

# Using Base Pairing Probabilities for MiRNA Recognition

Daniel Pasailă\*, Irina Mohorianu, Liviu Ciortuz\*  
Department of Computer Science, “Al. I. Cuza” University  
Iași, Romania  
{daniel.pasaila, irina.mohorianu, ciortuz}@info.uaic.ro

**Abstract**—We designed a new SVM for microRNA identification, whose novelty consist in the fact that many of its features incorporate the base-pairing probabilities provided by McCaskill’s algorithm. Comparisons with other SVMs for microRNA identification prove that our SVM obtains competitive results. One of the advantages of our approach is that it makes no use of so-called normalised features which are based on sequence shuffling, which is a sensitive issue from the biological point of view. This also makes our approach much less time consuming.

## I. INTRODUCTION

MicroRNAs (miRNAs) are short RNA molecules that play important gene regulatory roles. It is well known that most miRNA precursors (pre-miRNAs) fold as hairpins. However, many other RNA sequences in different genomes have a similar structure. Several methods have been proposed for miRNA recognition, among which support vector machines (SVMs) are the best. Most of these SVMs rely on the accuracy of RNA secondary structure prediction programs. We will describe another approach, also using SVM, in which most features are computed using the base-pair binding probabilities provided by McCaskill’s algorithm [21], based on thermodynamics principles. Such an approach seems promising because it does not rely on a single, predicted secondary structure. We prove this claim through direct comparisons with two other SVMs, namely Triplet-SVM [32] and miPred [26], the last of which has reported best results for pre-miRNA identification up to our knowledge.

The plan of this paper is as follows: Section 2 presents the biological background of the miRNA identification problem. Section 3 introduces the reader to existing work in the area of identifying new pre-miRNAs using machine learning techniques, especially support vector machines. Section 4 defines the features that we will use for building a new SVM, while section 5 will give the main results we obtained on different test datasets, and compare them with (some of) the best results available in the literature. Section 6 reports the results that we obtained when trying to find out whether another classifier, Random Forests, is capable of delivering better results than SVM when using the features presented in section 4. Section 7 draws the conclusions of our work and sketches some improvements that we plan to do in the coming future.

\*Joint first authors.

## II. BACKGROUND

MicroRNAs (miRNAs) are non-coding RNA molecules that regulate gene expression at post-transcriptional level. First, miRNAs are transcribed from DNA as *primary miRNAs*. Then the Microprocessor complex, containing the nuclease Drosha, interacting with a primary miRNAs cuts it down to a short hairpin, or stem-loop structure, that is called *precursor miRNA* (pre-miRNA), and has 70–100 nucleotides. Later, pre-miRNAs are processed to *mature miRNAs* (21-23 nucleotides) in the cytoplasm, by interaction with the Dicer enzyme. Figure 1 illustrates the structure of human precursory miRNA *mir-16*. It has been proved to be deleted or downregulated in more than two thirds of cases of chronic lymphocytic leukemia.

What led to miRNA discovery? In the early 1990s, plant scientists were trying to alter flower colours in petunias. Researchers introduced additional copies of a gene for a key enzyme responsible for flower pigmentation (chalcone synthase), thus aiming to obtain darker pink or violet petunias. Surprisingly, less pigmented, partially or fully white flowers were produced [24]. This indicated that the genes (both endogenes and transgenes) responsible for coding that enzyme were downregulated in the altered flowers, but no further explanation could be provided. Several years later, Andrew Z. Fire and Craig C. Mello published a paper in Nature [11], showing how a gene silencing effect can be obtained by injecting short fragments of double stranded RNA into a model organism, *C. elegans*. This gene silencing mechanism was named RNA-mediated interference, or simply RNA interference (RNAi). It easily explains the un-colouring effect on petunias in the above reported experiment: certain short RNAs (for instance miRNAs) produced by the plant itself suppressed the genes responsible for flower pigmentation, by interacting with the messenger RNA produced by these genes. It is now known that RNAi happens in many organisms. The current version (11.0) of miRBase [14], the database that registers all known miRNAs, contains 6396 pre-miRNAs and 6211 mature miRNAs from many species.

What makes the discovery of miRNAs very interesting and useful is that laboratory-made miRNAs can be injected into cells, thus triggering gene suppression, and therefore enabling inferences on targeted gene functions. This opens a new, very promising way for research in disease treatment and drug

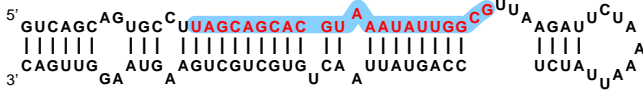


Fig. 1. The stem-loop structure of human precursory miRNA mir-16. The mature miRNA is shaded.

design [10]. For their discovery, Fire and Mello were awarded the Nobel Prize in Physiology and Medicine in 2006.

Bioinformatic methods can be successfully used for the identification of new microRNA genes in genomes. The miRNA identification problem is usually defined over pre-miRNAs because, since their length is larger than that of mature miRNAs, and therefore more information can be extracted from their sequences. Because pre-miRNAs usually have a stem-loop structure, but many other RNA sequences in different genomes have a similar structure, the real challenge is to differentiate real pre-miRNAs from other hairpin-shaped RNA sequences, which are usually called pseudo pre-miRNAs.

### III. RELATED WORK

The first bioinformatic attempts to miRNA identification used sequence alignment systems like BLASTN [2]. Because miRNAs often have non-conserved sequences, and instead they tend to conserve their secondary structure, this approach is not very promising, therefore the focus turned on using machine learning techniques, with a clear preference toward support vector machines, a powerful classification tool [8] [9].<sup>1</sup> For classification using SVM, a feature vector is extracted from the sequence. The selected features are usually statistical, structural, topological and thermodynamical. An RNA secondary structure prediction program, for instance *RNAfold* from the RNA Vienna package [17], is used and then many features are computed using the model predicted by this program. As stated in [22], this approach is limited by the secondary structure prediction accuracy. Therefore, relying on a probabilistic model is expected to be better than building features based on a single predicted structure. In this work we will follow this lead.

In the remaining part of this section we will briefly review the SVMs that have been created up to day for miRNA identification, and then starting with the next section we will develop our approach.

Since 2005 an impressive number of SVM-systems were built, aiming to get better and better results in recognizing miRNAs. The first two of these systems, *miR-abela* [29]

<sup>1</sup>Some of the precursors of ML-based systems for miRNA identification were: miRScan [20] that worked on the *C. elegans* and *H. sapiens* genomes, miRseeker [19] on *D. melanogaster*, and miRfinder [4] on *A. thaliana* and *O. sativa*.

<sup>2</sup>Examples of non-SVM machine learning systems for miRNA identification are BayesMiRNAfind [33] which is based on the naive Bayes classifier, and proMIR [23] that uses a Hidden Markov Model. [30] uses the *k*-NN clustering algorithm to learn how to distinguish between different categories of non-coding RNAs, while [31] introduces MiRank, a system that uses a ranking algorithm based on random walks, a stochastic process defined on weighted finite state graphs.

and Triplet-SVM [32] proved very inspiring. *MiR-abela's* authors have shown that their SVM-based predictions were really valuable to biologists: it turned out through laboratory work that about 30% of the proposed candidates were real pre-miRNAs. Triplet-SVM was instead remarkable due to its simplicity: the features employed are patterns over words of 3 consecutive nucleotides in the pre-miRNA sequence. These patterns gather informations from the first and secondary structure levels of the sequence.

Two other systems were basically derived from Triplet-SVM's approach: MiPred [26], and miREncoding [34]. MiPred, added a couple of thermodynamical features (minimum free energy MFE, and the so-called P-value [12]), and then succeeded to get better results by replacing SVM with Random Forests, an ensemble learning technique using decision trees. MiREncoding added several new features and tried to improve SVM's classification performances by using DFL, a feature selection algorithm.

Another SVM, RNAmicro [16], tried to explore similarities provided by multiple alignments of related miRNAs. [15] describes an SVM, called Microprocessor, that identifies the Drosha cutting site in the extended primary miRNA sequence, and then uses informations regarding this site to improve the performance of another SVM in charge with pre-miRNA recognition.

Recently, a new SVM called miPred [26] produced what seems to be the best results up to date, by making extensive use of thermodynamical features.<sup>3</sup> Despite its performances, miPred faces a two-fold criticism: it uses so-called normalised features, which are computed on a large number of shuffled versions of the given pre-miRNAs. This approach is not very welcome by biologists due to its lack of biological meaning. At the same time, working with normalised features is computationally very time consuming.<sup>4</sup> One of our aims when we started this work was to produce results comparable to those of miPred, without using normalised features.

### IV. OUR SVM

We propose a support vector machine built mainly upon features using the base-pair binding probabilities provided by McCaskill's algorithm [21], supplemented with some other, simple features. The first subsection will give the formal definition of base-pairing probabilities as introduced in [12], while the subsequent subsections will present our SVM's features.

#### A. Base-pairing probabilities

Given an RNA sequence,  $p_{ij}$ , the probability that the nucleotides  $i$  and  $j$  form a base-pair is defined as follows:

$$p_{ij} = \sum_{S_\alpha \in \mathcal{S}} P(S_\alpha) \delta_{ij}^\alpha$$

<sup>3</sup>The reader should not confuse the two miRNA identification systems that have very similar names: MiPred, cited above, and miPred.

<sup>4</sup>Supplementary materials published on the web for miPred [26] says that it uses 10,000 shuffled versions for each (real or pseudo) pre-miRNA. It is therefore expected that computing the features for our SVM, when using 100 pivots (see section 4.2) will be around 100 times faster.

where  $\mathcal{S}$  is the set of all possible secondary structures for the given sequence, and  $\delta_{ij}^\alpha$  is 1 if the nucleotides  $i$  and  $j$  form a base-pair in the structure  $S_\alpha$  and 0 otherwise. The probability of the structure  $S_\alpha \in \mathcal{S}$  follows a Boltzmann distribution:

$$P(S_\alpha) = \frac{e^{-MFE_\alpha/(R \cdot T)}}{Z}$$

with

$$\begin{aligned} Z &= \sum_{S_\alpha \in \mathcal{S}} e^{-MFE_\alpha/(R \cdot T)}, \\ R &= 8.31451 \text{ J mol}^{-1} \text{K}^{-1} \text{ (a molar gas constant), and} \\ T &= 310.15 \text{ K (37}^\circ \text{ C)}. \end{aligned}$$

The probabilities  $p_{ij}$  are efficiently computed using McCaskill's algorithm [21].

### B. A base-pairing profile similarity measure, and related features

We used the idea described in [22] for computing a similarity measure for two RNA sequences based on their pattern of base-pairing formation. To compute this similarity score, two steps are needed: first, a base-pair profile is calculated for each of the two sequences, and then the similarity score is obtained using the global alignment algorithm Needleman-Wunsch [25] with a modified match score and without gap penalties.

Given a pre-miRNA sequence, we apply McCaskill's algorithm, and then for every nucleotide  $i$  we compute the probability of  $i$  forming a base pairing upstream, downstream, or not forming a base pairing at all. Thus, we obtain a *profile* for the given sequence, under the form of an  $L \times 3$  matrix as follows:

$$PF[i, 0] = \sum_{j>i} p_{ij}$$

$$PF[i, 1] = \sum_{j<i} p_{ij}$$

$$PF[i, 2] = 1 - PF[i, 0] - PF[i, 1]$$

The global alignment of two computed profiles is calculated using the Needleman-Wunsch algorithm. We use zero gap penalties, and as match score the inner product of the two profile vectors associated to the corresponding positions in the input sequences. Here is the recurrence relation:

$$S[i, j] = \max \begin{cases} S[i-1, j] \\ S[i, j-1] \\ S[i-1, j-1] + \sum_{k=0}^2 PF[i, k] \cdot PF[j, k] \end{cases}$$

The result is the best alignment score of the profiles computed for the given pair of RNA sequences.

Now, we will show how this similarity measure will be used to compute a number of *profile-based features* for our SVM. First, we will construct a set of RNA sequences that we call *pivot sequences*. Then, the alignment scores of a given (training or testing) pre-miRNA with all the pivot sequences will be included in the pre-miRNA's feature vector. We conjecture that the way in which the pre-miRNA base-pairing profiles align to the profiles of pivot sequences can be successfully used as a discriminative factor in classifying

real vs. pseudo pre-miRNAs. In the developing phase of our system, we used pseudo-miRNAs and pre-miRNAs as pivots, but we saw that the prediction accuracy didn't significantly change when we used randomly generated sequences. Also, we noticed that about 50–200 pivot sequences were needed to achieve best performance. The length of the used pivot sequences seemed to affect the result. In practice we noticed that sequences of 45-65 nucleotides were most appropriate.

### C. Local contiguous structure-sequence probabilistic features

The Triplet-SVM [32] classifier used quite successfully a set of 32 local sequence features for pre-miRNA identification. It employed the *RNAfold* function in the Vienna RNA package [17] for the secondary structure prediction. Then features were computed by counting certain patterns on triplets of nucleotides in the given pre-miRNA sequence. We used the patterns proposed there, but instead of relying on the structure predicted by *RNAfold*, we worked with probabilities provided by the McCaskill algorithm.

In the secondary structure of RNAs, each nucleotide is either paired or unpaired. Let  $PNP[i] = PF[i, 2]$  store the probability that base on position  $i$  is unpaired. For any 3 consecutive nucleotides there are  $8 = 2^3$  possible structure patterns: 'ppp', 'pp.', 'p.p.', 'p..', '.pp', '.p.', '..p', and '...'. Here, 'p' denotes a paired nucleotide, and '.' an unpaired one. Further on, if we consider the middle nucleotide ( $A, C, G$  or  $U$ ) in a triplet, there will be  $32 = 8 \times 4$  possible combinations. Given a pre-miRNA, we will compute the probability of every such combination occurring inside the sequence.

First, we compute a two-dimensional matrix  $Pt[2..(L-1), 1..8]$  where  $Pt[i, j]$  stores the probability that the triplet centered of the  $i$ -th nucleotide has the pattern  $j$ . Making an obvious independence assumption,  $Pt(i, j)$  can be easily computed by multiplying the probabilities that correspond to the three positions inside that pattern. For example, the probability computed for the pattern 'p.p' for some  $i$  is  $(1 - PNP[i-1]) \cdot PNP[i] \cdot (1 - PNP[i+1])$ .

After having computed the matrix  $Pt$ , it is easy to calculate the two-dimensional matrix  $Pn[1..L, 1..8]$  where  $Pn[a, j]$  denotes the probability that nucleotide  $a$  appears in the middle position of occurrences of pattern  $j$ . For this, the following formula is used:

$$Pn[a, j] = \left( \sum_{S[i]=a} Pt[i, j] \right) / (cnt(a)/L)$$

where  $S[1..L]$  is the current sequence and  $cnt(a)$  denotes the number of nucleotides of type  $a$  in the sequence. The  $Pn[a, j]$  values are included in the feature vector we associate to a given pre-miRNA sequence. These 32 features are a natural generalisation to the local contiguous structure-sequence features defined for Triplet-SVM, now using base-pair binding probabilities.

### D. Other features using base-pairing probabilities

For every distinct pair of nucleotides ( $a, a$ ) (12 combinations) we also computed the sum of the base-pair probabilities

for all the corresponding positions in the sequence. We used the following formula:

$$\sum_{S[i]=a, S[j]=b} p_{ij}.$$

The *overall non base-pairing probability* was included in the feature vector. This value is given by:

$$\sum_{i=1}^L PNP[i]/L.$$

We also computed the non base-pairing probability for every nucleotide  $a \in \{A, C, G, U\}$  in the following way:

$$\sum_{S[i]=a} PNP[i]/cnt(a).$$

The output of *mean\_bp\_dist* function in the Vienna RNA package was also used as a feature. This value represents the mean base pair distance in the equilibrium state of a given RNA, which constitutes a measure of the structural diversity. It is also computed using the probabilities obtained with McCaskill’s algorithm.

#### E. Other features

As features not based on McCaskill’s probabilities we first added the folding *minimum free energy*. This was obtained using the *fold* function in the Vienna RNA package, which is based on Zuker’s algorithm [35]. Then, we added the *average frequencies* of  $A, C, G$  and  $U$  in the current sequence, calculated as  $cnt(a)/L$ , for each nucleotide  $a$ . Finally, the *average dinucleotide frequencies* (16 combinations) were also included in the feature vector.

### V. DATASETS AND MAIN RESULTS

The objective of this section is to evaluate the set of features presented in the previous section, by comparing the results it provides when using the SVM classifier with the results reported in the literature for Triplet-SVM and miPred. In order to make as fair comparisons as possible, we first trained our SVM on the same dataset as Triplet-SVM (TR-C, see below), and then retrained our SVM on the training dataset for miPred (TR-H).

As SVM implementation, we used the LibSVM package [7] version 2.84. The penalty parameter  $C$  and the RBF kernel parameter  $\gamma$  were selected using the grid search implemented by a Python script provided with LibSVM. The scaling was performed using the default parameters (-1, 1).

#### A. Comparison with Triplet-SVM

To train the Triplet-SVM classifier [32], its authors built a dataset called TR-C. As positive examples, 163 pre-miRNAs have been randomly selected from the 193 human pre-miRNAs in miRBase version 5.0. As negative examples, 168 pre-miRNA-like hairpins with a similar stem-loop structure to real pre-miRNAs have been randomly selected from CODING, a set of 8494 sequences chosen by Triplet-SVM’s authors from the NCBI RefSeq database [27]. On the TR-C training

TABLE I  
COMPARISON OF OUR SYSTEM WITH TRIPLET-SVM. THE RESULTS FOR TRIPLET-SVM ARE TAKEN FROM [32]. IN PARANTHESIS: THE RATIO OF CORRECTLY CLASSIFIED INSTANCES.

Test	Our accuracy(%)	Triplet-SVM acc.(%)
TE-C: Human pre-miRNAs	<b>96.6</b> (29/30)	93.3
TE-C: Pseudo pre-miRNAs	<b>96.5</b> (965/1000)	88.1
UPDATED	92.3 (36/39)	<b>92.3</b>
CROSS-SPECIES	<b>95.4</b> (554/581)	90.9
CONSERVED-HAIRPIN	<b>93.5</b> (2287/2444)	89.0

dataset, when doing 5-fold cross validation our SVM obtained a prediction accuracy of 96.07% following the grid parameter search.

For the test phase, the authors of Triplet-SVM built four datasets:

- The TE-C dataset included the 30 remaining human pre-miRNAs from miRBase version 5.0, and 1000 pseudo pre-miRNAs randomly selected from the CODING set, excluding those already allocated to the TR-C training set.
- The UPDATED dataset was made of 39 human pre-miRNAs, reported after the release of miRBase 5.0 and up to the time when Triplet-SVM was completed.
- The CROSS-SPECIES dataset consists of 581 pre-miRNAs from 11 species in miRBase 5.0, different from human.
- The CONSERVED-HAIRPIN dataset was built by extracting 2444 hairpins from the human chromosome 19, between positions 56000001 and 57000000, obtained from the UCSC database (hg17, May 2004) [18]. Of all these hairpins, 3 are real pre-miRNAs, while the others are pseudo pre-miRNAs.

Table I shows the results we obtained on the above four test datasets, compared to Triplet-SVM, after both SVMs were trained on the same dataset, TR-C. For the profile, we included 50 pivots, which are randomly generated sequences of 45-65 nucleotides. One can see that our SVM has a better accuracy than Triplet-SVM on all these four datasets. Detailed comparisons on the different species in the CROSS-SPECIES dataset are shown in Table II. These good results encouraged us to do further comparisons, this time with the miPred SVM.

#### B. Comparison with miPred

For miPred [26], the training set (called TR-H) included 200 human pre-miRNAs randomly selected from miRBase 8.2, and 400 pseudo-miRNAs from the CODING set, built by Triplet-SVM’s authors.

In order to test their classifier, the authors of miPred built four datasets: TE-H, IE-NH, IE-NC and IE-M:

- TE-H and IE-NH were designed similarly to the datasets TE-C and respectively CROSS-SPECIES used for testing Triplet-SVM: TE-H included the 123 human pre-miRNAs remaining from miRBase 8.2 after 200 such pre-miRNAs have been allocated for training (TR-H), while IE-NH contains 1918 pre-miRNAs from 40 non-human species from miRBase 8.2.

TABLE II

DETAILED COMPARISON OF OUR SYSTEM WITH TRIPLET-SVM: ACCURACY ON THE CROSS-SPECIES DATASET. THE RESULTS FOR TRIPLET-SVM ARE TAKEN FROM [32]. IN PARANTHESIS: THE RATIO OF CORRECTLY CLASSIFIED INSTANCES.

Test	Our accuracy(%)	Triplet-SVM accuracy(%)
Mus musculusi	<b>97.2</b> (35/36)	94.4
Rattus norvegicus	<b>84.0</b> (21/25)	80.0
Callus Gallus	<b>100.0</b> (13/13)	84.6
Dnio Rerio	<b>83.3</b> (5/6)	66.7
Caenorhabditis briggsae	<b>100.0</b> (73/73)	95.9
Caenorhabditis elegans	<b>92.7</b> (102/110)	86.4
Drosophila pseudoobscura	<b>94.3</b> (67/71)	90.1
Drosophila melanogaster	<b>95.7</b> (68/71)	91.5
Oryza sativa	<b>96.8</b> (93/96)	94.8
Arabidopsis thaliana	<b>97.3</b> (73/75)	92.0
Epstein Barr Virus	80.0 (4/5)	<b>100.0</b>
Total	<b>95.35</b> (554/581)	90.9

Both datasets included twice more negative examples than positives, randomly selected from the CODING set.

– IE-NC consists of 12387 non-coding RNAs (other than miRNAs) from the Rfam 7.0 database [13], and IE-M is made of 31 messenger RNAs selected from GenBank [3].

We recreated these five datasets according to the above specifications made by the authors of miPred, since they did not provide the datasets themselves.

We re-trained our SVM on the TR-H dataset, similarly to miPred, and then we run it on the above four test datasets. Table III shows comparative results with miPred and Triplet-SVM.<sup>5</sup> We used 100 randomly generated pivots. Our SVM not only outperformed again Triplet-SVM on all datasets, but it also definitely outperformed miPred on the IE-NC and IE-M datasets (82.75% vs. 68.68%, and respectively 100% vs. 87.09% accuracy), while on IE-NH it loses 1.53% in accuracy compared to miPred. On IE-NH, our SVM’s accuracy is slightly better than that of miPred. Note that Triplet-SVM misclassifies all 31 instances in the IE-M set, while our SVM correctly classifies them all.

Our conclusion is that our SVM is a very serious contender not only for Triplet-SVM, but also for miPred.

## VI. SEARCHING FOR FURTHER IMPROVEMENTS

The MiPred system [26] got better results by using the Random Forests [6] classifier instead of SVM, with the same features, namely, the Triplet-SVM features plus the folding minimum free energy and the P-value [12]. We wanted to see whether the same is true for our set of features. This section briefly presents Random Forests (RF), and then it reports on the tests we did using RF as classifier for our miRNA identification problem.

<sup>5</sup>Because we re-created the miPred’s train and test datasets, it is quite possible that there are slight differences between the published results and those that would be obtained by running miPred and Triplet-SVM on the re-created datasets.

TABLE III

COMPARISON OF OUR SYSTEM WITH miPred AND TRIPLET-SVM. THE RESULTS FOR miPred AND TRIPLET-SVM ARE TAKEN FROM [26]. ONLY ACCURACY IS GIVEN FOR IE-NC AND IE-M SINCE THESE DATASETS ARE MADE ONLY OF NON miRNAs; IN SUCH A CASE, SPECIFICITY IS EQUAL TO ACCURACY, AND SENSITIVITY IS UNDEFINED.

Test	Our accuracy(%)		miPred accuracy(%)		Triplet-SVM accuracy(%)	
	se.(%)	sp.(%)	se.(%)	sp.(%)	se.(%)	sp.(%)
TE-H	<b>93.77</b>		93.50		87.96	
	<b>87.80</b>	96.74	84.55	<b>97.97</b>	73.15	93.57
IE-NH	94.11		<b>95.64</b>		86.15	
	90.35	95.99	<b>92.08</b>	<b>97.42</b>	86.15	96.27
IE-NC	<b>82.75</b>		68.68		78.37	
IE-M	<b>100</b>		87.09		–	

Random Forests is an *ensemble learning* algorithm that was derived from *bagging*, also introduced by Leo Breiman [5]. Like *boosting* [28] too, these techniques use certain strategies for aggregating some simpler classification algorithms. In the sequel we will consider that the aggregated classifiers are *decision trees*.

The main idea behind *boosting* is the following: decision trees are constructed successively, and each time a new tree is built, the data points that have been incorrectly predicted by earlier trees are given some extra weight. The learner is thus forced to successively concentrate on more and more difficult cases. In the end, the classification of a given instance is decided by a linear (weighted) combination of the votes given by the decision trees.

In the *bagging* approach, whose name comes from *bootstrap aggregating*, new trees do not depend on earlier trees. Each tree is independently constructed using a bootstrap sample (i.e. sampling with replacing) from the training dataset. Classification of a test instance is done by taking a simple majority vote among the decision trees.

The Random Forests algorithm extends bagging with an additional layer of randomness, namely the random feature selection: while in standard decision trees each node is split using the best split among all variables, in RF each node is split using the best among a subset of features randomly chosen at that node. Thus, RF uses only two parameters: the number of variables in the random subset at each node, and the number of trees in the forest.

Although RF is a somehow counter-intuitive strategy, it proved to be robust against overfitting, and it produced some good results when compared to other machine learning techniques including SVMs, neural networks, discriminate analysis, etc. As implementation for RF, we used the *randomForest* (version 4.5-25) package for the R language [1]. Feature selection was done using the *importance* function from the R package, which is based on RF.

Table IV shows the accuracy results we obtained when training RF and respectively an SVM on TR-C, which was the Triplet-SVM’s training set, and did comparisons on its test sets: TE-C, UPDATED, CROSS-SPECIES, and

TABLE IV  
COMPARING THE PREDICTIVE ACCURACY(%) OF RF AND SVM ON TEST DATASETS FROM TRIPLET-SVM, USING OUR FEATURES.

Test	RF		SVM with feature sel.
	without feature sel.	with feature sel.	
TE-C	61.1	93.2	94.4
UPDATED	<b>94.9</b>	89.7	<b>97.4</b>
CROSS-SPECIES	<b>96.1</b>	89.5	89.8
CONSERVED-HAIRPIN	92.6	89.6	91.0

TABLE V  
COMPARING THE PREDICTIVE ACCURACY(%) OF RF AND SVM ON TEST DATASETS FROM MI-PRED, USING OUR FEATURES.

Test	RF		SVM with feature sel.
	without feature sel.	with feature sel.	
TE-H	92.14	92.14	91.86
IE-NH	93.82	92.72	91.87
IE-NC	63.46	63.30	88.31
IE-M	74.19	16.12	100

CONSERVED-HAIRPIN. Both classifiers used the features we described in Section 4. Profile similarities were computed on 50 pivots. RF produced a better results than our SVM described in Section 5 on the UPDATED and CROSS-SPECIES datasets: 94.9% vs. 92.3% and respectively 96.1% vs. 94.5% accuracy (SVM's results are from Table I). However, on the TE-C dataset, RF registered a very serious decrease of accuracy: 61.1% down from 96.5%. Another SVM, that used features selected following the analysis of the decision trees produced by RF on the full set of features, obtained a better result only on the UPDATED dataset: 97.4% vs. 92.3%. We note that feature selection did not improve the RF results on any of the four Triplet-SVM's test datasets.

We also performed a similar comparison between RF and SVM on the test datasets designed by miPred's authors: TE-H, IE-NH, IE-NC, and IE-M, after having the both classifiers trained on miPred's training dataset, TR-H. Profile similarities were computed on 125 pivots. Table V shows that unfortunately RF did not produce better results than our SVM described in Section 4, on any of these test datasets (see SVM's results in Table III). Instead, on the IE-NC and IE-M datasets, RF registered heavy losses of accuracy: 63.46% vs. 82.75%, and respectively 74.19% vs. 100%. The RF-supported feature selection procedure did not help RF (and neither SVM), to get any improvement. On the contrary, the RF classifier retrained after feature selection produced very bad results on the IE-M dataset: 16.12% vs 100%.

Our conclusion is that RF seems not to be a very good candidate to replace SVM for pre-miRNA identification using our set of features presented in Section 4.

## VII. CONCLUSIONS AND FURTHER WORK

We showed that the base pairing probabilities provided by McCaskill's algorithm combined with some other, simple statistical measures make a SVM classifier achieve high pre-miRNA prediction accuracy rates, comparable to the best published results up to our knowledge.

There is one issue that we need to address in the coming future: how to better choose the pivot sequences? Until now we performed several runs with different sets of randomly chosen pivots. Here we reported the results obtained for the set of pivots that produced the best overall accuracy on the Triplet-SVM and respectively the miPred test datasets. On the Triplet-SVM test datasets, our SVM produced no significant differences in the reported accuracy, specificity and selectivity. Unfortunately this did not held when using the miPred test datasets. Even if the situation were not this delicate, one could ask whether our SVM could get better results by "improving" its (set of) pivots. Using genetic programming could be an answer to this question. A second possibility would be to choose "representative" pivots among the training sets. Here clusterization might help, using a distance measure between vectors of features presented in Section 4, excepting features regarding the similarity with pivots. We plan to report soon on this issue.

We will also make direct comparisons with a very recent kNN-based classifier for non-coding RNAs that was documented while we were working on this paper [30]. Its reported results seem quite competitive, due to the use of certain topological features. We will see whether those features could be generalised using again the probabilities computed by McCaskill's algorithm. If so, we will check whether adopting them into the feature set of our SVM would further improve the quality of pre-miRNA prediction.

## ACKNOWLEDGMENT

We thank Ștefan Panțiru and Alina Sîrbu for their work on features used by the systems Triplet-SVM, miR-abela and miPred. LC thanks Mihaela Zavolan for introducing him to the problem of miRNA identification during his visit at Biozentrum, Basel in 2007. LC was partially supported by the CEEEX grant "ForMol" from the Romanian Ministry of Education and Research.

## REFERENCES

- [1] <http://www.r-project.org/>.
- [2] S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.
- [3] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler. GenBank. *Nucleic Acids Research*, 33(Database-Issue):34–38, 2005.
- [4] E. Bonnet, J. Wuyts, P. Rouzé, and Y. Van de Peer. Detection of 91 potential conserved plant microRNAs in *Arabidopsis thaliana* and *Oryza sativa* identifies important target genes. *Proc Natl Acad Sci U S A*, 101(31):11511–11516, 2004.
- [5] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.
- [6] L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.

- [7] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [8] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.
- [9] N. Cristianini and J. Shawe-Taylor. *An introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, New York, NY, USA, 2000.
- [10] G. A. Călin and C. M. Croce. MicroRNA-cancer connection: The beginning of a new tale. *Cancer Res*, 66(15):7390–7394, 2006.
- [11] A. Fire, S. Xu, M. Montgomery, S. Kostas, S. Driver, and C. Mello. Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature*, 391(6669):806–811, 1998.
- [12] E. Freyhult, P. P. Gardner, and V. Moulton. A comparison of RNA folding measures. *BMC Bioinformatics*, 6:241, 2005.
- [13] S. Griffiths-Jones, S. Moxon, M. Marshall, A. Khanna, S. R. Eddy, and A. Bateman. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Research*, 33:D121, 2005.
- [14] S. Griffiths-Jones, H. K. Saini, S. van Dongen, and A. J. Enright. miRBase: tools for microRNA genomics. *Nucleic Acids Research*, 36(Database-Issue):154–158, 2008.
- [15] S. Helvik, O. J. Snøve, and P. Sætrom. Reliable prediction of Droscha processing sites improves microRNA gene prediction. *Bioinformatics*, 23(2):142–149, 2007.
- [16] J. Hertel and P. F. Stadler. Hairpins in a haystack: recognizing microRNA precursors in comparative genomics data. *Bioinformatics*, 22(14):e197–e202, July 2006.
- [17] I. L. Hofacker. The Vienna RNA secondary structure server. *Nucleic Acids Res*, 31(13):3429–3431, 2003.
- [18] D. Karolchik, R. Baertsch, M. Diekhans, T. Furey, A. Hinrichs, Y. Lu, K. Roskin, M. Schwartz, C. Sugnet, D. Thomas, R. Weber, D. Haussler, and W. Kent. The UCSC Genome Browser Database. *Nucleic Acids Res*, 31(1):51–54, 2003.
- [19] E. C. Lai, P. Tomancak, R. W. Williams, and G. M. Rubin. Computational identification of *Drosophila* microRNA genes. *Genome Biology*, 4(7), 2003.
- [20] L. P. Lim, N. C. Lau, E. G. Weinstein, A. Abdelhakim, S. Yekta, M. W. Rhoades, C. B. Burge, and D. P. Bartel. The microRNAs of *Caenorhabditis elegans*. *Genes Dev*, 17(8):991–1008, 2003.
- [21] J. S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structures. *Biopolymers*, 29:1105–1119, 1990.
- [22] L. M. C. Meireles. Evaluation of a kernel function for recognizing microRNAs, 2006. Project report, School of Computer Science, CMU.
- [23] J.-W. Nam, K.-R. Shin, J. Han, Y. Lee, V. N. Kim, and B.-T. Zhang. Human microRNA prediction through a probabilistic co-learning model of sequence and structure. *Nucleic Acids Research*, 33(11):3570–3581, 2005.
- [24] C. Napoli, C. Lemieux, and R. Jorgensen. Introduction of a chimeric chalcone synthase gene into petunia results in reversible co-suppression of homologous genes in trans. *Plant Cell*, 2(4):279–289, 1990.
- [25] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453, 1970.
- [26] K. L. S. Ng and S. Mishra. De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures. *Bioinformatics*, 23(11):1321–1330, 2007.
- [27] K. D. Pruitt and D. R. Maglott. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Research*, 29(1):137–140, 2001.
- [28] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5):1651–1686, 1998.
- [29] A. Sewer, N. Paul, P. Landgraf, A. Aravin, S. Pfeffer, M. J. Brownstein, T. Tuschl, E. van Nimwegen, and M. Zavolan. Identification of clustered microRNAs using an *ab initio* prediction method. *BMC Bioinformatics*, 6:267, 2005.
- [30] W. Shu, X. Bo, Z. Zheng, and S. Wang. A novel representation of RNA secondary structure based on element-contact graphs. *BMC Bioinformatics*, 9(1):188, 2008.
- [31] Y. Xu, X. Zhou, and W. Zhang. MicroRNA prediction with a novel ranking algorithm based on random walks. *Bioinformatics*, 24(13), 2008.
- [32] C. Xue, F. Li, T. He, G. Liu, Y. Li, and X. Zhang. Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics*, 6(310), 2005.
- [33] M. Yousef, M. Nebozhyn, H. Shatkay, S. Kanterakis, L. Showe, and M. Showe. Combining multi-species genomic data for microRNA identification using a Naive Bayes classifier machine learning for identification of microRNA genes. *Bioinformatics*, 22(11):1325–1334, 2006.
- [34] Y. Zheng, W. Hsu, M.-L. Lee, and L. Wong. Exploring essential attributes for detecting microRNA precursors from background sequences. In M. M. Dalkilic, S. Kim, and J. Yang, editors, *2006 VDMB Workshop on Data Mining in Bioinformatics*, volume 4316 of *Lecture Notes in Computer Science*, pages 131–145. Springer, 2006.
- [35] M. Zuker and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, 9(1):133–148, 1981.