method for predicting miRNAs. The authors believe that the results from this study will potentially help in the design of new methods, which should then be subject to benchmarking against other state of the art methods; an analysis which is not applicable to this study. Moreover, since the adopted approach compared the predictive accuracies of classifiers obtained with a feature set in different species, the computational experiments were designed to produce classifiers whose predictive accuracies differ only by the species. For example, the low availability of positive examples for many

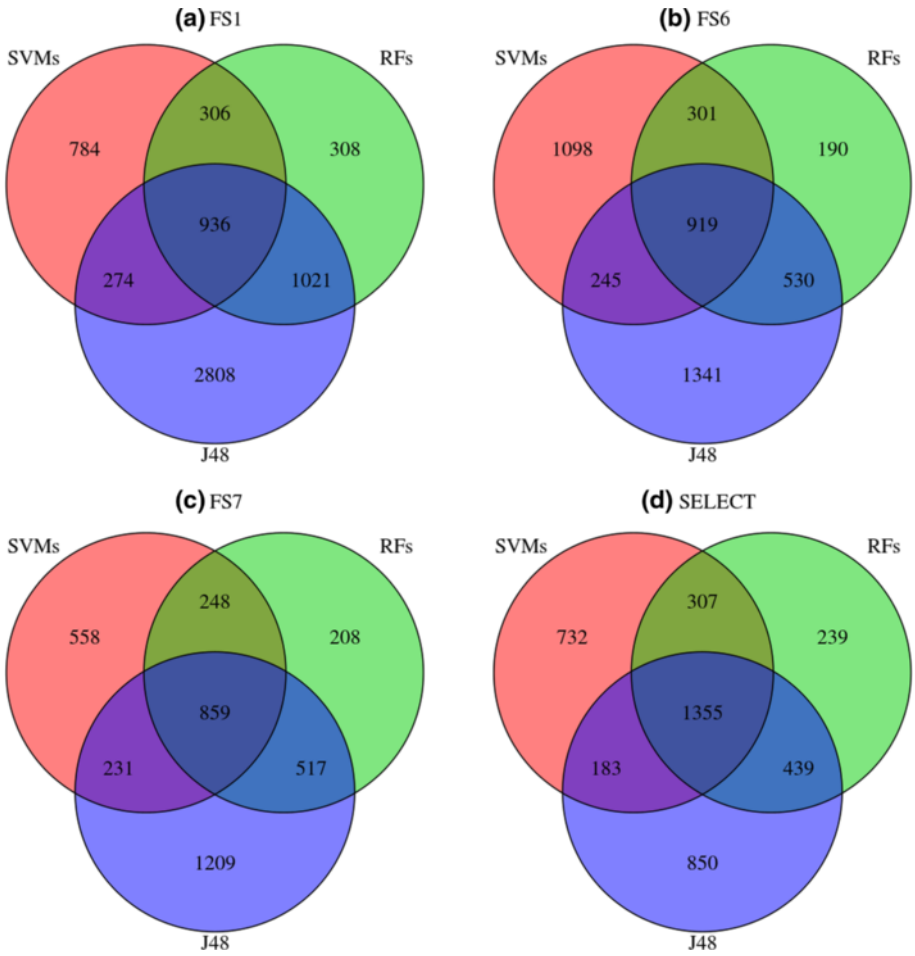

Fig. 7 Venn diagram of the classification errors of the classification algorithms, by feature set. Results were obtained from the classification of 27,000 = 45 (test species) ×10 (repetitions) ×60 (30+,30-). Figures (**a**), (**b**), (**c**) and (**d**) illustrate that the number of instances simultaneously misclassified by two or three learning algorithms depends on the feature set

species imposed the restriction of using small equal sized training sets (60+,60-) for all species, which certainly leads to lower predictive accuracies. Nevertheless, the ensembles evaluated in this work were benchmarked against 36 single classifiers, from which seven feature sets and algorithms combination are implemented in seven tools from the literature.

Finally, the scripts necessary scripts to reproduce the results presented in this work are publicly available from: http://dx.doi.org/10.5281/zenodo.49754.

Conclusion

The increase in sequencing capacity and the computational analysis of large amounts of sequencing data to detect miRNAs 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. As a consequence, the performance of these tools is usually reduced when data sets from species not used in their development are analyzed. Building species-specific miRNA discovery tools may not be always viable, for example for lack of training data. 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 between species relate to 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 one potential way to go forward, using ensemble methods built with computationally efficient features.

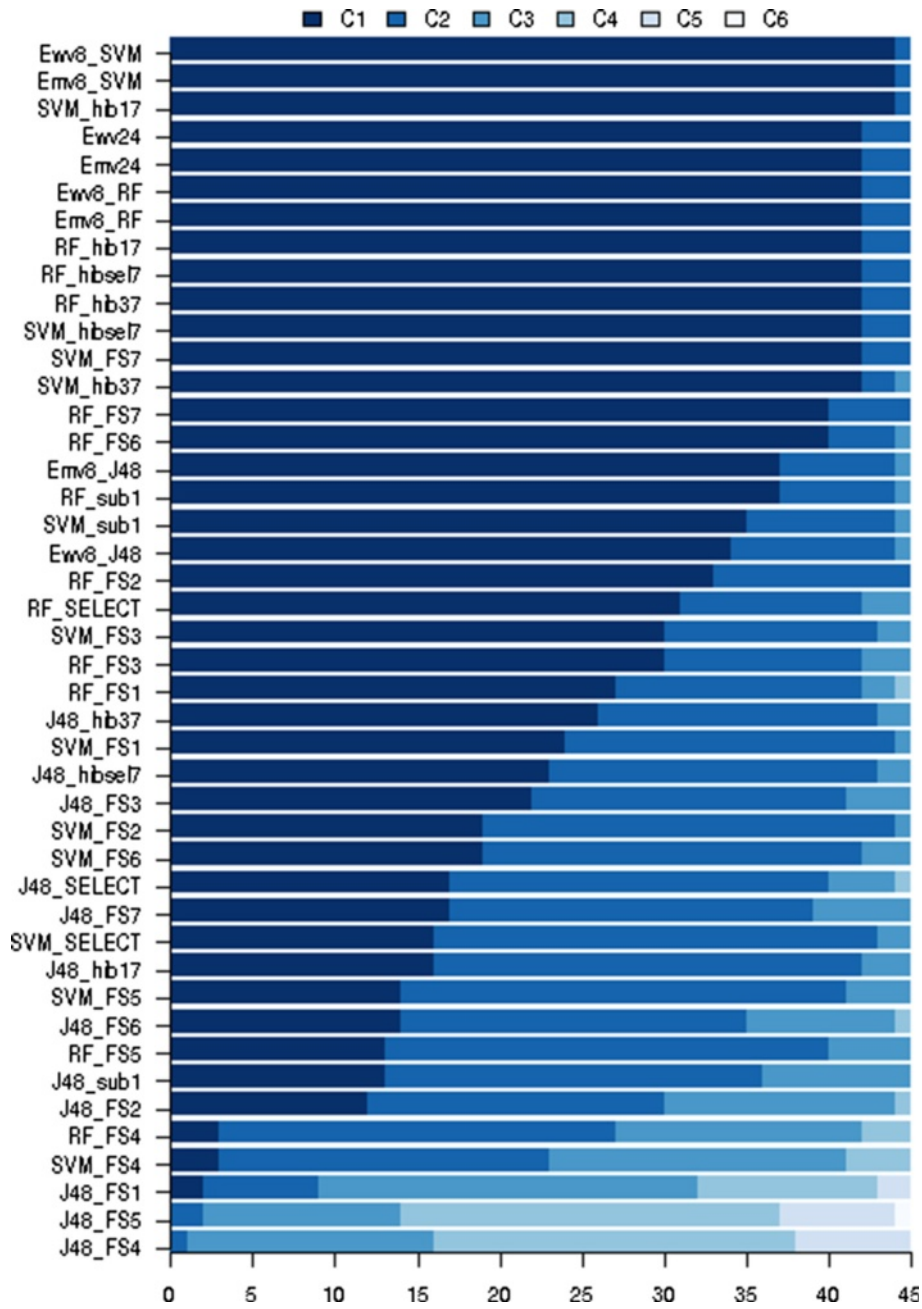

Fig. 8 Distribution of the accuracies of 44 classifiers within the accuracy clusters. $\text{Mean}_{C_1} > \ldots > \text{Mean}_{C_6}$

Additional files

Additional file 1: 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. (PDF 17.8 kb)

Additional file 2: Phylum/division, subphylum/class, species, acronyms, number of redundant negative examples out of 1,000 sequences excised from CDS or pseudo genes and the corresponding website link for download. (PDF 27.1 kb)

Additional file 3: Phylogeny and predictive accuracy of 26 metazoan species. Phylogeny extracted from the phylogenetic tree in TreeFam (http://www.treefam.org/browse#tabview=tab2) [36]. (JPG 719 kb)

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

Authors' contributions
AS and AC conceived and supervised the study. IL assembled the data, implemented the scripts, ran the experiments and summarized the results. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Empresa Brasileira de Pesquisa Agropecuária, Embrapa Soja, Caixa Postal 231, Londrina-PR, CEP 86001-970, Brasil. [2]Department of Computer Science, 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.

Received: 30 July 2015 Accepted: 12 April 2016
Published online: 28 May 2016

References

1. Lim LP, Lau NC, Weinstein EG, Abdelhakim A, Yekta S, Rhoades MW, Burge CB, Bartel DP. The microRNAs of Caenorhabditis elegans. Gene Dev. 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;116:281–97.
4. Berezikov E. Evolution of microRNA diversity and regulation in animals. Nat Rev Genet. 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 Ther. 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. Nat Biotechnol. 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 Res. 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 Res. 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 Res. 2011;39(Web Server issue):W132–8. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3125730&tool=pmcentrez&rendertype=abstract.
10. Ragan C, Mowry BJ, Bauer DC. Hybridization-based reconstruction of small non-coding RNA transcripts from deep sequencing data. Nucleic Acids Res. 2012;40(16):7633–43.
11. Jha A, Shankar R. miReader: Discovering novel miRNAs in species without sequenced genome. PLoS ONE. 2013;8(6):e66857. http://dx.doi.org/10.1371/journal.pone.0066857.
12. 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.
13. 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.
14. Scott AJ, Knott M. A cluster analysis method for grouping means in the analysis of variance. Biometrics. 1974;30(3):507–12. http://dx.doi.org/10.2307/2529204.
15. Jelihovschi EG, Faria JC, Allaman IB. The ScottKnott Clustering Algorithm. Ilheus, Bahia, Brasil: Universidade Estadual de, Santa Cruz - UESC; 2013.
16. Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res. 2014;42(Database issue):D68–73. http://nar.oxfordjournals.org/content/42/D1/D68.
17. Kamanu TKK, Radovanovic A, Archer JAC, Bajic VB. Exploration of miRNA families for hypotheses generation. Sci Rep. 2940;3:. http://www.nature.com/srep/2013/131015/srep02940/full/srep02940.html.
18. 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.
19. 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.
20. Toll-Riera M, Radó-Trilla N, Martys F, Albà MM. Role of low-complexity sequences in the formation of novel protein coding sequences. Mol Biol Evol. 2012;29(3):883–6. http://mbe.oxfordjournals.org/content/early/2011/12/08/molbev.msr263.full.
21. Batuwita R, Palade V. microPred: effective classification of pre-miRNAs for human miRNA gene prediction. Bioinformatics. 2009;25(8):989–95.
22. Bonnet E, Wuyts J, Rouzé P, Van de Peer Y. Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. Bioinformatics. 2004;20(17):2911–7.
23. 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;6:310.
24. Ng Kwang Loong S, Mishra SK. Unique folding of precursor microRNAsquantitative evidence and implications for de novo identification. RNA (New York, N.Y.) 2007;13(2):170–87. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1781370&tool=pmcentrez&rendertype=abstract.
25. Mathews DH. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. RNA. 2004;10(8):1178–90. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1370608&tool=pmcentrez&rendertype=abstract.
26. Nam JW, Shin KR, Han J, Lee Y, Kim NV, Zhang BT. Human microRNA prediction through a probabilistic co-learning model of sequence and structure. Nucleic Acids Res. 2005;33(11):3570–81.
27. Quinlan JR. C4.5: programs for machine learning. San, Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 1993.
28. Hornik K, Buchta C, Zeileis A. Open-source machine learning: R Meets Weka. Comput Stat. 2009;24(2):225–32.
29. Witten IH, Frank E. Data mining: practical machine learning tools and techniques, 2nd edition. San Francisco: Morgan Kaufmann; 2005.
30. Chang CC, Lin CJ. LIBSVM: A library for support vector machines. ACM Trans Intell Syst Technol. 2011;2:27:1–27:27. [Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm].
31. Liaw A, Wiener M. Classification and Regression by randomForest. R News. 2002;2(3):18–22. http://CRAN.R-project.org/doc/Rnews/.
32. 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.
33. 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.
34. Liu X, He S, Skogerbø G, 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.
35. 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. Nucleic Acid Res. 2007;35(suppl 2):W339–44.
36. Li H, Coghlan A, Ruan J, Coin LJ, Hériché JK, Osmotherly L, Li R, Liu T, Zhang Z, Bolund L, Wong GKS, Zheng W, Dehal P, Wang J, Durbin R. TreeFam: a curated database of phylogenetic trees of animal gene families. Nucleic Acids Res. 2006;34(Database issue):D572–80. http://europepmc.org/articles/PMC1347480.