

# RFMirTarget: a Random Forest Classifier for Human miRNA Target Gene Prediction

Mariana R. Mendoza<sup>1</sup>, Guilherme C. da Fonseca<sup>2</sup>, Guilherme L. de Morais<sup>2</sup>,  
Ronnie Alves<sup>3</sup>, Ana L. C. Bazzan<sup>1</sup> and Rogerio Margis<sup>2</sup>

<sup>1</sup>PPGC, UFRGS, P.O. Box 15064, Porto Alegre, RS, Brazil  
`{mrmendoza,bazzan}@inf.ufrgs.br`

<sup>2</sup>PPGBCM, UFRGS, P.O. Box 15005, Porto Alegre, RS, Brazil  
`{guicf13,guilherme.loss}@gmail.com,rogerio.margis@ufrgs.br`

<sup>3</sup>Vale Technological Institute Sustainable Development, Belém, PA, Brazil  
`ronnie.alves@vale.com`

**Abstract.** *MicroRNAs (miRNAs) are key regulators of eukaryotic gene expression whose fundamental role has been already identified in many cell pathways. The correct identification of miRNAs targets is a major challenge in bioinformatics. So far, machine learning-based methods for miRNA-target prediction have shown the best results in terms of specificity and sensitivity. However, despite its well-known efficiency in other classifying tasks, the random forest algorithm has not been employed in this problem. Therefore, in this work we present RFMirTarget, an efficient random forest miRNA-target prediction system. Our tool analyzes the alignment between a candidate miRNA-target pair and extracts a set of structural, thermodynamics, alignment and position-based features. Experiments have shown that RFMirTarget achieves a Matthew's correlation coefficient nearly 48% greater than the performance reported for the MultiMiTar, which was trained upon the same data set. In addition, tests performed with RFMirTarget reinforce the importance of the seed region for target prediction accuracy.*

**Keywords:** miRNA, target prediction, random forest, gene regulation

## 1 Introduction

MicroRNAs (miRNAs) are non-coding RNAs of  $\sim 22$  nucleotides in length that act as negative regulators of gene expression, thus playing an important role in gene regulation by targeting mRNAs with cleavage or translational repression [1]. The miRNA biogenesis is similar in both animals and plants. Mature miRNAs are formed from longer primary transcripts by two sequential processing steps mediated by a nuclear and a cytoplasmic RNase III endonuclease. In animals the responsible enzymes are Drosha and Dicer, respectively, while in plants both cleavages are performed by a Dicer homolog, DCL [1]. These cleavages generate a 60–70 nt stem-loop miRNA precursor (pre-miRNAs) and a mature miRNA duplex, respectively. Further, the mature miRNA duplex is assembled into an

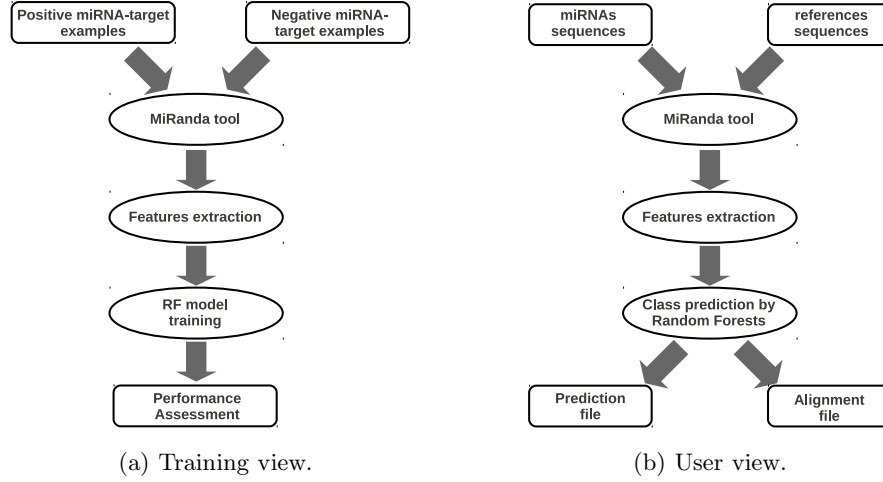
effector complex known as RNA-induced silencing complex (RISC). The gene repression, caused by miRNA silence, can occur in two different ways: i) by degradation of messenger RNA (mRNA) through the RNA interference (RNAi) pathway or ii) by inhibiting protein translation [2].

Since the discovery of miRNAs in *Caenorhabditis elegans* [3], in which they were related to control developmental timing, miRNAs have been characterized in both animals and plants in several metabolic processes, such as growth, apoptosis, cell proliferation, stress responses and defense against viruses and other diseases [4, 5]. In humans, miRNAs play a critical role in tumorigenesis, acting either as tumor suppressors or oncogenes. Moreover, recent studies have shown that miRNAs are highly related with cancer progression, including initiating, growth, apoptosis, invasion and metastasis [6].

Two main challenges are involved in the study of miRNAs: the identification of novel miRNAs and the prediction of miRNAs targets. The existence of a stem-loop is the key feature adopted by *ab initio* prediction methods to identify novel miRNAs. In this sense, machine learning (ML) algorithms, among which support vector machine (SVM) [7, 8], random forest [9] and naïve Bayes [10] stand out, have been extensively applied to the task of learning how to distinguish real pre-miRNAs from pseudo pre-miRNAs based on a set of descriptive features. Following this direction, ML methods can help in the computational prediction of miRNAs targets, although this is considered a more difficult problem.

Basically, the interaction of a miRNA and its target occurs by complementarity of their nucleotide sequences, with some functional differences between plants and animals miRNAs. Plant miRNAs bind their targets with perfect or near perfect complementarity and mostly in their open read frames (ORFs) [11]. In contrast, in most cases, animals miRNAs sequences have a partial complementarity to their targets and the hybridization occurs predominantly in the 3' untranslated regions (3' UTRs) [12]. Furthermore, in animals, a region of miRNA called *seed* plays an important role in the correct interaction between the miRNA and its target. This region comprehends six to eight nucleotides in the 5' end of the miRNA that have (almost) strict pairing with the mRNA target. The classification of miRNA target sites can be thus distinguished into three types: i) 5' dominant canonical, where a perfect match occurs in the seed region and an extensive base pairing is observed in the 3' end of the miRNA, ii) 5' dominant seed only, which presents an extensive base pairing only in the seed region and iii) 3' compensatory, in which the seed region does not contains a perfect base pairing, but the 3' end of the miRNA displays a more extensive base pairing [13, 14].

Several different computational tools have been already developed for target prediction analysis [15]. In general, these tools are based on features derived from the interaction between a miRNA and its potential targets, such as seed complementarity, thermodynamics stability, presence of multiple sites and evolutionary conservation among species [14]. Among these, ML-based algorithms have had the best results so far in terms of specificity and sensitivity [16]. One well-known classifier is the random forest algorithm [17], further explained in Section 2.1.



**Fig. 1.** Fluxogram of RFMirTarget training and functioning. (a) RFMirTarget is trained with positive and negative miRNA-target examples. This data set is analyzed by miRanda, whose output is processed for features extraction. A RF model is then built upon these features. (b) Once trained, RFMirTarget may be used to classify unknown miRNA-targets: given two files with miRNAs and references sequences, RFMirTarget outputs the confidence and alignment for all predicted miRNA-targets.

Its popularity is due mainly to its meaningful efficiency when compared to other classification methods. This efficiency, in turn, comes from the manner the algorithm profits from ensemble predictions. Although already known as an efficient approach for identifying novel miRNAs [9], random forests have not been explored in the context of miRNA-target prediction yet. Therefore, in the present paper, we introduce the RFMirTarget tool, a random forest model for the prediction of miRNAs targets. Tests with biologically validated examples have shown that the proposed model is indeed robust and has performance superior than the state-of-the-art tool MultiMiTar [16].

## 2 Materials and Methods

RFMirTarget is trained upon biologically validated miRNA-target pairs. This data set is processed by miRanda (see Section 2.3) in order to identify interacting sites between miRNAs and their respective targets and prepare the data set for feature extraction. The alignments provided by miRanda are the source for features definition, which in turn are used for training a random forest classifier. Once trained, the model can be applied to the classification of unknown instances of miRNA-targets (Fig. 1(b)). To perform such task, the user must provide two files with miRNAs and candidate reference sequences, and RFMirTarget outputs the prediction confidence and alignment for all predicted miRNA-target

pairs. In what follows we explain the methods involved in the training process, summarized in Fig. 1(a).

## 2.1 Random Forest

Random forest (RF) is a well-known ensemble approach for classification tasks [17], which has its basis on the combination of tree-structured classifiers with the randomness and robustness provided by bagging and random feature selection. In [18], the application of bagging as a means to enhance the performance of tree-structured classifiers and reduce their bias was proposed by Breiman. The author's approach consisted in training several classifiers with random bootstrap samples from the original data set and afterwards combining their results into a single prediction: for classification tasks, by means of voting; for regression tasks, by averaging all classifiers results.

Furthermore, Breiman has improved his previous model by aggregating random feature selection to the training process [17]. His proposal consisted in selecting from a random subset of features the one with the smallest impurity to split at each node when growing a tree. Tests run by Breiman have revealed that RF classifier always outperforms the bagging approach [18]. However, the benefits of RFs go beyond its good performance. The mechanism applied for growing trees allows an unbiased estimation of both the generalization error and the most important variables for classification during the growth process, using for such analysis the data left out of the bootstrap sample used as training set, named out-of-bag (OOB) data. Additionally, as RFs are tree-structure classifiers, they inherit the interpretability associated to this type of model [9]. In the present work, the RF model was implemented with the `randomForest` R package [19].

## 2.2 Data set

The RF model was trained with experimentally verified examples of human miRNA-target collected by Bandyopadhyay and Mitra in [20] and used in the training process of MultiMiTar [16], a SVM-based miRNA-target prediction system. The data set is composed of 289 biologically validated positive examples and 289 systematically identified tissue-specific negative examples. As the basic mechanism of RF renders the definition of training and testing data sets unnecessary, both classifier model and error estimative are drawn from the same data set. During the training process about 2/3 of the original data is sampled for growing the tree, while the remainder (OOB data) is used to test the generated model and estimate the generalization error. We refer the reader to [20] for more details about the data.

## 2.3 Data Preparation

The data set of positive and negative examples of miRNA-target pairs used for training MultiMiTar does not comprises information about the actual site of

```

                20                      1
hsa-let-7a: 3' ttGATATGTTGGATGATGGAGt 5'
                | | ||| |||||
HMGA2: 5' atCAAAACACACTACTACCTCt 3'

```

**Fig. 2.** Example of miRNA-target alignment predicted by miRanda. The miRanda software outputs all possible alignments and base pairings of a miRNA-target pair that scored above a given threshold. The highlighted nucleotides refer to the seed region. In addition, the figure illustrates the nucleotides numbering (1–20) for position-based features extraction.

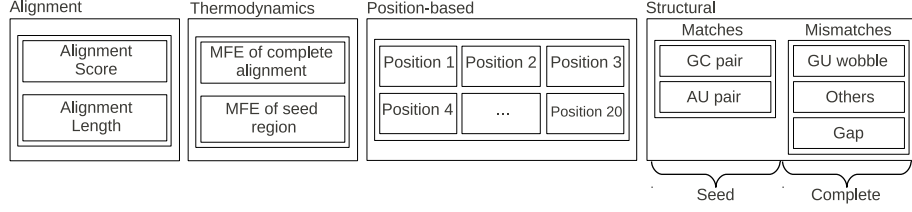
alignment between miRNAs and their targets. Due to the existence of multiple sites, which is related to some extent to the short length of miRNAs sequences, the extraction of this information by techniques such as BLAST could result in an extremely large data set, with many biologically unlikely miRNA:mRNA pairs. Therefore, to reduce the dimension of our problem and prepare our data set for feature extraction, we resort to miRanda software [21] to obtain the miRNA-target binding sites from the same examples used for MultiMiTar training.

The miRanda software runs a score-based algorithm to analyze the complementarity of nucleotides (A:U or G:C) between aligned sequences. The scoring matrix allows the occurrence of the non-canonical base-pairing G=U wobble, which is important for the accurate detection of RNA:RNA duplexes, and is based on the following parameters: +5 for G≡C, +5 for A=U, +2 for G=U and -3 for all other nucleotides pairing [21]. Additionally, there is a scaling factor for giving a higher weight to nucleotides within the first eleven positions.

Besides the scoring matrix, four empirical rules are applied for the identification of the miRNA binding sites, counting from the first position of the 5' end of the miRNA: i) no mismatches at positions 2 to 4; ii) fewer than five mismatches between positions 3-12; iii) at least one mismatch between positions 9 and L-5 (where L is length of the complete alignment); and iv) fewer than two mismatches in the last five positions of the alignment [21]. An example of the alignment output provided by miRanda is depicted in Fig. 2. After running miRanda algorithm on the data set described at Section 2.2, we obtained 1074 positive and 407 negative miRNA-target pairs, which consists of the training instances used for building the RF model with the `randomForest` R package [19]. At this point we emphasize that although the number of training instances we use is greater than the value reported in [16], they derive from the original data set used for training MultiMitar. The difference in the data set dimension is due to data processing by miRanda, which is part of our strategy to fulfill a lack of information on the actual binding site between miRNA-target pairs.

## 2.4 Features

The negative and positive examples predicted by miRanda algorithm consist of the alignment between both sequences, as well as properties such as the score



**Fig. 3.** RFMirTarget combines 34 features divided into four categories: alignment, thermodynamics, position-based and structural features. From these, six refer to the seed region: MFE value and five structural features.

and alignment length, based on which the classifier features are extracted. The RF features are divided into four categories: structural features, thermodynamics features, alignment features and position-based features. The first two categories are widely used for training classifier systems, while the later was introduced in the study held by [22] and the alignment features are proposed in the current study. In what follows we explain each of the defined categories.

- **Structural features.** Quantify the number of matches (G:C and A:U pairing) and mismatches (G:U wobble pair, gap and other mismatches) in the alignment.
- **Thermodynamics feature.** Minimum free energy (MFE) of the miRNA-target alignment computed by RNAfold [23].
- **Alignment features.** Properties of the miRNA-target alignment computed by the miRanda algorithm: alignment score and alignment length.
- **Position-based features.** Evaluation of each basepair from the 5'-most position of the miRNA up to the 20th position of the alignment, assigning nominal values to designate the kind of pairing in each position: a G:C match, an A:U match, a G:U wobble pair, a gap and a mismatch. In the previous study [22], the gap feature was not included in the set of position-based features.

The thermodynamics and structural features were extracted in twofold manner: for the complete alignment and for the seed region, which is composed by the nucleotides in positions 2-8, to count from the 5'-most position of the miRNA (Fig. 2). Following the approach in [22], the MFE for both seed and complete alignment was computed by using a linker sequence to connect the miRNA and target sequences into a single linear sequence and make possible the use of the RNAfold program (which requires a single linear RNA sequences as input). In the present work we used the same linker sequence applied in [22], "AAAGGGLL-LLCCCUUU", which according to authors ensures that each part of the subsequence extracted from the alignment will be paired and does not change the thermodynamics qualitatively. In total, 34 features were drawn from the miRanda output, six of which referring to thermodynamics and structural features of the seed region and the remainder concerning properties of the complete alignment, as shown in Fig. 3.

		Prediction		
		Non-Target	Target	Error
Real	Non-Target	346	61	0.1498
	Target	25	1049	0.0232

**Table 1.** Confusion matrix for RFMirTarget trained upon 28 features.

## 2.5 Performance Assessment

The tools's performance was assessed by computing the total prediction accuracy (ACC), specificity (SPE), sensitivity (SEN) and Matthew's correlation coefficient (MCC) based on the confusion matrix. This matrix is provided by the training process and quantifies the number of instances from the OOB data classified as false positive (FP), true positive (TP), false negative (FN) and true negative (TN).

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$SPE = \frac{TN}{TN + FP} \times 100\% \quad (2) \quad SEN = \frac{TP}{TP + FN} \times 100\% \quad (3)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TN + FN) \times (TP + FN) \times (TN + FP)}} \quad (4)$$

## 3 Results

### 3.1 RFMirTarget Prediction Performance Based on 28 Features

At first, the RF classifier was trained with a set of 28 features, which comprises all the features summarized in Fig. 3 except for those concerning the seed region, i.e., the MFE and five structure-based features that quantify the number of matches



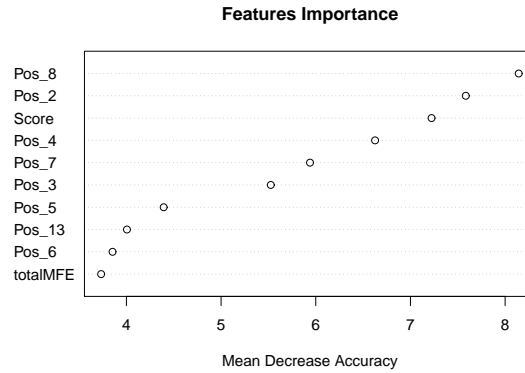
**Fig. 4.** Error rates for RFMirTarget trained with 28 features. The generalization error decreases as the number of trees in the ensemble prediction increases.

and mismatches between nucleotides 2-8 (see Fig. 2). We adopted the standard values suggested by the `randomForest` R package and trained the model with 500 trees and 5 predictors each. The number of predictors is computed as the square root of the number of variables and this formula is known to lead to results near optimal values [19].

Classification results in terms of the confusion matrix are reported in Table 1. RFMirTarget is very accurate, specially in what concerns the positive examples: the classification error for the target class was minimal, close to 0.02. From the 1074 positive examples, only 25 were misclassified, indicating an outstanding performance to detect true positive targets. The classification error for the non-target class was higher, but still satisfactory: 346 out of the 407 negative examples were correctly classified. The performance metrics for the 28-features model are ACC: 94.19, SEN: 97.67, SPE: 85.01 and MCC: 0.85.

Fig. 4 depicts the evolution of the error rate for the target class, non-target class and OOB data according to the number of trees used in the prediction. Error values start to stabilize from 200 trees. Yet, experiments have shown that there is still a performance gain when adopting 500 trees. In addition, the advantage of using ensemble predictions is clear: the error for the predictions based on 500 trees is much lower than the predictions of a single tree for all three cases.

As previously mentioned in Section 2.1, one important utility of RF classifiers is that they naturally provide an estimative of features importance computed as the forest building progresses. We analyzed this information and found that seven out of the ten most important features according to the average decrease in accuracy (Fig. 5) refer to the seed region, more specifically to the position-based features regarding the seed location (`Pos_2` to `Pos_8`). Hence, we decided to extend our model and include the seed thermodynamics and structural features, training a new RF classifier with the complete set of 34 features summarized in Fig. 3.



**Fig. 5.** Features importance for the 28-features RFMirTarget model. Features related to seed region play a crucial role in prediction accuracy.



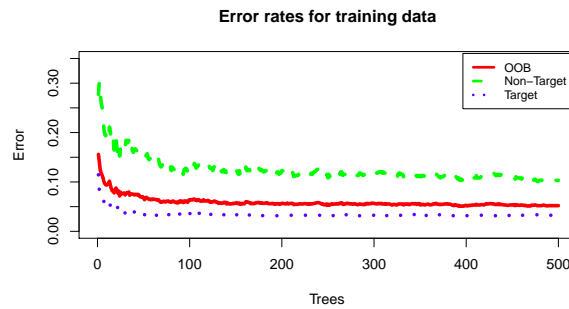
		Prediction		
		Non-Target	Target	Error
Real	Non-Target	365	42	0.1031
	Target	35	1039	0.0325

**Table 2.** Confusion matrix for the 34-features RFMirTarget model.

### 3.2 RFMirTarget Prediction Performance Based on 34 Features

The RF model trained with the set of 34 features comprising the complete alignment and seed properties have presented moderated improvement. Results are summarized in Table 2. Compared to the case with 28 features, despite the slight increase in the error rate for the target class, a significant decrease of 31% was observed in the misclassified negative examples: the error rate reduced to 0.10. An analysis of Table 2 in contrast to Table 1 shows that the overall error rate of RFMirTarget suffered a slight reduction, from 0.058 to 0.051. Fig. 6 illustrates the variation in the error rates for the target class, non-target class and OOB data according to the number of trees used in the training process: the error rates decrease as the number of trees increase.

A comparison between the 28-features and 34-features models in terms of the performance metrics discussed in Section 2.5 is given in Table 3. We include also the performance values reported for MultiMiTar [16]. A slight performance gain is observed for the 34-features model in relation to the 28-features model regarding all metrics except sensitivity. The greatest increase is related to the specificity, which suggests that information about the seed region is important to the correct identification of true negative instances, and thus to decrease the rate of false positives. Moreover, the analysis of the 34-features model results in contrast to MultiMiTar indicates that RFMirTarget has a balance between specificity and sensitivity as good as the latter. In addition, RFMirTarget shows an increase of 48% and 18.5% in MCC and accuracy, respectively, when compared



**Fig. 6.** Error rates for RFMirTarget trained upon 34 features. Again, the ensemble prediction provides a more accurate result than single classifier prediction.

	ACC (%)	SEN (%)	SPE (%)	MCC
RFMirTarget, 28-features model	94.19	97.67	85.01	0.85
RFMirTarget, 34-features model	94.80	96.74	86.68	0.86
MultiMiTar*	80.00	89.83	70.21	0.58

**Table 3.** Comparison between RFMirTarget models trained with 28 features and 34 features, in contrast to the competing tool MultiMiTar. All models are trained upon the same data set. \*Values reported in the tools’ original paper [16].

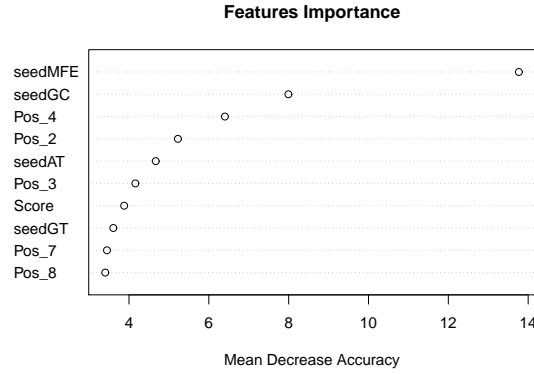
to MultiMiTar. Despite the fact that MultiMiTar performance metrics are computed based on an independent testing data set, the confusion matrix generated by RFMirTarget also provides an unbiased measurement of the performance [17] due to the use of bootstrap samples when training the RF model. Hence, this process generates a reliable method assessment.

The analysis of the ten most important features revealed that the features related to the seed region have the greatest impact on method’s performance, as expected (Fig. 7). According to this analysis, the most relevant features are the MFE value and the number of GC base pairs of the seed region, which is biologically plausible since GC pairings are more stable because they involve three hydrogen bonds.

## 4 Conclusion and Future Work

In the present paper we introduced RFMirTarget, a classifying system for miRNA targets prediction based on the RF algorithm. In [9], RFs were introduced as a tool for predicting miRNA precursors and identifying novel miRNAs, performing better than the well-known SVM classifier. Moreover, RF algorithms are of easy implementation, require less computational resources and time, and they are of more easy understanding for the final user when compared to other classification methods such as SVM. As RFs are ensembles of classification trees, they inherit the interpretability property of the latter and can be easily translated into rules, hence representing an interesting tool for exploring data [24]. Nonetheless, none of the tools for miRNA-target prediction proposed so far were built upon RFs.

The first RF model presented in the current work was trained with 28 structural, thermodynamics, alignment and position-based features extracted from the complete alignment between miRNA and candidate target. The results were encouraging: 94% of the examples were correctly classified. Seven out of the ten most important features are related to seed properties. Aggregating six thermodynamics and structural features of the seed region in the training process of the RF model, the classification performance improved in almost all senses. The list of the ten most important features for this model includes only seed features, except for the alignment score provided by the miRanda. Thus, our tool reinforces the importance of the seed region for target prediction accuracy, corroborating previous studies in the area [13, 14].



**Fig. 7.** Features importance for the 34-features RFMirTarget model. Besides the score, which is associated to the complete alignment, all relevant features concern the seed region. The MFE value of the seed region, in particular, causes a high impact in the prediction accuracy.

Comparing RFMirTarget with the competing tool MultiMiTar [16], which has presented the best predictions results for miRNAs targets so far, RFMirTarget has superior performance in all aspects. The MCC was significantly higher for the RFMirTarget: 0.86 against 0.58, which represents an improvement of 48%. The accuracy was also enhanced nearly to 18.5%: RFMirTarget has correctly predicted 94.80% of the examples, in contrast to 80.00% of MultiMiTar. MultiMiTar is a SVM classifier trained upon the same data set used in this work and its performance on an independent testing data set is superior than former methods [16]. Therefore, one can conclude that RFMirTarget is a reliable and robust strategy for miRNA target prediction when compared to other existing popular methods.

An interesting direction for future work is a deeper investigation of the impact of seed region for miRNAs targets prediction. Features related to the seed region are predominant in the analysis of the most impacting features regarding model's accuracy, such that we find worth investigating the predictive power of a classifying model trained based on this set of features, and perform comparisons with other available tools using an independent testing data set. RFMirTarget should be soon made freely available for download under a GPL license.

## References

1. Bartel, D.P.: MicroRNAs: Genomics, review biogenesis, mechanism, and function. *Cell* **116** (2004) 281–297
2. Betancur, J.G., Tomari, Y.: Dicer is dispensable for asymmetric RISC loading in mammals. *RNA* **18**(1) (2011) 1–7

3. Lee, R.C., Feinbaum, R.L., Ambrost, V.: The *C. elegans* Heterochronic Gene *lin-4* Encodes Small RNAs with Antisense Complementarity to *lin-14*. *Cell* **75** (1993) 843–854
4. Lu, M., Zhang, Q., Deng, M., Miao, J., Guo, Y., Gao, W., Cui, Q.: An analysis of Human microRNA and disease associations. *PLoS ONE* **3**(10) (2008) e3420
5. Chen, X.: microRNA biogenesis and function in plants. *FEBS Letters* **579** (2005) 5923–5931
6. Liu, J., Zheng, M., ling Tang, Y., hua Liang, X., Yang, Q.: microRNAs, an active and versatile group in cancers. *Int J Oral Sci* **3** (2011) 165–175
7. Xue, C., Li, F., He, T., Liu, G.P., Li1, Y., Zhang, X.: Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics* **6**(1) (2005) 310
8. Batuwita, R., Palade, V.: microPred: effective classification of pre-miRNAs for human miRNA gene prediction. *Bioinformatics* **25**(8) (2009) 989–995
9. 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 Research* **35** (2007) 339–344
10. Yousef, M., Nebozhyn, M., Shatkay, H., Kanterakis, S., Showe, L.C., Showe1, M.K.: Combining multi-species genomic data for microRNA identification using a Naïve Bayes classifier. *Bioinformatics* **22**(11) (2006) 1325–1334
11. Zhang, Y.: miRU: an automated plant miRNA target prediction server. *Nucleic Acids Research* **33** (2007) W701–W704
12. Lytle, J.R., Yario, T.A., Steitz, J.A.: Target mRNAs are repressed as efficiently by microRNA-binding sites in the 5' UTR as in the 3' UTR. *PNAS* **104**(23) (2007) 9667–9672
13. Mazière, P., Enright, A.J.: Prediction of microRNA targets. *Drug Discovery Today* **12**(11/12) (June 2007) 452–458
14. Lhakhang, T.W., Chaudhry, M.A.: Current approaches to microRNA analysis and target gene prediction. *Journal of Applied Genetics* (2011) 1–10
15. Witkos, T.M., Koscińska, E., Krzyzosiak, W.J.: Practical aspects of microRNA target prediction. *Current Molecular Medicine* **11** (2011) 93–109
16. Mitra, R., Bandyopadhyay, S.: MultiMiTar: A novel multi objective optimization based miRNA-target prediction method. *PLoS ONE* **6**(9) (2011) e24583
17. Breiman, L.: Random forests. *Machine Learning* **45**(1) (2001) 5–32
18. Breiman, L.: Bagging predictors. *Machine Learning* **24**(2) (1996) 123–140
19. Liaw, A., Wiener, M.: Classification and Regression by randomForest. *R News* **2**(3) (2002) 118–22
20. Bandyopadhyay, S., Mitra, R.: TargetMiner: microRNA target prediction with systematic identification of tissue-specific negative examples. *Bioinformatics* **25**(20) (2009) 2625–2531
21. Enright, A., John, B., Gaul, U., Tuschl, T., Sander, C., Marks, D.: MicroRNA targets in *drosophila*. *Genome Biology* **5**(1) (2003) R1
22. Kim, S.K., Nam, J.W., Rhee, J.K., Lee, W.J., Zhang, B.T.: mitarget: microRNA target gene prediction using a support vector machine. *BMC Bioinformatics* **7**(1) (2006) 411
23. Hofacker, I.L.: Vienna RNA secondary structure server. *Nucleic Acids Research* **31**(1) (2003) 3429–3431
24. Tkacz, A., Rychlewski, L., Uva, P., Plewczynski, D.: Supervised classification of genes and biological samples. In de Rinaldis, E., Lahm, A., eds.: *DNA Microarrays: Current Applications*. First edn. Taylor & Francis (2007) 101–120