algorithms

MDPI

Article

Classification of Precursor MicroRNAs from Different Species Based on K-mer Distance Features

Malik Yousef [1,*] **and Jens Allmer** [2,*]

[1] Department of Information Systems, Zefat Academic College, Zefat 13206, Israel
[2] Institute of Measurement and Sensor Technology, Hochschule Ruhr West University of Applied Sciences, 45479 Mülheim an der Ruhr, Germany
* Correspondence: malik.yousef@gmail.com (M.Y.); jens@allmer.de (J.A.)

Abstract: MicroRNAs (miRNAs) are short RNA sequences that are actively involved in gene regulation. These regulators on the post-transcriptional level have been discovered in virtually all eukaryotic organisms. Additionally, miRNAs seem to exist in viruses and might also be produced in microbial pathogens. Initially, transcribed RNA is cleaved by Drosha, producing precursor miRNAs. We have previously shown that it is possible to distinguish between microRNA precursors of different clades by representing the sequences in a k-mer feature space. The k-mer representation considers the frequency of a k-mer in the given sequence. We further hypothesized that the relationship between k-mers (e.g., distance between k-mers) could be useful for classification. Three different distance-based features were created, tested, and compared. The three feature sets were entitled inter k-mer distance, k-mer location distance, and k-mer first–last distance. Here, we show that classification performance above 80% (depending on the evolutionary distance) is possible with a combination of distance-based and regular k-mer features. With these novel features, classification at closer evolutionary distances is better than using k-mers alone. Combining the features leads to accurate classification for larger evolutionary distances. For example, categorizing *Homo sapiens* versus Brassicaceae leads to an accuracy of 93%. When considering average accuracy, the novel distance-based features lead to an overall increase in effectiveness. On the contrary, secondary-structure-based features did not lead to any effective separation among clades in this study. With this line of research, we support the differentiation between true and false miRNAs detected from next-generation sequencing data, provide an additional viewpoint for confirming miRNAs when the species of origin is known, and open up a new strategy for analyzing miRNA evolution.

Keywords: microRNA sequence; machine learning; differentiate miRNAs among species; k-mer miRNA categorization

check for updates

Citation: Yousef, M.; Allmer, J. Classification of Precursor MicroRNAs from Different Species Based on K-mer Distance Features. *Algorithms* **2021**, *14*, 132. https://doi.org/10.3390/a14050132

Academic Editor: Davide Verzotto

Received: 12 March 2021
Accepted: 21 April 2021
Published: 22 April 2021

1. Background

Dysregulation of gene expression is a hallmark of many diseases. MicroRNAs (miRNAs) are also expressed, and their differential regulation has been explored in many studies aimed at finding disease markers. MicroRNAs are post-transcriptional regulators that modify protein expression. Their production starts with expression as primary miRNAs (pri-miRNAs) followed by processing by Drosha in the nucleus [1]. This processing leads to precursor miRNAs (pre-miRNAs) that are exported into the cytosol. Mature miRNAs are single-stranded RNA sequences of 18–24 nt in length, which are incorporated as the recognition element into RISC. They are produced from precursor miRNAs via Dicer processing. Once incorporated into RISC, the loaded miRISC complex can act on its target messenger RNAs, with the mature miRNA providing targeting specificity.

Although microRNAs have been found in large parts of the phylogenetic tree, the molecular pathways of plants and animals may have evolved independently [2]. However, both pathways share the general processing from pri-miRNA to mature miRNA with the production of pre-miRNA and the final incorporation into a protein complex involved in translational silencing. While plants are eukaryotes and can be expected to have an miRNA pathway [3], finding miRNAs in viruses may be surprising, but it is efficient for them to encode a small RNA with a potentially large effect [4]. It is important to note that miRNAs are only functional if they are coexpressed with their targets [5]. When both miRNA and target mRNA are present, modulation of the targets' protein expression can occur [6]. MicroRNAs are not expressed at all times and in all tissues, and some may only be expressed in response to cellular stresses. Therefore, it is not possible to determine all miRNAs, their targets, and their interactions experimentally. Hence, computational approaches for miRNA prediction are important, and many such approaches are available [7–9]. A large part of the tools are based on machine learning (ML). With a few notable exceptions [10–12], two-class classification is the basis for ML-based miRNA prediction tools. Although it has been found that the complete procedure is essential for ML model establishment [13], one important factor is that negative data used in model establishment comes without any quality guarantee. Positive data, while also containing questionable examples [14–16], is generally of higher quality because positive interactions can be queried in the lab. In contrast, it would be difficult to do so for negative examples. MicroRNAs and their targets are collected in databases. Examples of such databases are miRTarBase [17], TarBase [18], and MirGeneDB [15], which generally depend on miRBase [19], which is the primary collection of all miRNAs.

To be able to apply machine learning, it is essential to represent pre-miRNAs in a vector space. Therefore, feature extraction and the types of features used are crucial for model performance [20]. An abundance of features for encoding miRNAs and their secondary structure have been proposed [21]. Since miRNA genesis is a multistep process involving several protein complexes, the structural features of pre-miRNAs are essential [22]. Krol et al. [23] evaluated some miRNA features experimentally using biochemical methods to discover precursor structures and compared their findings to predictive approaches. They found some differences between methods, which were more pronounced for the overall predicted precursor structure but did not strongly affect the stability of its termini.

Almost all published features have been implemented by Saçar Demirci et al. [20] and miRNAfe [24]. Features can be differentiated into sequence-based, thermodynamic, probabilistic, structural, or a mixture thereof. All features can further be normalized, e.g., using another feature. Typically, features such as stem length or the number of stems are used for normalization. The tool izMiR implements all ML-based approaches and especially dissects the various feature sets that are used in ML-based miRNA prediction [20].

K-mers are short nucleotide sequences with the length k, and they have been used for ML-based ab initio detection of pre-miRNAs from the onset [25]. We were interested in finding out whether the pre-miRNA sequence (not considering the secondary structure) contains enough discriminating power to categorize miRNAs among species. It could be hypothesized that there may be sequence-based recognition via the protein machinery of the miRNA pathway. Additionally, we took into account the phylogenetic relationship among species to investigate whether there is a discernable difference allowing the separation of the miRNAs of various species.

We set forth to answering these questions while removing the uncertainty of the quality of the negative data. For this, we employed the pre-miRNAs of one species/clade as positive data and the pre-miRNAs of another species/clade as negative data [26]. Hence, we only used positive data for two-class classification. For ML with the selected data, we used the random forest algorithm. We found that species that are phylogenetically distantly related could be distinguished by ML models established in this way. In another study [27], we had employed information-theoretic features to investigate their ability to categorize miRNAs among species/clades. However, we found that the usage of information-theoretic features did not outperform k-mers for this type of analysis. In this work, novel features based on k-mers, more specifically the distance among k-mers within a sequence, were used. The results were compared to k-mer performance and other previously published features. The average performance of the k-mer distance features was a bit higher than k-mers alone (~1%). On the other hand, they were somewhat less effective than selected features from all categories (~0.6% on average). We did not see any increase in performance when combining the k-mer distance-based feature set with the simple k-mer features. The novel k-mer distance features were, however, better at categorizing miRNAs at closer evolutionary distances. In conclusion, k-mer distance features can be useful for future studies aimed at categorizing miRNAs into their species of origin, especially for closer evolutionary distances. Categorizing miRNAs to their species of origin can aid in contaminant detection when predicting miRNAs from next-generation sequencing data. It further adds another line of evidence for miRNAs predicted from genomes and finally may present a different vantage point for the analysis of miRNA evolution. It is our aim for the future to create an automated system that can categorize miRNAs into their clade/species of origin.

2. Methods

2.1. Feature Space of Precursor miRNA

One important step in applying machine learning is the step of feature extraction and representation. In our data, the examples are given as a sequence of nucleotides where each sequence is a combination of four letters: A, T, C, and G. One possible representation of each sequence in a vector space is (freq of A, freq of T, freq of C, and freq of G), where freq is the frequency of the letter in the given sequence. However, this kind of representation was not successful for reaching high performance. In general, one would convert each miRNA sequence into vector $v = (v_1, v_2, \ldots, v_n)$, where each v_i corresponds to a specific feature. For precursor miRNA, different studies have considered k-mer representation. Recently, we have shown that k-mers are sufficient to allow categorizing of pre-miRNAs into species [9].

2.2. K-mer Features

Sequence-based features are commonly used for ML-based precursor miRNA analysis. Sequence-based features include patterns that can be derived from miRNAs or short words. Sequences typically consist of the nucleotides {A, U, C, G}. Short sequences with a particular length k can be referred to as k-mers or n-grams. For instance, 1-mers are the four possible "words" A, U, C, and G. Similarly, 2-mers consist of two adjacent nucleotides forming the words AA, AC, . . . , UU. Here, we used at most k = 3, which implies 64 short nucleotide sequences ranging from AAA to UUU. The number of k-mers up to and including k can be established by the following formular: $\sum_0^k 4^i$.

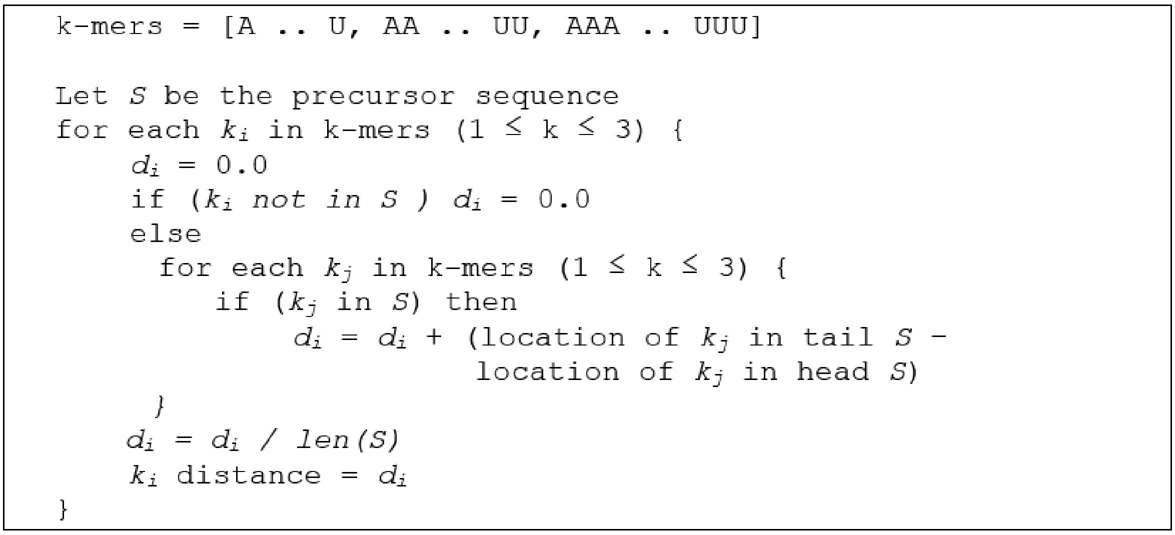
Although higher values for k have been explored [28], our preliminary tests showed that higher values of k did not add significant improvements but led to a dramatic increase of the feature space. Therefore, we chose 1-, 2-, and 3-mers as features for this work. The k-mer counts were not used directly, but their frequency was established via normalization by the length of the sequence (i.e., len(sequence) $-$ k + 1). In total, 84 features were calculated per miRNA given $k = \{1, 2, 3\}$. The k-mer frequency ranges between 0 and 1 (0 if the k-mer is not present in the sequence and one if the sequence is a repeat of a mononucleotide). The latter is unlikely to be judged as a miRNA as it would not form a secondary structure.

2.3. K-mer Distance Features

The new set of suggested features is based on the position of the k-mers within the pre-miRNA sequences. To capture the location of the k-mers and their relationship to other k-mers within the same sequence, we created three approaches. For each approach, we generated 84 features corresponding to one of the 84 k-mer features with $k \geq 1$ and $k \leq 3$.

2.3.1. Inter K-mer Distance

K-mers are distributed over the miRNA sequence, and the distance between k-mers may be important. Therefore, we designed a feature that calculates the distance between the first occurrence and last occurrence of each k-mer within the examples. The overall score is then the sum of these distances, which is further normalized by the sequence length (Figure 1).



```
k-mers = [A .. U, AA .. UU, AAA .. UUU]

Let $S$ be the precursor sequence
for each $k_i$ in k-mers ($1 \leq k \leq 3$) {
    $d_i$ = 0.0
    if ($k_i$ not in $S$ ) $d_i$ = 0.0
    else
      for each $k_j$ in k-mers ($1 \leq k \leq 3$) {
          if ($k_j$ in $S$) then
               $d_i$ = $d_i$ + (location of $k_j$ in tail $S$ -
                          location of $k_j$ in head $S$)
      }
    $d_i$ = $d_i$ / len(S)
    $k_i$ distance = $d_i$
}
```

Figure 1. The description of the algorithm calculating the inter k-mer distance features from the pre-miRNA sequence.

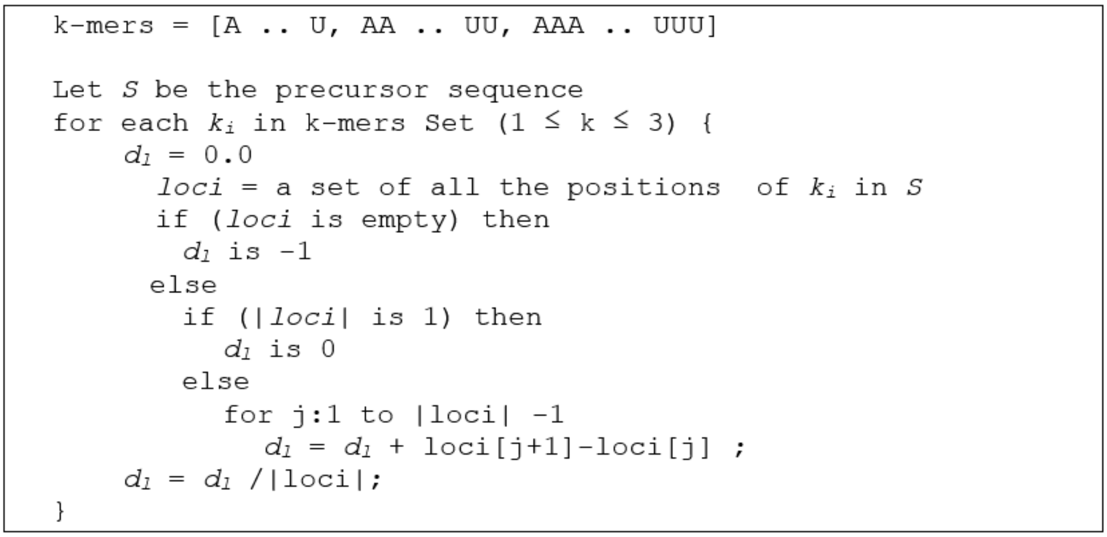
2.3.2. K-mer First–Last Distance

While the inter k-mer distance considers the distance between the first and last occurrence of a k-mer, this one measures the first and last occurrence of each k-mer directly. The measure is normalized using sequence length (S: sequence, dfl: distance first last):

$$\text{dfl} = (\text{last location of k-mer in S} - \text{first location of k-mer S})/\text{len(S)}$$

2.3.3. K-mer Location Distance

For each k-mer, all locations within the pre-miRNA sequence are recorded (loci in Figure 2). Then, the average distance among all locations is calculated (dl = dl/|loci|). In case a k-mer does not occur in the sequence, the location distance is ste to −1. In case there is only one occurrence of the k-mer in the sequence, the location distance is set to 0.



```
k-mers = [A .. U, AA .. UU, AAA .. UUU]

Let $S$ be the precursor sequence
for each $k_i$ in k-mers Set ($1 \leq k \leq 3$) {
     $d_l$ = 0.0
       loci = a set of all the positions  of $k_i$ in $S$
       if (loci is empty) then
         $d_l$ is -1
      else
         if (|loci| is 1) then
           $d_l$ is 0
         else
           for j:1 to |loci| -1
              $d_l$ = $d_l$ + loci[j+1]-loci[j] ;
     $d_l$ = $d_l$ /|loci|;
}
```

Figure 2. The description of the algorithm calculating the k-mer location distance features. |loci| depicts the number of elements in the set; loci[i] refers to the ith element in the set.

2.4. Secondary Features

Similar to what is described in [9], secondary-structure-based features were calculated for the miRNA examples. The features that were extracted were the number of loops in the structure, the number of base pairs in the stem, and the number of bulges. The number of loops was also extracted for a set size from 1 to 6. All loops greater than 6 were combined in one measure. If a loop has an even number it is symmetric, otherwise it is asymmetric.

We created a KNIME workflow [29] to extract those features. We did not calculate the secondary structure but used the one provided by miRBase [30].

2.5. Other Features Describing Pre-miRNAs

For the parameterization of pre-miRNAs, an abundance of features have been published [21]. These parameters can be categorized into sequence features such as k-mers [31], structural features such as the number of bulges [32], thermodynamic ones such as minimum free energy [33], and combinations like the triplet feature [34] consisting of one nucleotide and its adjacent secondary structure. Via normalizing or transforming the features, for example, into *p*-values [35], additional features are created, adding to the total number of parameters. The set of parameters used in different studies have influenced the prediction success, and most approaches have recently been compared [20]. All previous studies have used known pre-miRNAs as positive data. These studies used pseudo negative data to train ML classifiers. This significant contribution of this study is that no pseudo negative data were used in training ML classifiers. Instead, known miRNAs from one species were employed as positive data, while the negative data were the known miRNAs of another species. For this scenario, we have shown that it is effective to use k-mers or sequence motifs [26,36]. Performing information theory-based transformation of features [27], while effective by themselves, adds little to the classification accuracy. Here, we introduced three new sets of features (see above) and compared the previously described features for pre-miRNA detection. Out of the more than a thousand features for this (including normalized ones), we finally selected 100 in this study. The selection was based on the correlation among features (the lower the better) and the information gain (higher is better) of the features.

3. Datasets and Methods

Preprocessing the Data

Data from 15 clades were collected for this study (Table 1). The 16th example in Table 1 contains the sequences of *Homo sapiens* that were extracted from its clade Hominidae. Some miRNAs are highly conserved and exist in many copies throughout the data. This could create biased models, so we removed highly similar sequences, leaving just one representative. All data from all clades and *Homo sapiens* were combined and clustered using USEARCH [37]. From the resulting clusters, we chose one representative. This effectively created a dataset consisting of only nonhomologous sequences. Clades were re-established from the filtered dataset, but the homologous sequences between clades were not reintroduced after filtering.

Table 1. The data used in this study derives from the clades listed in the first column. The number of precursors available in miRBase at the time of download is listed in the second column. The third column shows how many precursors survive the homology filtering. Since *Homo sapiens* is well studied, it contains many examples, and therefore it was extracted from its clade Hominidae. Clade names are from miRBase. If two names are provided, the name in parentheses refers to the NCBI taxonomy.

	#Precursors	**#Uniques**
Hominidae	3629	1326
Brassicaceae	726	535
Hexapoda	3119	2050
Monocotyledons (Liliopsida)	1598	1402
Nematoda	1789	1632
Fabaceae	1313	1011
Pisces (Chondricthyes)	1530	682
Virus	306	295
Aves	948	790
Laurasiatheria	1205	675
Rodentia	1778	993
Homo sapiens	1828	1223
Cercopithecidae	631	503
Embryophyta	287	278
Malvaceae	458	419
Platyhelminthes	424	381
Total	21,569	14,195

4. Feature Vector and Feature Selection

Many features to parameterize pre-miRNAs are known, and this study employed 831 features used in a previous study [21]. Additionally, novel k-mer distance features were introduced. For feature selection, we employed information gain (IG) [38,39] as it is implemented in KNIME (version 3.1.2) [29].

Information gain was used for feature selection by evaluating each feature/variable's information content in the context of the target variable. The formula to compute IG is

$$I(C, A) = H(C) - H(C|A)$$

where $H(C) = \sum_{c \in C} p(C) \log p(C)$ represents the entropy of the class, and $H(C\,|\,A)$ is the conditional entropy of the class provided feature A:

$$H(C|A) = -\sum_{c \in C} p(C|A) \log p(C|A)$$

5. Classification Approach

Similar to the study of [26], we trained random forest (RF) classifiers [40] using the RF implementation of KNIME [29]. For training and testing, we split the overall data into 80% training and 20% testing data. We used undersampling of the majority class to force positive and negative examples to equal amounts. Cross-validation for model performance estimation was performed using 100-fold Monte Carlo cross-validation (MCCV) [41]. For training and testing, we used the default setting for the RF implementation by KNIME.

6. Model Performance Evaluation

Model performance was assessed using, for example, the Matthews's correlation coefficient (MCC) [42]. Sensitivity, specificity, and accuracy were other measures used for the evaluation of model performance. A number of performance measures were calculated from the confusion matrix containing true positive (TP), false positive (FP), true negative (TN), and false negative (FN) classifications:

$$\text{Sensitivity (SE, Recall)} = \text{TP}/(\text{TP} + \text{FN}),$$

$$\text{Specificity (SP)} = \text{TN}/(\text{TN} + \text{FP}),$$

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}),$$

$$\text{F-Measure} = 2 * (\text{precision} * \text{recall})/(\text{precision} + \text{recall}),$$

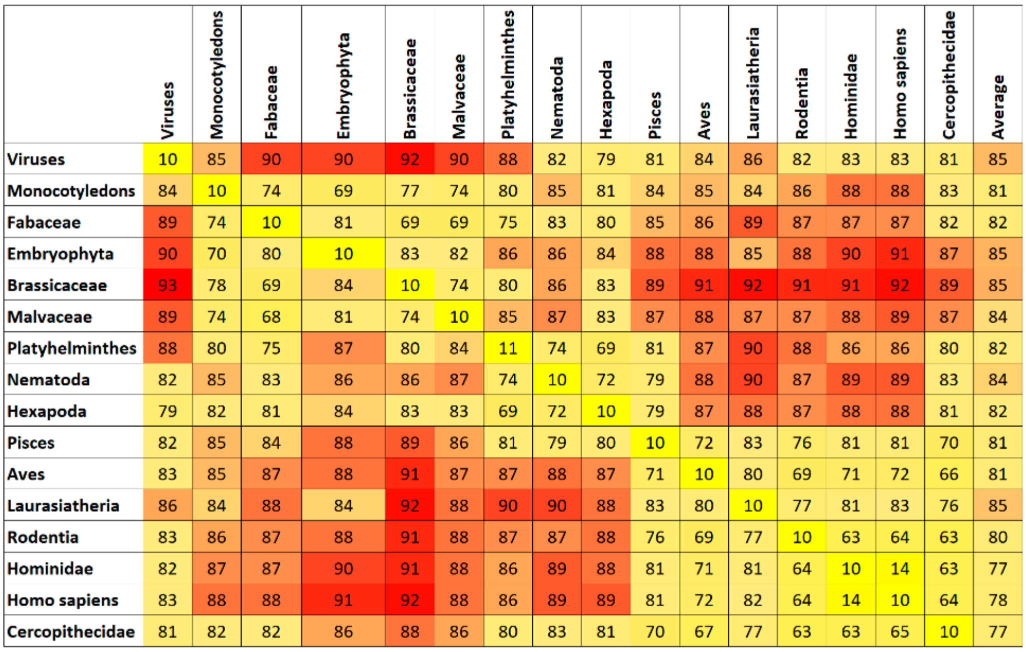
$$\text{Accuracy (ACC)} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN}), \text{ and}$$

$$\text{Matthews correlation coefficient (MCC)} = \frac{(\text{TP}/\text{TN} - \text{FP}/\text{FN})}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FN})(\text{TN}+\text{FP})}}$$

In the following, the presented results always refer to the average model performance for 100-fold MCCV.

7. Results and Discussion

We have previously shown that k-mers can categorize miRNAs into species [27]. To improve the categorization, we devised three new sets of features also based on k-mers. The data used for categorization derives from clades with sufficient amounts of examples for classification (Table 1). The data represents various evolutionary distances (Figure 1). Previously, we have shown that it is possible to categorize species accurately at larger evolutionary distances, but categorizing becomes more difficult with decreasing evolutionary distance [26]. Therefore, it is important to have examples from varying evolutionary distances, in this case, also from various kingdoms (Figure 3).



	Viruses	Monocotyledons	Fabaceae	Embryophyta	Brassicaceae	Malvaceae	Platyhelminthes	Nematoda	Hexapoda	Pisces	Aves	Laurasiatheria	Rodentia	Hominidae	Homo sapiens	Cercopithecidae	Average
Viruses	10	85	90	90	92	90	88	82	79	81	84	86	82	83	83	81	85
Monocotyledons	84	10	74	69	77	74	80	85	81	84	85	84	86	88	88	83	81
Fabaceae	89	74	10	81	69	69	75	83	80	85	86	89	87	87	87	82	82
Embryophyta	90	70	80	10	83	82	86	86	84	88	88	85	88	90	91	87	85
Brassicaceae	93	78	69	84	10	74	80	86	83	89	91	92	91	91	92	89	85
Malvaceae	89	74	68	81	74	10	85	87	83	87	88	87	87	88	89	87	84
Platyhelminthes	88	80	75	87	80	84	11	74	69	81	87	90	88	86	86	80	82
Nematoda	82	85	83	86	86	87	74	10	72	79	88	90	87	89	89	83	84
Hexapoda	79	82	81	84	83	83	69	72	10	79	87	88	87	88	88	81	82
Pisces	82	85	84	88	89	86	81	79	80	10	72	83	76	81	81	70	81
Aves	83	85	87	88	91	87	87	88	87	71	10	80	69	71	72	66	81
Laurasiatheria	86	84	88	84	92	88	90	90	88	83	80	10	77	81	83	76	85
Rodentia	83	86	87	88	91	88	87	87	88	76	69	77	10	63	64	63	80
Hominidae	82	87	87	90	91	88	86	89	88	81	71	81	64	10	14	63	77
Homo sapiens	83	88	88	91	92	88	86	89	89	81	72	82	64	14	10	64	78
Cercopithecidae	81	82	82	86	88	86	80	83	81	70	67	77	63	63	65	10	77

Figure 3. Results in the figure stem from k-mer features only. Average accuracy for 100-fold Monte Carlo cross-validation using a random forest classifier and a split of 80% training and 20% testing. Yellow shades indicate lower accuracy, while red shades show higher average classification accuracy.

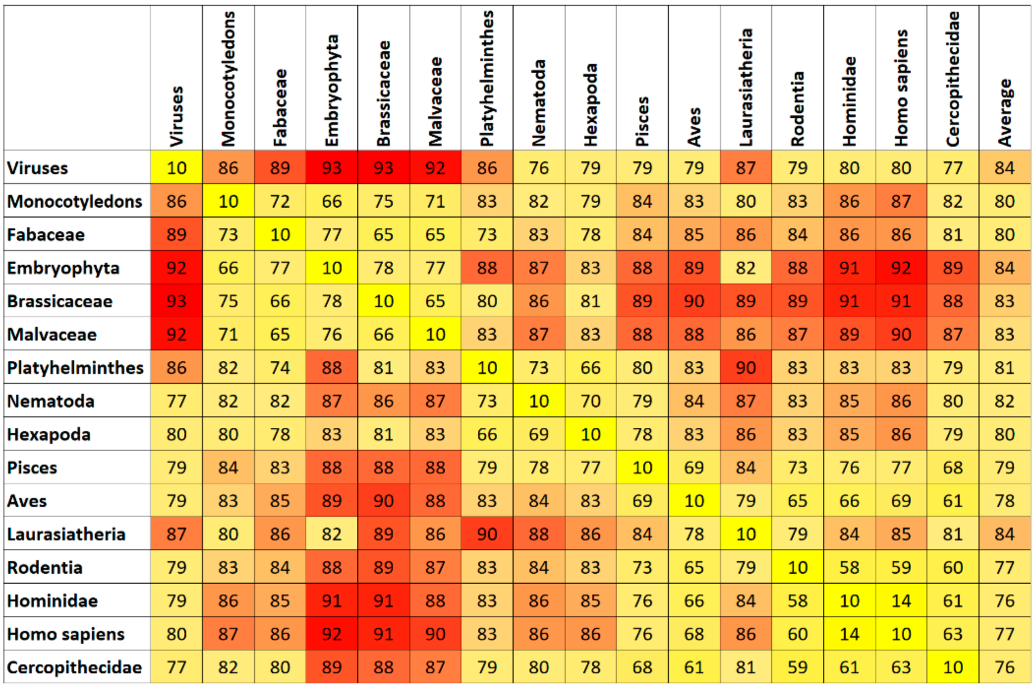
Each combination of species and clades needs a specific classifier, and these were trained using 100-fold MCCV. Initially, the known k-mer features were evaluated with respect to categorization accuracy (Figure 3). With one clade used for positive data and the other as negative data, 100 models were established using 80% training and 20% testing data. Classifier performance represents the average performance of 100-fold MCCV. These computations lead to a matrix where both columns and rows represent clades. Considering Aves versus Hexapoda as an example, the average accuracy amounts to 0.87 at 100-fold MCCV. This performance is 5% better than the general average of all established models. All the accuracy values on the diagonal of the table are the results of the clade with itself.

This categorization is obviously very difficult, and accordingly, all accuracy values along the diagonal are very low.

With the results in Figure 3, our previous observation regarding the effectiveness of k-mer features [26,27], namely that the average model accuracy also increases with increasing evolutionary distance, can be confirmed. Although we have removed similar sequences between *Homo sapiens* and its clade Hominidae, the performance is low (0.14 ACC), which indicates that some "hidden message" encoded in perhaps the triplet bias is conserved for the whole Hominidae clade. This observation is even more intriguing as the performance of Embryophyta versus its child clades is much better (0.70–0.78 ACC). Perhaps this is due to the larger evolutionary distance between Embryophyta and its child clades selected here, or the miRNAs may not be as strictly conserved within the plant kingdom.

Considering the closest clades in the lower side of the phylogenetic tree (Figure 1), Cercopithecidae, Homo sapiens, and Rodentia, the performance is also the lowest (63–65 ACC; Table 1). Surprisingly, the performance of Brassicaceae and Malvaceae, which are also very closely related, is much better (74).

K-mers have been introduced to parameterize pre-miRNAs early on [25], and many other sequence-based features have followed. Here, we did not use k-mers directly but transformed them to create new parameters do describe pre-miRNAs. Differentiation between pre-miRNAs from different species/clades was tested in the same manner as for k-mers (Figure 4).

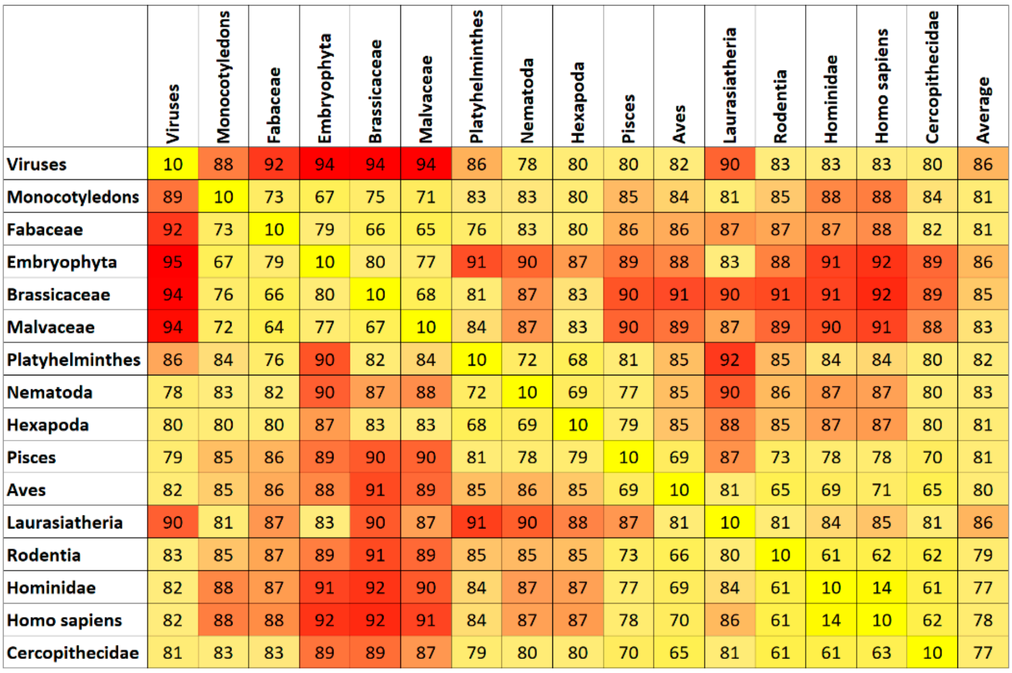


	Viruses	Monocotyledons	Fabaceae	Embryophyta	Brassicaceae	Malvaceae	Platyhelminthes	Nematoda	Hexapoda	Pisces	Aves	Laurasiatheria	Rodentia	Hominidae	Homo sapiens	Cercopithecidae	Average
Viruses	10	86	89	93	93	92	86	76	79	79	79	87	79	80	80	77	84
Monocotyledons	86	10	72	66	75	71	83	82	79	84	83	80	83	86	87	82	80
Fabaceae	89	73	10	77	65	65	73	83	78	84	85	86	84	86	86	81	80
Embryophyta	92	66	77	10	78	77	88	87	83	88	89	82	88	91	92	89	84
Brassicaceae	93	75	66	78	10	65	80	86	81	89	90	89	89	91	91	88	83
Malvaceae	92	71	65	76	66	10	83	87	83	88	88	86	87	89	90	87	83
Platyhelminthes	86	82	74	88	81	83	10	73	66	80	83	90	83	83	83	79	81
Nematoda	77	82	82	87	86	87	73	10	70	79	84	87	83	85	86	80	82
Hexapoda	80	80	78	83	81	83	66	69	10	78	83	86	83	85	86	79	80
Pisces	79	84	83	88	88	88	79	78	77	10	69	84	73	76	77	68	79
Aves	79	83	85	89	90	88	83	84	83	69	10	79	65	66	69	61	78
Laurasiatheria	87	80	86	82	89	86	90	88	86	84	78	10	79	84	85	81	84
Rodentia	79	83	84	88	89	87	83	84	83	73	65	79	10	58	59	60	77
Hominidae	79	86	85	91	91	88	83	86	85	76	66	84	58	10	14	61	76
Homo sapiens	80	87	86	92	91	90	83	86	86	76	68	86	60	14	10	63	77
Cercopithecidae	77	82	80	89	88	87	79	80	78	68	61	81	59	61	63	10	76

Figure 4. Results in the figure stem from k-mer distance features only. Average accuracy for 100-fold Monte Carlo cross-validation using a random forest classifier and a split of 80% training and 20% testing. Red shades indicate higher average classification accuracy, while yellow shades indicate lower accuracy.

The inter k-mer distance was tested first (Figure 4). Like k-mers (Figure 3), roughly three hotspots with high accuracy for differentiation among species/clades can be observed. The most striking result is that virus pre-miRNAs seem to be very different from pre-miRNAs from the plant kingdom but much closer to pre-miRNAs from the metazoan clade. Since most viruses in the list infect animals, this may imply that their pre-miRNAs are functional in the host rather than for virus regulative activities, similar to what we had earlier proposed for *Toxoplasma gondii* [41,42]. The other hotspots separate kingdoms from each other, which supports the previous observation that the performance of each pair of clades is correlated to the phylogenetic distance between the two clades. In summary, k-mer features alone are on average about 2% better than inter k-mer distance features.

Employing our new k-mer first–last distance (Figure 5) leads to similar results as for k-mers (Figure 3) and inter k-mer distance (Figure 4). The average performance for k-mer first–last distance is equal to k-mers. Interestingly, k-mer first–last distance is slightly better for viruses, Embryophyta, and Laurasiatheria. Therefore, we checked the difference between k-mer and k-mer first–last distance results (Figure 6). It is interesting to see that k-mer first–last distance performance is better for more closely related species. At the same time, k-mers seem to be better at larger evolutionary distances. This difference can be leveraged when a rough categorization has already been achieved with some prior method.

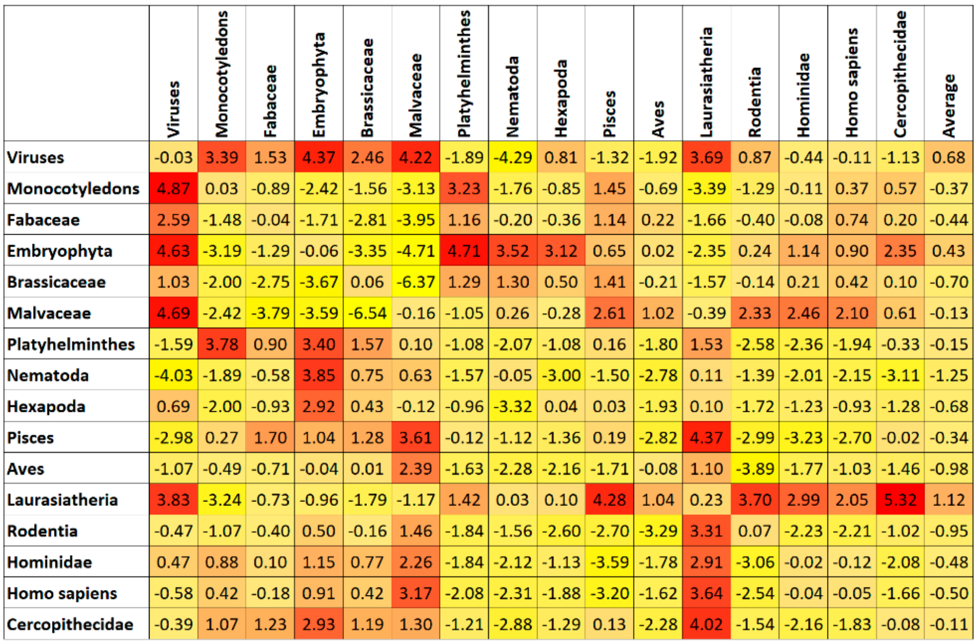


	Viruses	Monocotyledons	Fabaceae	Embryophyta	Brassicaceae	Malvaceae	Platyhelminthes	Nematoda	Hexapoda	Pisces	Aves	Laurasiatheria	Rodentia	Hominidae	Homo sapiens	Cercopithecidae	Average
Viruses	10	88	92	94	94	94	86	78	80	80	82	90	83	83	83	80	86
Monocotyledons	89	10	73	67	75	71	83	83	80	85	84	81	85	88	88	84	81
Fabaceae	92	73	10	79	66	65	76	83	80	86	86	87	87	87	88	82	81
Embryophyta	95	67	79	10	80	77	91	90	87	89	88	83	88	91	92	89	86
Brassicaceae	94	76	66	80	10	68	81	87	83	90	91	90	91	91	92	89	85
Malvaceae	94	72	64	77	67	10	84	87	83	90	89	87	89	90	91	88	83
Platyhelminthes	86	84	76	90	82	84	10	72	68	81	85	92	85	84	84	80	82
Nematoda	78	83	82	90	87	88	72	10	69	77	85	90	86	87	87	80	83
Hexapoda	80	80	80	87	83	83	68	69	10	79	85	88	85	87	87	80	81
Pisces	79	85	86	89	90	90	81	78	79	10	69	87	73	78	78	70	81
Aves	82	85	86	88	91	89	85	86	85	69	10	81	65	69	71	65	80
Laurasiatheria	90	81	87	83	90	87	91	90	88	87	81	10	81	84	85	81	86
Rodentia	83	85	87	89	91	89	85	85	85	73	66	80	10	61	62	62	79
Hominidae	82	88	87	91	92	90	84	87	87	77	69	84	61	10	14	61	77
Homo sapiens	82	88	88	92	92	91	84	87	87	78	70	86	61	14	10	62	78
Cercopithecidae	81	83	83	89	89	87	79	80	80	70	65	81	61	61	63	10	77

Figure 5. Results in the figure stem from k-mer first–last distance features only. Average accuracy for 100-fold MCCV using a random forest classifier and a split of 80% training and 20% testing. Yellow shades refer to lower accuracy, while red shades indicate higher average classification accuracy.

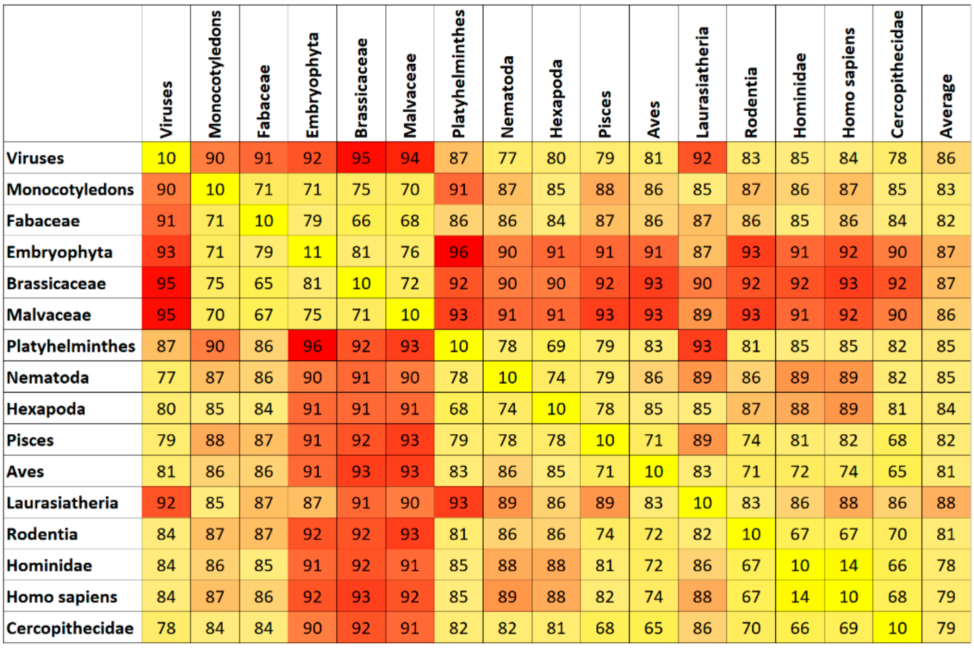
However, combining k-mer and the three k-mer distance feature sets and selecting the top 100 features according to information gain leads to only a slight increase in accuracy over k-mer features alone (~1%).

For comparison, we used other features, including structural features such as the number of bulges [32], thermodynamic ones such as minimum free energy [33], and combinations such as the triplet feature [34] consisting of one nucleotide and its adjacent folding structure. The same procedure that was used for the establishment of k-mer and k-mer distance models was employed. The results are reported in Figure 7. The average accuracy is 0.83, just like the one for k-mer and k-mer distance features combined. The areas of Figure 7 with high accuracy are also mostly similar to the areas for k-mers (Figure 3) and the combination of k-mer features. Some of the features, especially the probabilistic ones, are computationally expensive and lead to long run times. It can be concluded that their calculation is not warranted when aiming to categorize pre-miRNAs to their species.



	Viruses	Monocotyledons	Fabaceae	Embryophyta	Brassicaceae	Malvaceae	Platyhelminthes	Nematoda	Hexapoda	Pisces	Aves	Laurasiatheria	Rodentia	Hominidae	Homo sapiens	Cercopithecidae	Average
Viruses	-0.03	3.39	1.53	4.37	2.46	4.22	-1.89	-4.29	0.81	-1.32	-1.92	3.69	0.87	-0.44	-0.11	-1.13	0.68
Monocotyledons	4.87	0.03	-0.89	-2.42	-1.56	-3.13	3.23	-1.76	-0.85	1.45	-0.69	-3.39	-1.29	-0.11	0.37	0.57	-0.37
Fabaceae	2.59	-1.48	-0.04	-1.71	-2.81	-3.95	1.16	-0.20	-0.36	1.14	0.22	-1.66	-0.40	-0.08	0.74	0.20	-0.44
Embryophyta	4.63	-3.19	-1.29	-0.06	-3.35	-4.71	4.71	3.52	3.12	0.65	0.02	-2.35	0.24	1.14	0.90	2.35	0.43
Brassicaceae	1.03	-2.00	-2.75	-3.67	0.06	-6.37	1.29	1.30	0.50	1.41	-0.21	-1.57	-0.14	0.21	0.42	0.10	-0.70
Malvaceae	4.69	-2.42	-3.79	-3.59	-6.54	-0.16	-1.05	0.26	-0.28	2.61	1.02	-0.39	2.33	2.46	2.10	0.61	-0.13
Platyhelminthes	-1.59	3.78	0.90	3.40	1.57	0.10	-1.08	-2.07	-1.08	0.16	-1.80	1.53	-2.58	-2.36	-1.94	-0.33	-0.15
Nematoda	-4.03	-1.89	-0.58	3.85	0.75	0.63	-1.57	-0.05	-3.00	-1.50	-2.78	0.11	-1.39	-2.01	-2.15	-3.11	-1.25
Hexapoda	0.69	-2.00	-0.93	2.92	0.43	-0.12	-0.96	-3.32	0.04	0.03	-1.93	0.10	-1.72	-1.23	-0.93	-1.28	-0.68
Pisces	-2.98	0.27	1.70	1.04	1.28	3.61	-0.12	-1.12	-1.36	0.19	-2.82	4.37	-2.99	-3.23	-2.70	-0.02	-0.34
Aves	-1.07	-0.49	-0.71	-0.04	0.01	2.39	-1.63	-2.28	-2.16	-1.71	-0.08	1.10	-3.89	-1.77	-1.03	-1.46	-0.98
Laurasiatheria	3.83	-3.24	-0.73	-0.96	-1.79	-1.17	1.42	0.03	0.10	4.28	1.04	0.23	3.70	2.99	2.05	5.32	1.12
Rodentia	-0.47	-1.07	-0.40	0.50	-0.16	1.46	-1.84	-1.56	-2.60	-2.70	-3.29	3.31	0.07	-2.23	-2.21	-1.02	-0.95
Hominidae	0.47	0.88	0.10	1.15	0.77	2.26	-1.84	-2.12	-1.13	-3.59	-1.78	2.91	-3.06	-0.02	-0.12	-2.08	-0.48
Homo sapiens	-0.58	0.42	-0.18	0.91	0.42	3.17	-2.08	-2.31	-1.88	-3.20	-1.62	3.64	-2.54	-0.04	-0.05	-1.66	-0.50
Cercopithecidae	-0.39	1.07	1.23	2.93	1.19	1.30	-1.21	-2.88	-1.29	0.13	-2.28	4.02	-1.54	-2.16	-1.83	-0.08	-0.11

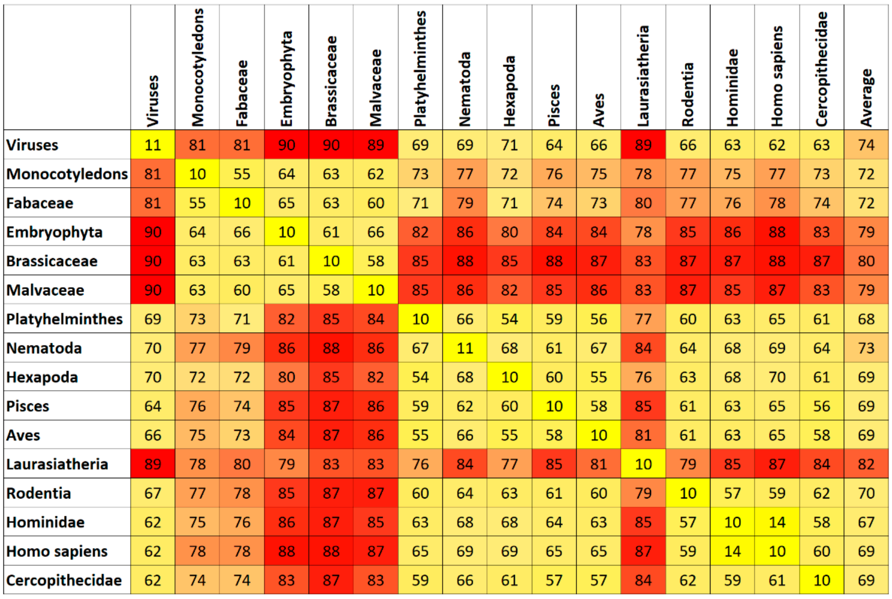
Figure 6. Difference between k-mer first–last distance and k-mers only. Yellow shades indicate better performance for k-mer first–last distance, while red shades indicate higher average classification accuracy for k-mers.



	Viruses	Monocotyledons	Fabaceae	Embryophyta	Brassicaceae	Malvaceae	Platyhelminthes	Nematoda	Hexapoda	Pisces	Aves	Laurasiatheria	Rodentia	Hominidae	Homo sapiens	Cercopithecidae	Average
Viruses	10	90	91	92	95	94	87	77	80	79	81	92	83	85	84	78	86
Monocotyledons	90	10	71	71	75	70	91	87	85	88	86	85	87	86	87	85	83
Fabaceae	91	71	10	79	66	68	86	86	84	87	86	87	86	85	86	84	82
Embryophyta	93	71	79	11	81	76	96	90	91	91	91	87	93	91	92	90	87
Brassicaceae	95	75	65	81	10	72	92	90	90	92	93	90	92	92	93	92	87
Malvaceae	95	70	67	75	71	10	93	91	91	93	93	89	93	91	92	90	86
Platyhelminthes	87	90	86	96	92	93	10	78	69	79	83	93	81	85	85	82	85
Nematoda	77	87	86	90	91	90	78	10	74	79	86	89	86	89	89	82	85
Hexapoda	80	85	84	91	91	91	68	74	10	78	85	85	87	88	89	81	84
Pisces	79	88	87	91	92	93	79	78	78	10	71	89	74	81	82	68	82
Aves	81	86	86	91	93	93	83	86	85	71	10	83	71	72	74	65	81
Laurasiatheria	92	85	87	87	91	90	93	89	86	89	83	10	83	86	88	86	88
Rodentia	84	87	87	92	92	93	81	86	86	74	72	82	10	67	67	70	81
Hominidae	84	86	85	91	92	91	85	88	88	81	72	86	67	10	14	66	78
Homo sapiens	84	87	86	92	93	92	85	89	88	82	74	88	67	14	10	68	79
Cercopithecidae	78	84	84	90	92	91	82	82	81	68	65	86	70	66	69	10	79

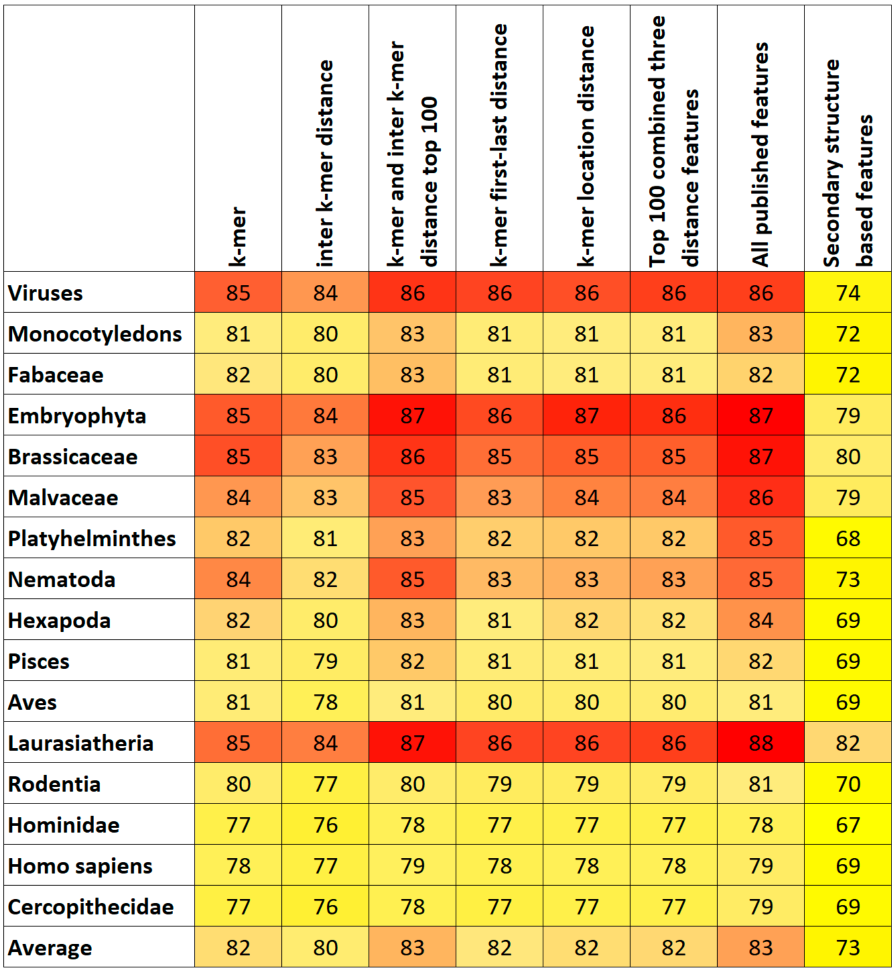
Figure 7. Most published features describing a pre-miRNA (about 700), excluding the features that only rely on sequence information. The results presented in the figure stem from models trained with the top 100 features selected from previously published features. Yellow shades indicate better performance for k-mer first–last distance, while red shades indicate higher average classification accuracy for k-mers.

Since categorization was very successful using non-sequence-based features, we decided to extract features that focus on structure, similar to Yousef et al., 2006 [9]. The results are presented in Figure 8. While the higher accuracy areas in Figure 8 are similar to the above results, the overall accuracy is far below (10%) the accuracy achieved with k-mer or k-mer distance-based features (Figure 9).



	Viruses	Monocotyledons	Fabaceae	Embryophyta	Brassicaceae	Malvaceae	Platyhelminthes	Nematoda	Hexapoda	Pisces	Aves	Laurasiatheria	Rodentia	Hominidae	Homo sapiens	Cercopithecidae	Average
Viruses	11	81	81	90	90	89	69	69	71	64	66	89	66	63	62	63	74
Monocotyledons	81	10	55	64	63	62	73	77	72	76	75	78	77	75	77	73	72
Fabaceae	81	55	10	65	63	60	71	79	71	74	73	80	77	76	78	74	72
Embryophyta	90	64	66	10	61	66	82	86	80	84	84	78	85	86	88	83	79
Brassicaceae	90	63	63	61	10	58	85	88	85	88	87	83	87	87	88	87	80
Malvaceae	90	63	60	65	58	10	85	86	82	85	86	83	87	85	87	83	79
Platyhelminthes	69	73	71	82	85	84	10	66	54	59	56	77	60	63	65	61	68
Nematoda	70	77	79	86	88	86	67	11	68	61	67	84	64	68	69	64	73
Hexapoda	70	72	72	80	85	82	54	68	10	60	55	76	63	68	70	61	69
Pisces	64	76	74	85	87	86	59	62	60	10	58	85	61	63	65	56	69
Aves	66	75	73	84	87	86	55	66	55	58	10	81	61	63	65	58	69
Laurasiatheria	89	78	80	79	83	83	76	84	77	85	81	10	79	85	87	84	82
Rodentia	67	77	78	85	87	87	60	64	63	61	60	79	10	57	59	62	70
Hominidae	62	75	76	86	87	85	63	68	68	64	63	85	57	10	14	58	67
Homo sapiens	62	78	78	88	88	87	65	69	69	65	65	87	59	14	10	60	69
Cercopithecidae	62	74	74	83	87	83	59	66	61	57	57	84	62	59	61	10	69

Figure 8. Average accuracy for 100-fold MCCV using a random forest classifier and a split of 80% training and 20% testing employing only secondary-structure-based features. Yellow shades indicate lower accuracy, while red shades show higher average classification accuracy.



	k-mer	inter k-mer distance	k-mer and inter k-mer distance top 100	k-mer first-last distance	k-mer location distance	Top 100 combined three distance features	All published features	Secondary structure based features
Viruses	85	84	86	86	86	86	86	74
Monocotyledons	81	80	83	81	81	81	83	72
Fabaceae	82	80	83	81	81	81	82	72
Embryophyta	85	84	87	86	87	86	87	79
Brassicaceae	85	83	86	85	85	85	87	80
Malvaceae	84	83	85	83	84	84	86	79
Platyhelminthes	82	81	83	82	82	82	85	68
Nematoda	84	82	85	83	83	83	85	73
Hexapoda	82	80	83	81	82	82	84	69
Pisces	81	79	82	81	81	81	82	69
Aves	81	78	81	80	80	80	81	69
Laurasiatheria	85	84	87	86	86	86	88	82
Rodentia	80	77	80	79	79	79	81	70
Hominidae	77	76	78	77	77	77	78	67
Homo sapiens	78	77	79	78	78	78	79	69
Cercopithecidae	77	76	78	77	77	77	79	69
Average	82	80	83	82	82	82	83	73

Figure 9. Summary of the results from Figures 3–5, 7 and 8. Yellow shades indicate lower accuracy, while red shades show higher average classification accuracy.

In summary (Figure 9), it can be seen that secondary-structure-based features display less performance for the categorization of miRNAs among species/clades. There appears to be little difference in employing all published features, the top 100 selected ones (based on low correlation among features and high information gain), k-mer, or k-mer distance features. There still appears to be sequence information in the selected features. For example, the triplet structure features consist of a nucleotide and the local hybridization pattern: N, where N is any nucleotide, and (non)bonds are represented by dots and parentheses, respectively. Among the novel features, the k-mer location distance performed best. It achieved a comparable accuracy to using k-mers alone (data not shown).

Top Features

We ranked all the distance features according to information gain. For more details, see our GitHub repository, which lists the ranked features for each pair of clades. We merged the distance features “Inter k-mer distance” and “k-mer first–last distance” for each pair for this analysis.

Table 2 confirms the expectation that different features are important for the categorization of different pairs of clades. While Inter k-mer distance features are most prominent, the particular sequence that is most discriminating varies. A complete analysis can be found on our GitHub repository.

Table 2. An example of ranked features based on IG. The features ending with “_all” are related to “Inter k-mer distance”, while the rest are related to the k-mer first–last distance.

Aves vs. Brassicaceae		**Hexapoda vs. Embryophyta**	
Feature	**IG**	**Feature**	**IG**
AU_all	0.45	G_all	0.36
U_all	0.43	U_all	0.36
A_all	0.42	A_all	0.34
UU_all	0.41	UG_all	0.33
UA_all	0.41	C_all	0.33
AA_all	0.39	CU_all	0.30
AUU_all	0.39	AG_all	0.29
AAU_all	0.37	CA_all	0.29
G_all	0.37	GC_all	0.28
AUA_all	0.35	UU_all	0.27
UAU_all	0.34	GA_all	0.26
AU_AU	0.34	AU_all	0.25
GA_all	0.33	GU_all	0.24
AUU_AUU	0.33	UC_all	0.23
UUU_UUU	0.32	AC_all	0.22
UU_UU	0.32	GG_all	0.21
C_all	0.32	UGC_all	0.20
UA_UA	0.32	CC_all	0.19
AAU_AAU	0.31	UA_all	0.19
GU_all	0.31	AA_all	0.19

8. Conclusions

MicroRNAs are small noncoding RNA sequences that are involved in post-transcriptional gene regulation, modulating protein abundance. Representatives of these miRNAs are found throughout the tree of life. How miRNA evolved is under current investigation [42]. This is related to our previous research, where we had shown the possibility of categorizing miRNAs to their species of origin [26,27].

In this study, we designed three transformations of k-mers (k-mer location distance, k-mer first–last, and inter k-mer) for miRNA parameterization. These features were used in machine learning to differentiate between miRNAs from different species/clades. The distinction performance was compared to using k-mer features and previously published features. Random forest models were established using 100-fold MCCV, with one species or clade contributing the positive class and another posing as the negative one. To assess the categorization effectiveness into species/clades, examples needed to be selected from a wide range of phylogenetic distances (Table 1, Figure 1). More than 100 models were established per feature set comparison, leading to the establishment and comparison of about 100,000 models for this study.

As can be expected, more distant species/clades can be categorized more effectively, which can be seen by the clustering of higher average accuracy (Figures 3–5, 7 and 8). The finding that k-mer features can categorize well at larger evolutionary distances confirms our previous results [26,27]. Using only the features k-mer inter or k-mer location distance leads to a similar performance compared to employing only k-mer features. Combining all features followed by the selection of the top 100 features (based on low correlation among features and high information gain) slightly increases the average accuracy by 1% (Figure 9). The latter is equal to the effectiveness of using all published features. However, calculating all these features is very time-consuming and amounted to thousands of CPU hours for this study. Naturally, these features contain many sequence-based ones. In an attempt to evaluate the contribution of the secondary structure to the categorization of miRNAs, we selected parameters describing the secondary structure of pre-miRNAs (Figure 8). Categorization with secondary-structure-based features is less accurate (~10% on average) compared to other approaches (Figure 9). This finding is in line with the conservation of structure over the conservation of sequence paradigm. A more in-depth comparison of results showed that the effectiveness is not equal when considering k-mers and k-mer distance features (Figure 6). K-mer distance features are slightly more effective for categorization at closer evolutionary distances.

In conclusion, k-mer and k-mer distance features together lead to more accurate categorization at larger evolutionary distances. An example could be the categorization of miRNAs into *Homo sapiens* versus Brassicaceae, which achieves an average accuracy of 93%. The novel distance-based transformation of the k-mer features increase the average accuracy at closer evolutionary distances. The use of secondary-structure-based features did not lead to a favorable performance in this study (Figure 9). We hope that these findings will provide a new angle to study miRNA evolution. More practically, categorization of miRNAs to their species of origin can help ensure that predicted miRNAs conform to the expectation of the organism they are predicted for. Similarly, when predicting miRNAs from NGS data, categorizing the predictions into their species of origin can reveal whether the predictions are contaminations or not. To better support these processes, we aim to implement a fully automatic categorization method in the future.

Author Contributions: Methodology, M.Y. and J.A.; Writing—original draft, M.Y. and J.A. All authors have read and agreed to the published version of the manuscript.

Funding: The work was supported by Zefat Academic College to M.Y.

Data Availability Statement: Not Applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Availability and Implementation: We performed all analysis using the KNIME platform and have made the analysis workflow available at GitHub: https://github.com/malikyousef/microRNAs-based-on-K-mer-Distance-Features- (doi:10.5281/zenodo.4662625) (accessed on 21 April 2021).

References

1. Erson-Bensan, A.E. Introduction to MicroRNAs in Biological Systems. *Methods Mol. Biol.* **2014**, *1107*, 1–14. [CrossRef]
2. Chapman, E.J.; Carrington, J.C. Specialization and Evolution of Endogenous Small RNA Pathways. *Nat. Rev. Genet.* **2007**, *8*, 884–896. [CrossRef] [PubMed]
3. Yousef, M.; Allmer, J.; Khalifa, W. Plant MicroRNA Prediction Employing Sequence Motifs Achieves High Accuracy. 2016. Available online: https://www.researchgate.net/publication/320402782_Plant_microRNA_prediction_employing_sequence_motifs_achieves_high_accuracy (accessed on 21 April 2021).
4. Grey, F. Role of MicroRNAs in Herpesvirus Latency and Persistence. *J. Gen. Virol.* **2015**, *96*, 739–751. [CrossRef] [PubMed]

5. Saçar, M.D.; Allmer, J. Current Limitations for Computational Analysis of MiRNAs in Cancer. *Pak. J. Clin. Biomed. Res.* **2013**, *1*, 3–5.
6. Yousef, M.; Trinh, H.V.; Allmer, J. Intersection of MicroRNA and Gene Regulatory Networks and Their Implication in Cancer. *Curr. Pharm. Biotechnol.* **2014**, *15*, 445–454. [CrossRef]
7. Allmer, J.; Yousef, M. Computational Methods for Ab Initio Detection of MicroRNAs. *Front. Genet.* **2012**, *3*, 209. [CrossRef]
8. Saçar, M.; Allmer, J. Machine Learning Methods for MicroRNA Gene Prediction. In *miRNomics: MicroRNA Biology and Computational Analysis SE-10*; Yousef, M., Allmer, J., Eds.; Methods in Molecular Biology; Humana Press: Tortowa, NJ, USA, 2014; Volume 1107, pp. 177–187, ISBN 978-1-62703-747-1.
9. Yousef, M.; Nebozhyn, M.; Shatkay, H.; Kanterakis, S.; Showe, L.C.; Showe, M.K. Combining Multi-Species Genomic Data for MicroRNA Identification Using a Naive Bayes Classifier. *Bioinformatics* **2006**, *22*, 1325–1334. [CrossRef] [PubMed]
10. Dang, H.T.; Tho, H.P.; Satou, K.; Tu, B.H. Prediction of MicroRNA Hairpins Using One-Class Support Vector Machines. In Proceedings of the 2nd International Conference on Bioinformatics and Biomedical Engineering, iCBBE 2008, Shanghai, China, 16–18 May 2008; pp. 33–36.
11. Khalifa, W.; Yousef, M.; Demirci, M.D.S.; Allmer, J. The Impact of Feature Selection on One and Two-Class Classification Performance for Plant MicroRNAs. *PeerJ* **2016**, *4*, e2135. [CrossRef] [PubMed]
12. Yousef, M.; Jung, S.; Showe, L.C.; Showe, M.K. Learning from Positive Examples When the Negative Class Is Undetermined—MicroRNA Gene Identification. *Algorithms Mol. Biol. AMB* **2008**, *3*, 2. [CrossRef]
13. Demirci, M.D.S.; Allmer, J. Delineating the Impact of Machine Learning Elements in Pre-MicroRNA Detection. *PeerJ* **2017**, *5*, e3131. [CrossRef]
14. Saçar, M.D.; Hamzeiy, H.; Allmer, J. Can MiRBase Provide Positive Data for Machine Learning for the Detection of MiRNA Hairpins? *J. Integr. Bioinform.* **2013**, *10*, 215. [CrossRef] [PubMed]
15. Fromm, B.; Billipp, T.; Peck, L.E.; Johansen, M.; Tarver, J.E.; King, B.L.; Newcomb, J.M.; Sempere, L.F.; Flatmark, K.; Hovig, E.; et al. A Uniform System for the Annotation of Vertebrate MicroRNA Genes and the Evolution of the Human MicroRNAome. *Annu. Rev. Genet.* **2015**, *49*, 213–242. [CrossRef] [PubMed]
16. Duygu, M.; Demirci, S.; Allmer, J. Improving the Quality of Positive Datasets for the Establishment of Machine Learning Models for Pre- MicroRNA Detection. *J. Integr. Bioinform.* **2017**, *14*, 20170032.
17. Hsu, S.-D.; Tseng, Y.-T.; Shrestha, S.; Lin, Y.-L.; Khaleel, A.; Chou, C.-H.; Chu, C.-F.; Huang, H.-Y.; Lin, C.-M.; Ho, S.-Y.; et al. MiRTarBase Update 2014: An Information Resource for Experimentally Validated MiRNA-Target Interactions. *Nucleic Acids Res.* **2014**, *42*, D78–D85. [CrossRef] [PubMed]
18. Vergoulis, T.; Vlachos, I.S.; Alexiou, P.; Georgakilas, G.; Maragkakis, M.; Reczko, M.; Gerangelos, S.; Koziris, N.; Dalamagas, T.; Hatzigeorgiou, A.G. TarBase 6.0: Capturing the Exponential Growth of MiRNA Targets with Experimental Support. *Nucleic Acids Res.* **2012**, *40*, D222–D229. [CrossRef] [PubMed]
19. Kozomara, A.; Griffiths-Jones, S. MiRBase: Integrating MicroRNA Annotation and Deep-Sequencing Data. *Nucleic Acids Res.* **2011**, *39*, D152–D157. [CrossRef] [PubMed]
20. Demirci, M.D.S.; Baumbach, J.; Allmer, J. On the Performance of Pre-MicroRNA Detection Algorithms. *Nat. Commun.* **2017**, *8*, 330. [CrossRef]
21. Sacar, M.D.; Allmer, J. Data Mining for Microrna Gene Prediction: On the Impact of Class Imbalance and Feature Number for Microrna Gene Prediction. In Proceedings of the 2013 8th International Symposium on Health Informatics and Bioinformatics, Ankara, Turkey, 25–27 September 2013.
22. Sewer, A.; Paul, N.; Landgraf, P.; Aravin, A.; Pfeffer, S.; Brownstein, M.J.; Tuschl, T.; van Nimwegen, E.; Zavolan, M. Identification of Clustered MicroRNAs Using an Ab Initio Prediction Method. *BMC Bioinform.* **2005**, *6*, 267. [CrossRef] [PubMed]
23. Krol, J.; Sobczak, K.; Wilczynska, U.; Drath, M.; Jasinska, A.; Kaczynska, D.; Krzyzosiak, W.J. Structural Features of MicroRNA (MiRNA) Precursors and Their Relevance to MiRNA Biogenesis and Small Interfering RNA/Short Hairpin RNA Design. *J. Biol. Chem.* **2004**, *279*, 42230–42239. [CrossRef]
24. Yones, C.A.; Stegmayer, G.; Kamenetzky, L.; Milone, D.H. MiRNAfe: A Comprehensive Tool for Feature Extraction in MicroRNA Prediction. *BioSystems* **2015**, *138*, 1–5. [CrossRef]
25. Lai, E.C.; Tomancak, P.; Williams, R.W.; Rubin, G.M. Computational Identification of Drosophila MicroRNA Genes. *Genome Biol.* **2003**, *4*, R42. [CrossRef] [PubMed]
26. Yousef, M.; Khalifa, W.; Acar, I.E.; Allmer, J. MicroRNA Categorization Using Sequence Motifs and K-Mers. *BMC Bioinform.* **2017**, *18*, 170. [CrossRef] [PubMed]
27. Yousef, M.; Nigatu, D.; Levy, D.; Allmer, J.; Henkel, W. Categorization of Species Based on Their MicroRNAs Employing Sequence Motifs, Infor-Mation-Theoretic Sequence Feature Extraction, and k-Mers. *EURASIP J. Adv. Signal Process.* **2017**, *2017*. [CrossRef]
28. Cakir, M.V.; Allmer, J. Systematic Computational Analysis of Potential RNAi Regulation in Toxoplasma Gondii. In Proceedings of the 2010 5th International Symposium on Health Informatics and Bioinformatics, Ankara, Turkey, 20–22 April 2010.
29. Berthold, M.R.; Cebron, N.; Dill, F.; Gabriel, T.R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. KNIME: The Konstanz Information Miner. Available online: https://www.knime.com/sites/default/files/knime_whitepaper.pdf (accessed on 21 April 2021).
30. Griffiths-Jones, S. MiRBase: MicroRNA Sequences and Annotation. *Curr. Protoc. Bioinform.* **2010**, *29*, 12.9.1–12.9.10. [CrossRef]

31. Ng, K.L.S.; Mishra, S.K. De Novo SVM Classification of Precursor MicroRNAs from Genomic Pseudo Hairpins Using Global and Intrinsic Folding Measures. *Bioinformatics* **2007**, *23*, 1321–1330. [CrossRef]
32. Ritchie, W.; Gao, D.; Rasko, J.E.J. Defining and Providing Robust Controls for MicroRNA Prediction. *Bioinformatics* **2012**, *28*, 1058–1061. [CrossRef]
33. 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 Acids Res.* **2007**, *35*, W339–W344. [CrossRef]
34. Xue, C.; Li, F.; He, T.; Liu, G.-P.; Li, Y.; Zhang, X. Classification of Real and Pseudo MicroRNA Precursors Using Local Structure-Sequence Features and Support Vector Machine. *BMC Bioinform.* **2005**, *6*, 310. [CrossRef]
35. Yousef, M.; Allmer, J.; Khalifa, W. Sequence Motif-Based One-Class Classifiers Can Achieve Comparable Accuracy to Two-Class Learners for Plant MicroRNA Detection. *J. Biomed. Sci. Eng.* **2015**. [CrossRef]
36. Edgar, R.C. Search and Clustering Orders of Magnitude Faster than BLAST. *Bioinformatics* **2010**, *26*, 2460–2461. [CrossRef]
37. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
38. Xu, Q.-S.; Liang, Y.-Z. Monte Carlo Cross Validation. *Chemom. Intell. Lab. Syst.* **2001**, *56*, 1–11. [CrossRef]
39. Matthews, B.W. Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. *BBA Protein Struct.* **1975**, *405*, 442–451. [CrossRef]
40. Saçar Demirci, M.D.; Bağci, C.; Allmer, J. Differential Expression of Toxoplasma Gondii MicroRNAs in Murine and Human Hosts. Available online: https://openaccess.iyte.edu.tr/xmlui/bitstream/handle/11147/7918/10.1007@978-3-319-39496-19.pdf;jsessionid=D7A7AB90CE83A13466B77615F319E128?sequence=1 (accessed on 21 April 2021).
41. Saçar, M.D.; Bağcı, C.; Allmer, J. Computational Prediction of MicroRNAs from Toxoplasma Gondii Potentially Regulating the Hosts' Gene Expression. *Genom. Proteom. Bioinform.* **2014**, *12*, 228–238. [CrossRef] [PubMed]
42. Tanzer, A.; Stadler, P.F. Evolution of MicroRNAs. *Methods Mol. Biol.* **2006**, *342*, 335–350. [CrossRef] [PubMed]