EURASIP Journal on Advances
in Signal Processing

RESEARCH Open Access

Categorization of species based on their microRNAs employing sequence motifs, information-theoretic sequence feature extraction, and *k*-mers

Malik Yousef[1*], Dawit Nigatu[2], Dalit Levy[1], Jens Allmer[3,4] and Werner Henkel[2]

Abstract

Background: Diseases like cancer can manifest themselves through changes in protein abundance, and microRNAs (miRNAs) play a key role in the modulation of protein quantity. MicroRNAs are used throughout all kingdoms and have been shown to be exploited by viruses to modulate their host environment. Since the experimental detection of miRNAs is difficult, computational methods have been developed. Many such tools employ machine learning for pre-miRNA detection, and many features for miRNA parameterization have been proposed. To train machine learning models, negative data is of importance yet hard to come by; therefore, we recently started to employ pre-miRNAs from one species as positive data versus another species' pre-miRNAs as negative examples based on sequence motifs and *k*-mers. Here, we introduce the additional usage of information-theoretic (IT) features.

Results: Pre-miRNAs from one species were used as positive and another species' pre-miRNAs as negative training data for machine learning. The categorization capability of IT and *k*-mer features was investigated. Both feature sets and their combinations yielded a very high accuracy, which is as good as the previously suggested sequence motif and *k*-mer based method. However, for obtaining a high performance, a sufficiently large phylogenetic distance between the species and sufficiently high number of pre-miRNAs in the training set is required. To examine the contribution of the IT and *k*-mer features, an information gain-based feature ranking was performed. Although the top 3 are IT features, 80% of the top 100 features are *k*-mers. The comparison of all three individual approaches (motifs, IT, and *k*-mers) shows that the distinction of species based on their pre-miRNAs *k*-mers are sufficient.

Conclusions: IT sequence feature extraction enables the distinction among species and is less computationally expensive than motif calculations. However, since IT features need larger amounts of data to have enough statistics for producing highly accurate results, future categorization into species can be effectively done using *k*-mers only. The biological reasoning for this is the existence of a codon bias between species which can, at least, be observed in exonic miRNAs. Future work in this direction will be the ab initio detection of pre-miRNA. In addition, prediction of pre-miRNA from RNA-seq can be done.

Keywords: MicroRNA, Sequence motifs, Pre-microRNA, Machine learning, Differentiate miRNAs among species, *k*-mer, miRNA categorization, Information theory

* Correspondence: malik.yousef@gmail.com
[1]Community Information Systems, Zefat Academic College, 13206 Zefat, Israel
Full list of author information is available at the end of the article

1 Introduction

Proteins define a phenotype, and their dysregulation often leads to a disease. Protein abundance is highly regulated, and microRNAs are responsible for its post-transcriptional modulation. Mature microRNAs (miRNAs), which act as recognition sequences for their target messenger RNAs, are produced from a molecular pathway which is different for plants and animals [1]. They have in common that pri-miRNAs are transcribed from the genome and that hairpins (pre-miRNAs) are excised from these transcripts. Each pre-miRNA can have multiple mature miRNAs (18–24 nucleotides in length) which are incorporated into a protein complex, responsible for modulating the translation efficiency of multiple targets. MicroRNAs have been shown to exist in a variety of species ranging from viruses [2] to plants [3]. MicroRNAs need to be co-expressed with their targets [4] in order to be functional, and many transcripts in an organism are only produced in response to internal or external stresses. Thus, it may not be possible to experimentally determine all miRNAs, their targets, and their interactions. Computational approaches to detect miRNAs have been developed to overcome the limitation, and most methods for pre-miRNA detection are based on machine learning [5–7]. With the exception of few approaches based on one class classification [8–10], most methods rely on two class classification. While all parts contributing to model establishment are important [11], the selection of negative data is crucial since no gold standard is available. Although other databases like miRTarBase [12], TarBase [13], and MirGeneDB [14] are available, positive data is generally derived from miRBase [15]. While negative data is of unknown quality, also positive data from miRBase contains questionable entries [14, 16] and even MirGeneDB which filters miRBase entries is not free from questionable examples [17].

Parameterization of pre-miRNAs is important for applying machine learning algorithms, and numerous features have been proposed [18]. Short sequences (*k*-mers) have been used early on for the machine learning-based ab initio detection of pre-miRNAs [19]. Since miRNA genesis depends on a pathway involving several protein complexes, structural features of pre-miRNAs have been found to be important [20]. Additionally, we have recently established the use of sequence motifs as features enabling the detection of pre-miRNAs [21, 22]. Many machine learning models for pre-miRNA detection have been established using a variety of learning algorithms and training schemes [23–26]. All the established models suffer from the selection of arbitrary examples for the negative class. Gao and colleagues [27], for example, reasoned that exons and other non-coding RNAs would be useful as negative data, but miRNAs can be derived from anywhere in the genome, including exons [28].

Due to the unknown quality of the negative data, in a previous work, we successfully used the one-class classification approach for the detection of pre-miRNAs [29, 30]. However, we realized that using positive examples to represent the negative class from different species holds a number of promises [31]. One of the promises is that it enables the categorization of pre-miRNAs into species. Hairpins can be structurally classified fairly well, and many approaches are available despite data quality issues [32–35]. Categorization of the identified pre-miRNAs into their species of origin or a very closely related one adds a further line of evidence to their identification. We established random forest machine learning models using two-class classification with the positive class being pre-miRNAs from one species and the negative pre-miRNAs from a different species. Therefore, both positive and negative classes for training and testing were derived from known pre-miRNAs, effectively removing the need for pseudo negative data. We have previously proposed the same strategy [31] using sequence motifs and *k*-mers. In this study, we further introduced information-theoretic approaches and important additional analyses. In our previous work, we showed that discrimination among miRNAs from different species is possible which is likely due to alleged fast evolution for some miRNAs [36–38], supporting the possibility to differentiate among evolutionary distant species based on miRNAs. We then focused on sequence motifs since structure is evolutionarily more conserved than sequence. Due to the large impact of *k*-mers on the categorization in our previous work, in an attempt to add more discriminating power, here, we added information theory (IT)-based features. Apart from our previous study [31], only Lopes and colleagues attempted to use pre-miRNAs to discriminate between species [33]. However, they resorted to establishing ab initio pre-miRNA detection models with the same bias on negative data as existing pre-miRNA detection methods [26, 39–42]; using the same training and testing strategies [32, 42–44]. Furthermore, a large part of the features they used assesses structural features of pre-miRNAs, which poses problems when analyzing closely related species since structure is more conserved than sequence. In this work, we analyzed the discriminative power of sequence motifs, information-theoretic quantities, and *k*-mers for the categorization of pre-miRNAs into species. It became clear that *k*-mers alone can separate between species that are not strongly related. We also showed that the number of examples is important for the establishment of suitable machine learning models. If enough examples are not available for a species, a model can be established for the next higher level (genus), which may even outperform all species-based models. Since sequence motifs and IT features are computationally expensive compared to *k*-mers, it would be extremely difficult to establish models for all pairs of species for automatic categorization. However, since we were able to show that *k*-mers have enough discriminative power, automatic species categorization will become possible in the future. All in all, this work not only provides an important additional line of evidence for detecting pre-miRNAs but is also useful for studies depending on deep sequencing data which often contains contaminated sequences [45].

2 Methods

2.1 Datasets

All data were downloaded from miRBase [46] Release 21. From the family Hominidae (3629 hairpins), *Gorilla gorilla* (ggo, 352), *Homo sapiens* (has, 1881), *Pan paniscus* (ppa, 88), *Pongo pygmaeus* (ppy, 642), *Pan troglodytes* (ptr, 655), and *Symphalangus syndactylus* (ssy, 11) were acquired. From the clade Nematoda (1856 hairpins), 10 species were downloaded: *Ascaris suum* (asu, 97), *Brugia malayi* (bma, 115), *Caenorhabditis brenneri* (cbn, 214), *Caenorhabditis briggsae* (cbr, 175), *Caenorhabditis elegans* (cel, 250), *Caenorhabditis remanei* (crm, 157), *Haemonchus contortus* (hco, 188), *Pristionchus pacificus* (ppc, 354), *Panagrellus redivivus* (prd, 200), and *Strongyloides ratti* (str, 106). From the clade which miRBase still calls pisces (1623 hairpins), the following species' data hairpins were attained: *Cyprinus carpio* (ccr, 134), *Danio rerio* (dre, 346), *Fugu rubripes* (fru, 131), *Hippoglossus hippoglossus* (hhi, 40), *Ictalurus punctatus* (ipu, 281), *Oryzias latipes* (ola, 168), *Paralichthys olivaceus* (pol, 20), *Salmo salar* (ssa, 371), and *Tetraodon nigroviridis* (tni, 132). Finally, from the group of hexapoda (3119 hairpins), *Aedes aegypti* (aae, 101), *Anopheles gambiae* (aga, 66), *Apis mellifera* (ame, 254), *Acyrthosiphon pisum* (api, 123), *Bombyx mori* (bmo, 487), *Culex quinquefasciatus* (cpu, 74), *Drosophila ananassae* (dan, 76), *Drosophila erecta* (der, 81), *Drosophila grimshawi* (dgr, 82), *Drosophila melanogaster* (dme, 256), *Drosophila mojavensis* (dmo, 71), *Drosophila persimilis* (dpe, 75), *Drosophila pseudoobscura* (dps, 210), *Drosophila sechellia* (dse, 78), *Drosophila simulans* (dsi, 135), *Drosophila virilis* (dvi, 134), *Drosophila willistoni* (dwi, 77), *Drosophila yakuba* (dya, 76), *Heliconius melpomene* (hme, 92), *Locusta migratoria* (lmi, 7), *Manduca sexta* (mse, 98), *Nasonia giraulti* (ngi, 32), *Nasonia longicornis* (nlo, 28), *Nasonia vitripennis* (nvi, 53), *Plutella xylostella* (pxy, 133), and *Tribolium castaneum* (tca, 220). In addition to these data, several clades from miRBase (e.g., the fabaceae dataset consisting of *Acacia auriculiformis*, *Arachis hypogaea*, *Acacia mangium*, *Glycine max*, *Glycine soja*, *Lotus japonicus*, *Medicago truncatula*, *Phaseolus vulgaris*, and *Vigna unguiculata* with a total of about 1400 pre-miRNAs) were used by combining all the hairpins of the species within the clade.

All hairpins were filtered for sequence similarity as in Yousef et al. [31] before training machine learning models using the Usearch tool [47].

2.2 Parameterization of pre-miRNAs

In order to allow the application of machine learning, biological features need to be translated into mathematical parameters. It is our hypothesis that structural and thermodynamic features which have previously been described [18] are evolutionarily more conserved than sequence features. Therefore, only sequence-based features were used for parameterization in this study. Sequence motifs (200) as in [31] were used as well as 84 *k*-mers and their information-theoretic transformations (91). In the following, the parameters used in this study are detailed.

2.3 *k*-mer features

Many studies performing pre-miRNA detection based on machine learning include simple sequence-based features. These features are words, *k*-mers, or *n*-grams, all of which describe a short sequence of nucleotides. Here, we use *k*-mers to describe a short nucleotide sequence of length *k*. For example, a 1-mer over the alphabet {A, U, C, G} can produce the words A, U, C, and G; a 2-mer can generate AA, AC, …, UU, and a 3-mer leads to 64 short nucleotide sequences ranging from AAA to UUU. Higher *k* have also been used [48], but here, we chose 1-, 2-, and 3-mers as features. The *k*-mer counts in a given sequence were normalized by the total number of *k*-mers in the sequence (i.e., len(sequence) − k + 1) [49]. Hence, for *k*-mers with $k = \{1, 2, 3\}$, 84 features were calculated per example. The *k*-mer frequency ranges between 0 (if the *k*-mer is not present in the sequence) and 1 (if the sequence is a repeat of a mononucleotide which is not observed since such a sequence does not fold into secondary structures).

2.4 Motif features

Motif features differ from *k*-mers since they are approximate sequence matches instead of an exact match. Motifs are discovered by searches for short overrepresented approximate sequences within a larger pool of sequences. The MEME Suite (Multiple Expectation Maximization for Motif Elicitation) [50] was used for motif discovery in our previous study [31], and the discovered motifs were used. For positive and negative data, 100 motifs were discovered, and thus, 200 features were created.

2.5 Information-theoretic features

Information-theoretic (IT) features have been widely used in computational biology and bioinformatics to measure, analyze, and model the structural and organizational properties of biological sequences. In [51], we used theses IT features for the classification of essential and non-essential genes. The IT features used in this study are 4 entropy (E), 17 mutual information (MI), 65 conditional mutual information (CMI), 1 Kullback-Leibler divergence (DKL), and 4 Markov model (M) related. Next, we will present a brief description of the information-theoretic quantities used in this study. For a more detailed explanations, we refer the reader to [52].

2.6 Mutual information

We used mutual information to measure the information between consecutive bases X and Y. The mutual information measures the dependency between two random variables and is mathematically defined as

$$I(X;Y) = \sum_{x\in\Omega}\sum_{y\in\Omega} P(x,y)\log_2\frac{P(x,y)}{P(x)P(y)}, \tag{1}$$

where $P(x)$ and $P(y)$ are the marginal probabilities and $P(x,y)$ is the joint probability and Ω is the set of nucleotides {A, C, G, U}. The probabilities are estimated from the relative frequencies in the corresponding pre-miRNA sequences. Along with the total mutual information computed according to Eq. (1), for each base pair (x,y), the quantity $P(x,y)\log_2(P(x,y)/P(x)P(y))$ is calculated and used as a feature. Thus, 17 mutual information related features are defined in this manner.

2.7 Conditional mutual information

The mutual information between two random variables X and Y conditioned on a third random variable Z having a probability mass function (pmf) $P(z)$ is given by

$$\begin{aligned} I(X;Y|Z) &= \sum_{z\in\Omega}P(z)\sum_{x\in\Omega}\sum_{y\in\Omega}P(x,y|z)\log_2\frac{P(x,y|z)}{P(x|z)P(y|z)}, \\ &= \sum_{x\in\Omega}\sum_{y\in\Omega}\sum_{z\in\Omega}P(x,y,z)\log_2\frac{P(z)P(x,y,z)}{P(x,z)P(y,z)}, \end{aligned} \tag{2}$$

where $P(xyz)$, $P(xz)$, and $P(yz)$ are the joint pmfs of the random variables shown in parentheses. The three positions in a triplet are regarded as the random variables X, Z, and Y. The mutual information between the bases at the first and the third position conditioned on the base in the middle is calculated according to Eq. (1) and used as a feature. In addition, for each possible triplet, we computed the quantity $p(x,y,z)\log_2\frac{P(z)P(x,y,z)}{P(x,z)P(y,z)}$. A total of 65 conditional mutual information based features are, therefore, considered.

2.8 Entropy

The Shannon [53] and Gibbs [54] entropies were used to measure the average information content and the thermodynamic stability of the miRNA sequences, respectively. In [55] and [56], we used these entropy measures to quantify digital information content and thermodynamic stability of bacterial genomes. The Shannon entropy for a block size of N is defined as

$$H_N = -\sum_i P_S^N(x_i)\log_2 P_S^{(N)}(x_i) \tag{3}$$

$P_S^N(x_i)$ is the probability of the $i$thword of block size N. Likewise, the Gibbs entropy is defined as

$$S_G = -k_B\sum_i P_G^N(x_i)\ \ln P_G^{(N)}(x_i) \tag{4}$$

where $P_G^N(x_i)$ is the probability to be in the x_ith state and k_B is the Boltzmann constant (1.38 × 10 ^ − 23 J/K). Gibbs' entropy is similar to Shannon's entropy except for the Boltzmann constant (k_B = 1.38 × 10 ^ − 23 J/K). Nevertheless, unlike the Shannon case, where the probability is defined according to the frequency of occurrence, we associated the probability distribution with the thermodynamic stability quantified by the nearest-neighbor free energy parameters. The probability distribution, $P_G^{(N)}$, was modeled by the Boltzmann distribution [57], which provides a functional relationship between energy and temperature

$$P_G^{(N)}(x_i) = \frac{n_{x_i}e^{\frac{-E(x_i)}{k_BT}}}{\sum_j n_{x_j}\ e^{\frac{-E\left(x_j\right)}{k_BT}}}. \tag{5}$$

T is the temperature in Kelvin, n_{x_i} the frequency, and $E(x_i)$ the energy of the ith word of block size N. We used SantaLucia's unified free energy parameters for di-nucleotide steps at 37°C [58]. For block sizes greater than two, the energies were computed by adding the involved di-nucleotides. Shannon and Gibbs entropies for block size of 2 and 3 were calculated and used as features.

2.9 Kullback-Leibler divergence

The Kullback-Leibler divergence or distance (DKL) [59] is a quantitative measure of how similar a probability distribution $P(x)$ is to a model distribution $Q(x)$:

$$D_{KL} = -\sum_i P(x)\log_2\frac{P(x)}{Q(x)}. \tag{6}$$

The relative frequencies of the nucleotides in the given miRNA sequence, $P(x)$, were compared against a uniform distribution $Q(x)$, i.e., the divergence from a uniform distribution is computed.

2.10 Markov model

Assuming that the sequences in the positive and negative classes were generated by two separate Markov sources, we construct a Markov chain and use the scores of miRNA sequences as Markov features. The training set is subdivided into a subset containing the positive and negative samples. Thereafter, each subset is used to generate a Markov chain of a preselected order m ($MC_+(m)$ and $MC_-(m)$). The transition probabilities of the two Markov chains are empirically estimated using the so-called Lidstone estimator [60]. Let $N_x(\mathbf{v})$ denote the number of times a word $\mathbf{v}$ of length m appears in a sequence $\mathbf{x}$. The probability that the next nucleotide is a, where $a \in \Omega = \{A, C, G, U\}$, conditioned on the context $\mathbf{v} \in \Omega^m$ is

$$p_{\mathbf{v},a} = \frac{N_x(\mathbf{v}a) + \delta}{N_x(\mathbf{v}) + 4\delta}. \tag{7}$$

The parameter δ assigns a pseudo count to unseen symbols. In this work, we experimentally checked and found that better results were obtained using smaller values for δ and consequently set $\delta = 0.001$. After the Markov chains for the positive and negative classes were constructed, they were used to score each miRNA sequence. If we represent the sequence as $b_1 b_2 b_3 \dots b_L$, the score is calculated as

$$Score = \sum_{i=1}^{L-m} p(b_i b_{i+1} \dots b_{i+m}) \log_2 \left(\frac{p(b_{i+m} | b_i b_{i+1} \dots b_{i+m-1})}{p(b_{i+m})} \right). \tag{8}$$

The score gives an indication of how likely the miRNA sequence is generated by the given mth order Markov chain. The scores of the miRNA sequence on the Markov chains $MC_+(m)$ and $MC_-(m)$ were used as features. In a previous work [51], we estimated the Markov orders from the training set. However, due to the very short length of the miRNA sequences, the results of order estimation were too poor. Hence, to capture both short and relatively longer dependencies, we decided to select two Markov orders. A combination of orders 1 and either 4 or 5 (i.e., $m = 1, 4$) were found to give better results. Thus, we used four Markov features obtained from scoring the miRNA sequences with the Markov chains $MC_+(1)$, $MC_-(1)$, $MC_+(4)$, and $MC_-(4)$.

2.11 Feature vector and feature selection

For feature selection on a per experiment basis, we have considered the information gain measurement [61] implemented in KNIME (version 3.1.2) [62]. We defined four feature sets, one consists of sequence motifs combined with k-mers (284 features) of which 100 features with highest information gain were used during model training, the second is a combination of IT features with k-mers (175 features), the third comprises of IT features (91 features), and the last considered only k-mers (84 features). Previously [18], it was shown that 50 features might be enough to establish successful models, but we chose to be more conservative here and used 100 features.

2.12 Classification approach

Following the study of [31], we used the random forest (RF) classifiers implemented by the platform KNIME [62]. Classifiers were trained and tested with a split into 80% training and 20% testing data. Negative and positive examples were forced to equal amounts while performing a 100-fold Monte Carlo cross-validation (MCCV) [63] for model establishment.

2.13 Performance evaluation

For each established model, we calculated a number of statistical measures like the Matthews's correlation coefficient (MCC) [64], sensitivity, specificity, and accuracy for evaluation of model performance. The following formulations were used to calculate the statistics (with TP true positive, FP false positive, TN true negative, and FN referring to false negative classifications):

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

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

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

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

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

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

All reported performance measures refer to the average of 100-fold MCCVs.

3 Results and discussion

We have previously shown that sequence motifs and k-mers together (k-mers + motifs) can be used to categorize pre-miRNAs into their species of origin using a machine learning approach [31, 49]. Here, we wanted to test whether IT features and k-mers alone (or their combination k-mers + IT) would be able to achieve better or equal performance. Therefore, we trained a number of classifiers using a 100-fold MCCV with the data split into 80% training and 20% testing ensuring equal shares of positive and negative examples. Random forest is a successful machine learning methodology and was used for setting up all models. In Table 1, we present the performance of machine learning models trained with Hominidae pre-miRNAs as the positive class versus pre-miRNAs from a variety of other groups as negative

Table 1 Average performance of models trained to classify into Hominidae or one of the listed clades

Hominidae	***K*-mers and motifs**			**IT**			***K*-mers only**			***K*-mers and IT**		
vs.	**F-measure**	**ACC**	**MCC**	**F-measure**	**ACC**	**MCC**	**F-measure**	**ACC**	**MCC**	**F-measure**	**ACC**	**MCC**
Hexapoda	0.932	0.931	0.862	0.914	0.914	0.828	0.926	0.926	0.853	0.934	0.934	0.868
Brassicaceae	0.915	0.915	0.830	0.929	0.928	0.857	0.925	0.925	0.850	0.941	0.941	0.883
Monocotyle	0.905	0.903	0.807	0.900	0.897	0.795	0.896	0.898	0.965	0.918	0.917	0.835
Nematoda	0.907	0.906	0.813	0.889	0.889	0.779	0.900	0.901	0.803	0.916	0.916	0.832
Fabaceae	0.864	0.864	0.728	0.892	0.891	0.782	0.894	0.893	0.787	0.907	0.907	0.814
Pisces	0.875	0.877	0.755	0.847	0.851	0.702	0.880	0.877	0.755	0.882	0.885	0.771
Virus	0.823	0.823	0.648	0.827	0.832	0.665	0.821	0.819	0.639	0.846	0.846	0.692
Aves	0.736	0.735	0.472	0.715	0.711	0.424	0.722	0.726	0.454	0.738	0.736	0.473
Laur-asiatheria	0.746	0.731	0.466	0.741	0.720	0.448	0.737	0.729	0.459	0.754	0.736	0.479
Rodentia	0.713	0.722	0.446	0.693	0.697	0.395	0.723	0.719	0.440	0.705	0.710	0.421
Homo sapiens	0.576	0.596	0.190	0.602	0.604	0.208	0.608	0.602	0.205	0.616	0.610	0.221
Cercopithecidae	0.520	0.508	0.017	0.511	0.502	0.004	0.489	0.499	0.000	0.505	0.495	-0.007
Average	0.793	0.793	0.586	0.788	0.786	0.574	0.793	0.793	0.601	0.805	0.803	0.607

The table shows a comparison between IT and *k*-mers + sequence motif features. For the *k*-mers + motif, the best 100 features were selected based on IG (information gain). For the IT, the whole set of features was used. The training/testing was performed with 80%/20% training-testing split, and the average results of 100-fold MCCVs are presented. The table is sorted according to average model accuracy. Note that for the test Hominidae versus H. sapiens, the H. sapiens examples were removed from Hominidae. Highest performance per statistics is highlighted in gray

classes. Similarly to [31], classification between Hominidae and Hexapoda was very accurate (~ 0.93 average accuracy) while classification into Hominidae and Cercopithecidae was impossible (~ 0.50 average accuracy) which is likely due to the very close evolutionary relationship. Moreover, compared to *k*-mers + motifs (0.793 on average), the average accuracy of IT (0.786) is almost as equal whereas *k*-mers + IT (0.803) perform slightly better. The difference between highest and lowest accuracies among feature sets is quite similar. *K*-mers + motifs (0.423), IT (0.426), *k*-mers (0.427), and *k*-mers + IT (0.500). The latter range is the largest which we interpret as being best suited for discriminating between species. The distribution of accuracies for categorization into different clades is similar and with increasing phylogenetic distance the average model accuracy also increased for all feature sets, in general (Table 1). Due to the smaller evolutionary distance, a classification between *Homo sapiens* and Hominiae (*without Homo sapiens*) yielded a relatively low accuracy.

Interestingly, *k*-mers only and IT features achieved a similar performance on average and the results were close to that of IT + *k*-mers as well as sequence motifs + *k*-mers. This is very promising observation which could mean that one can rely only on IT and/or *k*-mer features thereby dropping the computationally expensive generation of sequence motifs and in turn simplifying the process of miRNA categorization. *K*-mers can be calculated in O(n), but motif discovery is NP complete [65] and it is likely that IT features are probably at most O(n^2). A classification of Hominidae as the positive class and the combination of all the other data as the negative class using *k*-mers lead to an accuracy of 0.751 which is close to the average accuracy (0.793; Table 1).

To assess the contribution of IT and *k*-mer features, we performed a feature selection experiment on the combined *k*-mer + IT feature set selecting the top ranked features using information gain (Fig. 1). Among the top 100 features, IT features constitute between 22 and 45%, depending on the groups used to establish the categorization model. However, the contribution of IT features to selected features is high for a smaller number of features (Fig. 1), i.e., they are ranked higher. On average, 76% of the top 2 features are IT. Regardless of the groups used to establish a model for categorization, feature selection considers similar amounts of IT features among the top features (Fig. 1). However, only a small amount of IT features are initially selected (most notably Markov features; Additional file 1: Table S1) while other IT features are included later during feature selection (e.g., MI_CG and Shannon; Additional file 1: Table S1). Additionally, we realized that since miRNAs can stem from any part of a genome, with a significant

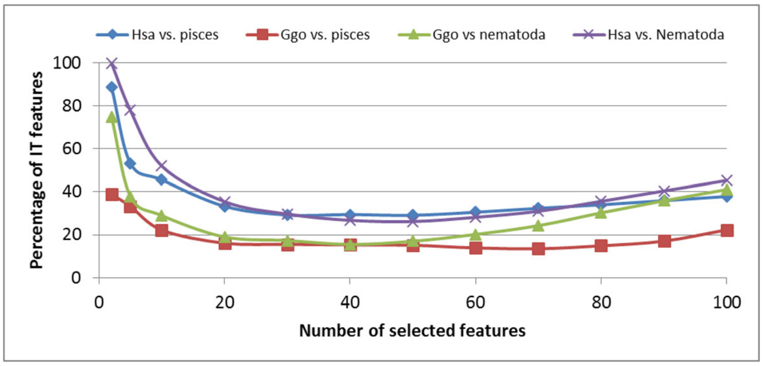

Fig. 1 Percentage of IT features within the top selected features (IT and *k*-mers, no motifs). Ranked using information gain

amount harbored in exons [28], that codon bias [66] would be able to explain how *k*-mer features attain their distinction power between species (*k*-mers were co-dominating the top 10).

Table 1 considers categorization into Hominidae and other groups but apart from *Homo sapiens*, individual species were not considered. The aim, however, is to categorize pre-miRNAs into their species of origin. Therefore, *Homo sapiens* and *Gorilla gorilla* were used as positive data to create models with negative data coming from species from the Pisces class or Nematoda genus (Tables 2 and 3). *G. gorilla* and *H. sapiens*, both in the Hominidae group, have a sufficient amount of pre-miRNA examples to establish a model and are very closely related so that they should show similar average model accuracies when trained against the same set of species. Model establishment, training, and testing were performed as before and the same four feature sets were considered (Tables 2 and 3).

Similar average accuracy over all species was obtained with all four feature sets. However, with the exception of *k*-mers + motifs for ggo, the results were less than the accuracy measured for the Nematoda versus Hominidae experiment (Table 1). To mention one example, the achieved accuracy using *k*-mers + motifs is 0.906 for Hominidae versus Nematoda whereas the average accuracies for categorizing *Homo sapiens* (hsa) and *Gorilla* (ggo) using models trained on the Nematoda species are

Table 2 Average accuracy (ACC) for 100-fold MCCV model training using Homo sapiens (hsa) or Gorilla gorilla (ggo) as the target class and species nematoda as the other class (sorted by accuracy of *k*-mers and motifs for hsa). Species are abbreviated according to miRBase and the expansions are available in our Section 2

	Nematoda							
	hsa (average accuracy)				ggo (average accuracy)			
Species	*K*-mers and motifs	IT	*K*-mers only	*K*-mers and IT	*K*-mers and motifs	IT	*K*-mers only	*K*-mers and IT
cbn	0.938	0.871	0.935	0.914	0.966	0.945	0.941	0.957
prd	0.935	0.896	0.924	0.909	0.975	0.943	0.925	0.950
ppc	0.927	0.889	0.932	0.925	0.938	0.900	0.922	0.916
str	0.912	0.881	0.894	0.894	0.950	0.932	0.905	0.933
crm	0.887	0.856	0.899	0.895	0.881	0.834	0.890	0.877
hco	0.880	0.810	0.870	0.860	0.871	0.798	0.856	0.838
cbr	0.877	0.837	0.874	0.870	0.886	0.851	0.865	0.869
asu	0.865	0.855	0.867	0.864	0.890	0.824	0.886	0.871
cel	0.863	0.847	0.867	0.867	0.938	0.842	0.876	0.871
bma	0.829	0.830	0.843	0.857	0.827	0.793	0.824	0.824
Average	0.891	0.857	0.890	0.885	0.912	0.866	0.889	0.890

Table 3 Average accuracy (ACC) for 100-fold MCCV model training using Homo sapiens (hsa) or Gorilla gorilla (ggo) as target class and species from pisces as other class (sorted by *k*-mers and motif result for hsa). Species are abbreviated according to miRBase, and the expansions are available in our Section 2

	Pisces							
	hsa (average accuracy)				ggo (average accuracy)			
Species	*K*-mers and motifs	IT	*K*-mers only	*K*-mers and IT	*K*-mers and motifs	IT	*K*-mers only	*K*-mers and IT
hhi	0.931	0.894	0.793	0.919	0.972	0.936	0.940	0.949
dre	0.893	0.835	0.874	0.874	0.965	0.936	0.907	0.940
tni	0.838	0.846	0.632	0.833	0.826	0.643	0.820	0.793
pol	0.815	0.738	0.756	0.780	0.943	0.843	0.826	0.865
ccr	0.809	0.792	0.919	0.793	0.801	0.671	0.803	0.773
ipu	0.784	0.743	0.780	0.756	0.821	0.743	0.805	0.797
ola	0.778	0.756	0.765	0.752	0.833	0.740	0.804	0.752
ssa	0.763	0.762	0.752	0.765	0.798	0.684	0.782	0.755
fru	0.687	0.649	0.833	0.632	0.734	0.624	0.673	0.675
Average	0.811	0.780	0.789	0.789	0.855	0.758	0.818	0.811

0.891 and 0.912, respectively. The generated models mostly agree on the order of species based on the sorted accuracies, which can be seen from the highlighted highest performance per feature set (Table 2). As expected, the performance of the hsa and ggo trained models are comparable. In summary, the distinction into Hominidae and Nematoda can be performed with a very high accuracy and the categorization into specific species is also satisfactory. Nematoda are evolutionary distant to Hominidae and may, therefore, perform particularly well. Fish are evolutionary closer and were tested in the same manner (Table 3).

As observed for Nematoda, the average accuracy of discriminating the individual species from the Pisces clade against hsa and ggo (Table 3) is lower than that of classifying between Pisces and Hominidae (Table 1). For example, *k*-mers and motifs averaged over species achieve an accuracy of 0.811 while the model Hominidae versus Pisces achieved 0.877, hence, about 0.06 points more accurate. The Pisces species yielded a similar result in both hsa and ggo models (see highest performing model per feature set highlighted in gray; Table 3). This confirms the hypothesis that species from one group should lead to similar models when trained against species from a different group. A notable outlier is the *k*-mer feature set for hsa which lead to a different sorting of species compared to all other models (Table 3).

Comparing Table 1 with Tables 2 and 3 leads to the observation that categorizing based on the combined species, e.g., for Nematoda (0.906 average accuracy) is more successful than using individual species (average for Nematoda species 0.891). The same holds true for all feature sets and Pisces as well. Some of the species contained only few hairpins, so we wanted to investigate the impact of the amount of available hairpins on the categorization performance. Figure 2 shows that species/clades with more example hairpins tend to perform better. Drosophila (1351 hairpins) and Hexapoda (2014 hairpins) have a large number of hairpins compared to some individual species (lmi 7, dme 256) and consequently produce models which perform well. This analysis further shows that about 100 hairpin examples are needed to establish a successful model and that performance increases when models are created based on the genus rather than individual species (e.g., averaged drosophila species 0.85, genus drosophila, 0.94). This confirms the findings supported by Tables 1, 2, and 3. The Hexapoda model, however, is performing worse than the Drosophila and Nasonia models, which we attribute to an increasing share of non-miRNAs among the hairpin examples. It has been shown before that entries in miRBase are questionable and the chance of incorporating a large share of low-quality hairpins increases with the number of hairpins available when using MCCV. Additional files 1 and 2 contain all the results of this study.

4 Conclusions

Machine learning is important for pre-miRNA detection, but negative data is of an unknown quality [5], which

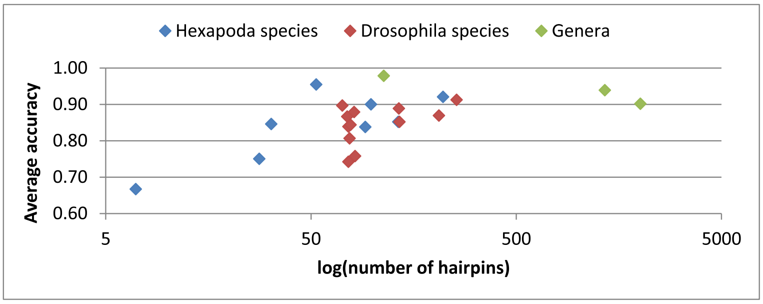

Fig. 2 Number of hairpins (log) available for training a model and the resulting average accuracy. All species derive from Hexapoda (blue and red). Drosophila species are colored in red but also belong to the Hexapoda clade. Three clades (green), Drosophila, Nasonia, and Hexapoda, were combined, and the average model accuracy determined

highlights the need for models that do not depend on negative data. The usage of an arbitrary negative data can be avoided by utilizing the one-class classification approach, or as pointed out in [29, 22], by using known pre-miRNAs of other species. Detection of pre-miRNAs in next generation sequencing data or directly from a genome is the main aim of the field, and many approaches have been developed. For this, hundreds of features have been described and the success rate for pre-miRNA detection is high. With this study, we confirmed that it is possible to distinguish between miRNAs from different species using sequence motifs, IT features, and *k*-mers. Precious studies have also classified miRNAs, for example, into families [67], but it has been shown that miRNAs can evolve rapidly [36, 37, 66], which is detrimental for their categorization into families. However, categorization based on miRNA sequence features (motifs, *k*-mers, and IT) may be possible due to the rapid evolution. Especially, miRNAs originating from exons can be good discriminators when using 3-mer features as they describe codon bias very well (6 3-mer features in top 10; Additional file 1: Table S1). Additionally, we found that *k*-mers are performing almost on a par with the combination of *k*-mer and other features (Tables 1, 2, and 3). Due to the low complexity of calculating *k*-mers and their discriminative power, we suggest that future attempts at categorizing miRNAs into their species can be based on *k*-mer features only. By using both species data directly (Tables 2 and 3) and groups (Table 1), it became clear that models trained are showing consistent performance and are clearly biased by evolutionary distance. The trained models can reliably distinguish between distantly related species. However, the accuracy of the classifiers reduces when the species are closely related (Table 1). To solve this problem, in the future, models will be created for each pair of species and groups in miRBase. Then, for a new example, a distance vector can be determined using the confidence levels of all established models. This confidence vector can be used to categorize any new example to their species of origin or very close to it (e.g., genus). Such a system offers an independent line of evidence for pre-miRNA detection. In addition, this supports pre-miRNA detection from genomes and even more so from next generation sequencing data which is often contaminated [45].

5 Additional files

Additional file 1: Table S1. Contains 38 feature selection experiments. (XLSX 141 kb)

Additional file 2: Table S2. Contains all the results in this study. (XLSX 72 kb)

Abbreviations

ACC: Accuracy; FN: False negative; FP: False positive; IT: Information-theoretic; MCC: Matthews correlation coefficient; MCCV: Monte Carlo cross-validation; MEME: Multiple expectation maximization for motif elicitation; miRNA: MicroRNA; RF: Random forest; RISC: RNA-induced silencing complex; TN: True negative; TP: True positive

Acknowledgements

Not applicable.

Funding

The work was supported by the Scientific and Technological Research Council of Turkey (grant number 113E326), Zefat Academic College, and German Research Foundation (DFG).

Availability of data and materials

All of the miRNA data was obtained from www.mirbase.org.

Authors' contributions

MY formulated the idea of using motifs as features and configured them accordingly for the data used in this study. DN and WH contribute in the idea of using IT features. MY created the workflow, and DN and MY run the experiments. JA and MY jointly made strategic decisions for the machine learning approach and data analysis. DL contributed in designing some more experiments. JA and MY wrote the manuscript while DN and WH wrote the IT section. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]Community Information Systems, Zefat Academic College, 13206 Zefat, Israel. [2]Transmission Systems Group (TrSys), Jacobs University Bremen, Bremen, Germany. [3]Molecular Biology and Genetics, Izmir Institute of Technology, Urla, Izmir, Turkey. [4]Applied Bioinformatics, Wageningen University and Research, Wageningen, the Netherlands.

Received: 17 May 2017 Accepted: 2 October 2017
Published online: 13 October 2017

References

1. EJ Chapman, JC Carrington, Nat Rev Genet **8**, 884 (2007)
2. F Grey, J. Gen. Virol. **96**, 739 (2015)
3. M. Yousef, J. Allmer, and W. Khalifa, Plant microRNA prediction employing sequence motifs achieves high accuracy (2016).
4. MD Saçar, J Allmer, J Pakistan, Clin. Biomed. Res. **1**, 3 (2013)
5. J Allmer, M Yousef, Front. Genet. **3**, 209 (2012)
6. M Saçar, J Allmer, ed. by M Yousef, J Allmer, *miRNomics MicroRNA Biol. Comput. Anal. SE - 10*, vol 2014 (Humana Press), pp. 177–187
7. M Yousef, M Nebozhyn, H Shatkay, S Kanterakis, LC Showe, MK Showe, Bioinformatics **22**, 1325 (2006)
8. HT Dang, HP Tho, K Satou, BH Tu, *2nd Int. Conf. Bioinforma. Biomed. Eng. iCBBE 2008* (2008), pp. 33–36
9. W Khalifa, M Yousef, MD Sacar Demirci, J Allmer, PeerJ **4**, e2135 (2016)
10. DH Tran, TH Pham, K Satov, TB Ho, *2nd Int. Conf. Bioinforma Biomed. Eng.* (2008), pp. 33–36
11. MD Saçar Demirci, J Allmer, PeerJ **5**, e3131 (2017)
12. S-D Hsu, Y-T Tseng, S Shrestha, Y-L Lin, A Khaleel, C-H Chou, C-F Chu, H-Y Huang, C-M Lin, S-Y Ho, T-Y Jian, F-M Lin, T-H Chang, S-L Weng, K-W Liao, I-E Liao, C-C Liu, H-D Huang, Nucleic Acids Res. **42**, D78 (2014)
13. T Vergoulis, IS Vlachos, P Alexiou, G Georgakilas, M Maragkakis, M Reczko, S Gerangelos, N Koziris, T Dalamagas, AG Hatzigeorgiou, Nucleic Acids Res. **40**, D222 (2012)
14. B Fromm, T Billipp, LE Peck, M Johansen, JE Tarver, BL King, JM Newcomb, LF Sempere, K Flatmark, E Hovig, KJ Peterson, Annu. Rev. Genet. **49**, 213 (2015)
15. A Kozomara, S Griffiths-Jones, Nucleic Acids Res. **39**, D152 (2011)
16. MD Saçar, H Hamzeiy, J Allmer, J. Integr. Bioinform. **10**, 215 (2013)
17. M Duygu, S Demirci, J Allmer, J. Integr. Bioinform. (2017)
18. MD Sacar, J Allmer, *2013 8th Int. Symp. Heal. Informatics Bioinforma* (IEEE, 2013), pp. 1–6
19. EC Lai, P Tomancak, RW Williams, GM Rubin, Genome Biol. **4**, R42 (2003)
20. A Sewer, N Paul, P Landgraf, A Aravin, S Pfeffer, MJ Brownstein, T Tuschl, E van Nimwegen, M Zavolan, BMC Bioinformatics **6**, 267 (2005)
21. M Yousef, J Allmer, W Khalifa, J. Intell. Learn. Syst. Appl. **08**, 9 (2016)
22. M Yousef, J Allmer, W Khalifa, J. Biomed. Sci. Eng. **08**, 684 (2015)
23. J Ding, S Zhou, J Guan, BMC Bioinformatics **11**, S11 (2010)
24. D Song, Y Yang, B Yu, B Zheng, Z Deng, B-L Lu, X Chen, T Jiang, BMC Bioinformatics **10**(Suppl 1), S36 (2009)
25. Y. Xu, X. Zhou, and W. Zhang, 24, i50 (2008).
26. KLS Ng, SK Mishra, Bioinformatics **23**, 1321 (2007)
27. Z Gao, X Luo, T Shi, B Cai, Z Zhang, Z Cheng, W Zhuang, Mol. Cells **34**, 239 (2012)
28. VN Kim, J Han, MC Siomi, Nat. Rev. Mol. Cell Biol. **10**, 126 (2009)
29. M Yousef, S Jung, LC Showe, MK Showe, Algorithms Mol. Biol. **3**, 2 (2008)
30. M Yousef, J Allmer, W Khalifa, *Proc. 9th Int. Jt. Conf. Biomed. Eng. Syst. Technol* (Rome, 2016), pp. 219–225
31. M Yousef, W Khalifa, IE Acar, J Allmer, BMC Bioinformatics **18**, 170 (2017)
32. R Batuwita, V Palade, Bioinformatics **25**, 989 (2009)
33. I. D. O. N. Lopes, A. Schliep, A. C. P. de L. F. de Carvalho, I. de On Lopes, A. Schliep, A. C. de Lf de Carvalho, I. D. O. N. Lopes, A. C. P. de L. F. de Carvalho, A. Schliep, and A. C. P. de L. F. de Carvalho, BMC Bioinformatics 15, 124 (2014).
34. W Ritchie, D Gao, JEJ Rasko, Bioinformatics **28**, 1058 (2012)
35. J Chen, X Wang, B Liu, Sci Rep **6**, 19062 (2016)
36. H Liang, W-H Li, Mol. Biol. Evol. **26**, 1195 (2009)
37. J Lu, Y Shen, Q Wu, S Kumar, B He, S Shi, RW Carthew, SM Wang, C-I Wu, Nat. Genet. **40**, 351 (2008)
38. N Fahlgren, S Jogdeo, KD Kasschau, CM Sullivan, EJ Chapman, S Laubinger, LM Smith, M Dasenko, SA Givan, D Weigel, JC Carrington, Plant Cell Online **22**, 1074 (2010)
39. J-H Teune, G Steger, J. Nucleic Acids 2010, 2010
40. Y Wu, B Wei, H Liu, T Li, S Rayner, BMC Bioinformatics **12**, 107 (2011)
41. D Gerlach, EV Kriventseva, N Rahman, CE Vejnar, EM Zdobnov, Nucleic Acids Res. **37**, D111 (2009)
42. C Xue, F Li, T He, G-P Liu, Y Li, X Zhang, BMC Bioinformatics **6**, 310 (2005)
43. P Jiang, H Wu, W Wang, W Ma, X Sun, Z Lu, Nucleic Acids Res. **35**, W339 (2007)
44. A van der Burgt, MWJE Fiers, J-P Nap, RCHJ van Ham, BMC Genomics **10**, 204 (2009)
45. C Bağcı, J Allmer, PLoS One **11**, e0145065 (2016)
46. S Griffiths-Jones, RJ Grocock, S van Dongen, A Bateman, AJ Enright, Nucleic Acids Res. **34**, D140 (2006)
47. RC Edgar, Bioinformatics **26**, 2460 (2010)
48. MV Cakir, J Allmer, *Heal. Informatics Bioinforma. (HIBIT), 2010 5th Int. Symp* (IEEE, Ankara, Turkey, 2010), pp. 31–38
49. M. Yousef, W. Khalifa, I. E. Acar, and J. Allmer, in Proc. BIOSTEC 2017, 10th Int. Jt. Conf. Biomed. Eng. Syst. Technol. (2017).
50. TL Bailey, M Boden, FA Buske, M Frith, CE Grant, L Clementi, J Ren, WW Li, WS Noble, Nucleic Acids Res. **37**, W202 (2009)
51. D Nigatu, W Henkel, *8th Int. Conf. Bioinforma. Model. Methods Algorithms* (2017), pp. 81–92
52. TM Cover, JA Thomas, *Elements of Information Theory*, 2nd edn. (Wiley, 2006)
53. CE Shannon, Bell Syst. Tech. J. **27**, 379 (1948)
54. F Reif, *Fundamentals of Statistical and Thermal Physics*, 56946th edn. (Waveland Pr Inc, 2008)
55. D Nigatu, A Mahmood, W Henkel, P Sobetzko, G Muskhelishvili, IEEE Glob. Conf. Signal Inf. Process. **1338**(2014) (2014)
56. D Nigatu, W Henkel, P Sobetzko, G Muskhelishvili, EURASIP J. Bioinforma. Syst. Biol. **2016**, 4 (2016)
57. J Kovac, J. Chem. Educ. **79**, 1322 (2002)
58. J SantaLucia, Proc. Natl. Acad. Sci. **95**, 1460 (1998)
59. S Kullback, RA Leibler, Ann. Math. Stat. **22**, 79 (1951)
60. GJ Lindstone, Trans. Fac. Actuar. **8**, 182 (1920)
61. NAN Shaltout, M El-Hefnawi, A Rafea, A Moustafa, *Proc. World Congr. Eng* (Newswood Limited, 2014), pp. 625–631
62. MR Berthold, N Cebron, F Dill, TR Gabriel, T Kötter, T Meinl, P Ohl, C Sieb, K Thiel, B Wiswedel, *SIGKDD Explor* (2008), pp. 319–326
63. Q-S Xu, Y-Z Liang, Chemom. Intell. Lab. Syst. **56**, 1 (2001)
64. BW Matthews, BBA - Protein Struct. **405**, 442 (1975)
65. U Keich, PA Pevzner, Bioinformatics **18**, 1374 (2002)
66. M Burset, R Guigó, Genomics **34**, 353 (1996)
67. J Ding, S Zhou, J Guan, BMC Bioinformatics **12**, 216 (2011)