miRNA Target Prediction through Mining of miRNA Relationships

Yanju Zhang, Jeroen S. de Bruin and Fons J. Verbeek

Abstract— **miRNAs are small regulators that mediate gene expression and each miRNA regulates specific target genes. In animals, target prediction of the miRNAs is accomplished through several computational methods, i.e. miRanda, TargetScan and PicTar. Typically, these methods predict targets from features of miRNA-target interaction such as sequence complementarity, free energy of RNA duplexes and conservation of target sites. They are constructed for high throughput and also result in a large amount of predictions and a high estimated false-positive rate. To date, specific rules to capture all known miRNA targets have not been devised. We observed that miRNAs sometimes share targets. Therefore, in this paper we present an approach which analyzes miRNA-miRNA relationships and utilizes them for target prediction. We use machine learning techniques to reveal the feature patterns between known miRNAs. Different data setups are evaluated and compared to achieve the best performance. Furthermore, the derived rules are applied to miRNAs of which the targets are not yet known so as to see if new targets could be predicted. In the analysis of functionally similar miRNAs, we found that genomic distance and seed similarity between miRNAs are dominant features in the description of a group of miRNAs binding the same target. Application of one specific rule resulted in the prediction of targets for seven miRNAs for which the targets were formerly unknown. Some of these targets were also detected by the existing methods. Our method contributes to the improvement of target identification by predicting targets with high specificity and without conservation limitation.**

I. INTRODUCTION

MicroRNAs (miRNAs) are ∼22 nucleotide single-stranded noncoding RNA molecules that serve as post-transcriptional regulators of gene expression in plants, animals and viruses. They bind to target messengerRNAs (mRNAs) and block the target expression by repressing mRNA translation or mediate mRNA degradation [5], [1]. Recent studies revealed the key roles of miRNAs in diverse regulatory functions including developmental timing regulation [19], apoptosis [3] and cell proliferation [13]. Some of them are even implied as potential tumor suppressors [9] and oncogenes [8].

A central goal for understanding biological function of miRNAs has been to identify miRNA targets since miRNAs act as regulators by binding to mRNAs. Currently, specific rules for functional miRNA-target pairing that capture all known functional targets have not been devised [4]. In animals in particular, the loose sequence complementarity in miRNA-target interaction has complicated computational approaches for target site identification.

Manuscript received: 05 July 2008. This research has been partially supported by the BioRange program of the Netherlands Bioinformatics Centre (NBIC, BSIK grant).

The authors are with Leiden Institute of Advanced Computer Science, Leiden Univeristy, Niels Bohrweg 1, 2333 CA Leiden, The Netherlands. {yanju, jdebruin, fverbeek}@liacs.nl

Nowadays, to establish possible miRNA targets, a number of high throughput computational methods/tools have been developed i.e. miRanda [5], TargetScan [15], Pictar [12], miTarget [11] and RNAhybrid [18]. All of these predict miRNA targets from different methodologies. MiRanda takes position weight into account to estimate sequence complementarity, uses RNAFold for free-energy calculation, and relies on evolutionary conservation of the binding sites [5], [2]. TargetScan seeks a strong 7-nucleotide seed, uses RNAFold to calculate the thermodynamic free-energy of the binding, and scores both single and multiple binding sites [15], [2]. PicTar takes sets of coexpressed miRNAs and searches for combinations of miRNA binding sites in each 3'UTR [12]. miTarget is a support vector machine classifier for miRNA target-gene prediction, which utilizes a radial basis function kernel to characterize targets by structural, thermodynamic, and position-based features [11]. RNAhybrid predicts multiple potential binding sites of miRNAs in large target RNAs, and finds the energetically most favorable hybridizations of a small RNA in a large RNA [18].

The aforementioned methods however, predict large amounts of targets per miRNA and include lots of false-positives in the result. The estimated false-positive rate (FPR) for PicTar, miRanda and TargetScan is about 30%, 24-39% and 22-31% respectively [2], [22], [15]. It has been reported that miTarget have similar performace like TargetScan [11]. In addition to the relatively high FPR, *Enright et al.* observed that many real targets are not predicted by these methods and this seems to be largely due to requirements for evolutionary conservation of the putative miRNA target-site across different species [5], [17].

In general we notice that in all of these algorithms, the target prediction is based on features that consider the miRNA-target interaction such as sequence complementarity and stability of miRNA-target duplex. Through the observations in the population of confirmed miRNAs-targets we became aware that some miRNAs are validated as binding the same target. Subsequently, we considered that this observation would allow target identification from the analysis of functionally similar miRNAs. Based on this idea, in this paper, we propose a method which predicts miRNA targets from analyzing the relationships between miRNAs instead of miRNA-target relations. By deducting the rules which characterize the properties of functionally similar miRNAs, targets can be found for these miRNAs for which no regulatory function was known. Our method avoids considering evolutionary conservation and only produces a small number of predictions; for such numbers, it becomes feasible to test them through biological experiment.

II. MATERIALS AND METHODS

A. Data collection

Currently, our study focuses on human miRNAs. The input data set is collected from Tarbase, which is the repository for a manually curated collection of experimentally tested miRNA targets [21]. The latest version (TarBase-V4) includes 99 human genes as translationally repressed miRNA targets and 359 human genes as cleaved miRNA targets. Among the cleaved miRNA targets, 335, about 93% of the total, are supported by the microarray assay cited from one publication [16]. Considering the fact that microarray results tend to be rather noisy, we think that it is unreliable to treat them as a training set. Thus in this stage only the miRNAs with translationally repressed targets are used in this study.

According to our observations about the functionally similar miRNAs, we pair the miRNAs as positive if they bind the same target, and couple the rest randomly as the negative data set. In total, 18 positive pairs and 18 negative pairs are generated. For quality control reasons, the data generation step is repeated 10 times and each set is tested individually in the following analysis. Here we clarify two notions; known miRNAs are those whose function is known and have been validated for having at least one target, unknown miRNAs refer to those for which the targets are unknown.

B. Feature selection

We predefine four features: overall sequence ($\sim$ 22 nt) similarity, seed (position 2-8) similarity, nonseed (position 9-end) similarity and distance. Seed has been proven to be an important region in miRNA-target interaction [10], thus we suggest that seed similarity between miRNAs is a potentially important feature. Additionally, including nonseed and sequence similarity features enables us to investigate the property behaviours of these two regions. The idea of investigating genomic distance between miRNAs is derived from our former study. Previously, through statistical methods and heterogeneous data support, we demonstrated that the genomic location feature plays a role in miRNA-target interaction for a selection of miRNA families [27]. Here we transform this idea to the study of miRNAs relationships based on the genomic distance.

We calculated similarity using the EBI pairwise global sequence alignment tool called *needle* [20], and retrieved genomic sequence and location from the miRBase Sequence Database [6]. The distance between two miRNAs is calculated by genomic position subtraction when they are located on the same chromosome, otherwise it is set to undefined.

C. Methods

The idea of this approach is to investigate feature patterns shared by the functionally similar miRNAs and use it to identify new targets. The workflow is presented in Fig. 1. In the following, we will mainly explain rule generation procedure illustrated in the framework of Fig. 1. We start with description of each algorithm followed by detailed configurations.

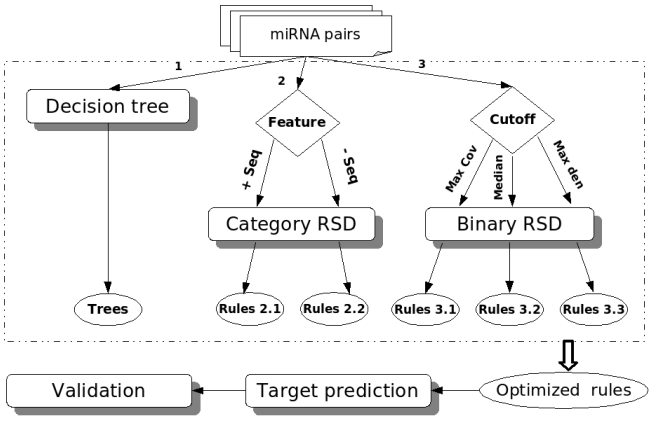


Fig. 1. Our method is composed of three phases: rule generation, target prediction and validation. In the first stage, rules are discovered from three methods with respect to decision tree and relational subgroup discovery techniques. Through combining the results from these methods, the optimized rules describing functionally alike miRNAs are generated which are used for final targets prediction and validation.

1) Decision Trees: The decision tree [25] is a common machine learning algorithm used for classification and prediction. It represents rules in the form of a tree structure consisting of leaf nodes, decision nodes and edges. This algorithm starts with finding the attribute with the highest information gain which best separates the classes, then it is split into different groups. Ideally, this process will be repeated until all the leaves are pure.

As the first step (labelled as 1 in Fig. 1), decision tree learning is utilized to build a classifier discriminating two classes of miRNA pairs. In our experiments, we used the decision tree from the Weka software [25]. The features were tested using the J48 classifier and evaluated by 8 fold cross-validation considering the fact that our dataset is relatively small.

2) Relational subgroup discovery: Subgroup discovery is well suited for finding interesting subsets from the overall example set. In our experiments, we used the propositionalization based Relational Subgroup Discovery (RSD) algorithm [26]. The main improvement compared to previous subgroup discovery algorithms is achieved through the use of a weighted relative accuracy heuristic:

$$WRAcc(H \leftarrow B) = p(B) \cdot (p(H \mid B) - p(H))$$

In rule $H \leftarrow B$, H stands for Head representing classes, while B denotes the Body which consists of one or a conjunction of first-ordered features. p is the probability function. As shown in the equation, weighted relative accuracy consists of two components: weight $p(B)$, and relative accuracy $p(H \mid B) - p(H)$. The second term, relative accuracy, is the relative accuracy gain between the conditional probability of class H given that features B is satisfied and the probability of class H. A rule is only interesting if it improves over this default rule $H \leftarrow true$ accuracy [26].

Due to the fact that not all the determinant features are known at this stage, we are interested in finding rules for subgroups of functionally similar miRNAs with respect to

TABLE I

CATEGORY RSD RESULTS. RULES GENERATED FROM TWO DATA STRUCTURES: CONSIDERING OVERALL SEQUENCE, SEED, NONSEED SIMILARITIES AS WELL AS DISTANCE (LEFT) AND ONLY SEED, NONSEED SIMILARITIES AND DISTANCE (RIGHT).

+Overall sequence YES Rules 2.1	$\overline{Sig}$	-Overall sequence YES Rules 2.2	$\overline{Sig}$
distance $\leq$ 1kb	9.5	distance $\leq$ 1kb	9.5
seed = very high	8	seed = very high	8
distance $\leq$ 1kb AND seed = very high	7	distance $\leq$ 1kb AND seed = very high	7
nonseed = high AND seed = very high	4	seed = very high AND (seed, nonseed) = (very high, high)	4
nonseed = high	3.2	(seed, nonseed) = (very high, high)	4
distance $\leq$ 1kb AND nonseed = low	2.7	distance$\leq$1kb AND nonseed = low	2.7
seqence = high	2	nonseed = high AND seed = very high	0.8

our predefined features. We prefer rules that contain only the positive pairs and have high coverage. Thus the repetitive rules are selected, if their E-value is greater than 0.01 and at the same time the significance is above 0.

Both the category RSD (method 2) and the binary RSD (method 3) marked as 2 and 3 in Fig. 1 reveal feature patterns by utilizing the relational subgroup discovery algorithm. The main difference is that method 2 analyzes the data in a categorized format, whereas in method 3 the data is transformed to a binary form. Aiming to discover the most significant rules, different data structures and feature thresholds are evaluated and compared.

In method 2, the similarity percentage is evenly divided into 5 groups: very low (0-20%), low (20-40%), medium (40-60%), high (60-80%), very high (80-100%); Distance is categorized into 5 regions: 0-1kb[1], 1-10kb, 10-100kb, 100kb-end, undef (if miRNAs that are paired are located in two different chromosomes). Two relational input tables (with and without the overall sequence similarity feature) are constructed and further tested with the purpose of verifying whether the sequence has a global effect or only contributes as the combination of seed and nonseed parts.

Through the observation of density graphs of the features, as displayed in Fig. 2, we concluded that distance and seed similarity feature densities match a bimodal distribution. The same conclusion can, however, not be drawn easily for overall and nonseed sequence similarities. Moreover, as a result of the previous method 2, overall and nonseed sequence similarities have been proven as irrelevant (shown in TABLE I). Therefore, in the third method, we apply a decision tree algorithm to discriminate input data into binary values with respect to only distance and seed similarity properties of miRNA pairs. Each feature is calculated individually and only the root classifier value in the tree is used for establishing the cutoff. After that, we generate binary tables according to three criteria: maximum coverage where the value covers the most positive pairs ($Max\_coverage_{(distance,seed)} = 8947013b, 50\%$), maximum density which is the region with the highest positive pair density ($Max\_density_{(distance,seed)} = 836b, 75\%$) and median value ($Median_{(distance,seed)} = 4473924.5b, 62.5\%$) among the data sets.

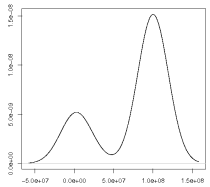
[1]Distance unit is base pair abbreviated as b, kb = kilo base pairs.

III. RESULTS

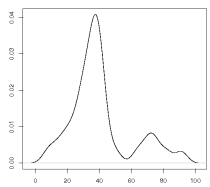
A. Rule generation

From the decision tree analysis, six different tree structures are generated from 10 replications of the training data. Among them, the root attribute or the first depth of the tree is mainly associated with distance, sequence and seed similarity properties, while nonseed feature appeared only near the leaf nodes. This inconsistency in the tree structures indicated that none of the predefined features, or any combination of them, can significantly classify miRNAs.

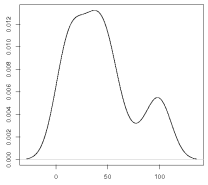
The feature patterns discovered from category RSD are listed in TABLE I where the rules in the left column take overall sequence into account but those in the right column do not. "YES"-rules describe functionally similar miRNAs characterized by our predefined features. $\overline{Sig}$ denotes the average significance over 10 replications. As for choosing the significant rules, it can be seen that the maximum gap in the sequence of $\overline{Sig}$ is between 7 and 4, therefore rules with $\overline{Sig}$ value equal or above 7 are considered as significant and stable. In the table, the insignificant ones are grayed out. Further inspection of TABLE I shows that the most significant rules from both sides are the same and they only include the distance and seed similarity features. These indicate that



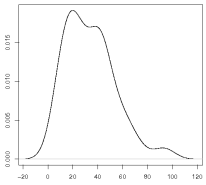
(a) Distance (basepair)



(b) Sequence (%)



(c) Seed (%)



(d) Nonseed (%)

Fig. 2. Density plot for the four features. The plots of distance (a) and seed similarity (c) match bimodal distribution indicating two main groups in each feature. However it is not straightforward to judge sequence (b) and nonseed similarity (d) distributions.

TABLE II

BINARY RSD RESULTS. RULES GENERATED FROM 3 SETS OF PARAMETERS FOR DISTANCE AND SEED SIMILARITY ARE SHOWN IN A SEQUENCE OF MAX COVERAGE (LEFT), MEDIAN (MIDDLE) AND MAX DENSITY (RIGHT).

Max coverage		Median		Max density	
YES Rules 3.1	$\overline{Sig}$	YES Rules 3.2	$\overline{Sig}$	YES Rule 3.3	$\overline{Sig}$
distance $\leq$ 8947013b	8.5	distance $\leq$ 4473924.5b	8	distance $\leq$ 836b	9.5
distance $\leq$ 8947013b AND seed $>$ 50%	6.6	distance $\leq$ 4473924.5b AND seed $>$ 62.5%	7	seed $>$ 75%	8
seed $>$ 50%	5.4	seed $>$ 62.5%	6	distance $\leq$ 836b AND seed $>$ 75%	7
distance $\leq$ 8947013b AND seed$\leq$ 50%	2.4	distance $\leq$ 4473924.5b AND seed $\leq$ 62.5%	1.8	distance $\leq$ 836b AND seed $\leq$ 75%	2.7

genomic location and seed similarity between miRNAs are dominant features when deciding which miRNAs bind to the same target.

TABLE II shows the rules generated by method 3, thereby using three cutoff criteria: Max coverage, Max density and Median. As can be seen, three rule sets have similar structures but different feature cutoffs which leads to different significance. In the comparison of these rules, we find that the Max density criterion has the best performance indicated by the highest significance value.

When comparing the rules generated from method 2 and 3, we noticed that the category and the best binary rule sets describe the same three subgroups of positive miRNA pairs. Two features with 2 values: Distance $\leq$ 1kb or $\leq$ 836b and Seed similarity $>$ 80% (very high) or $>$ 75% cover the same amount of positive miRNA pairs illustrated by the same significance.

To evaluate our methods, as a reference, a permutation test is performed. We repeat the learning procedure for each training set with the labels randomly shuffled. Using Max density as a cutoff criterion, we obtained that all the rules have the average significance lower than 1. This test therefore demonstrates that the rules derived from the original data are more significant than in the random situation.

B. Prediction and validation

Integrating the results from both category and binary RSD with the aim to achieve the best accuracy, we optimize the rules as following. Each rule describes a subgroup of functionally similar miRNAs.

Rule 1: IF distance between two miRNAs $\leq$ 1kb,
Rule 2: IF seed similarity between two miRNAs $>$ 75%,
Rule 3: IF distance $\leq$ 1kb AND seed similarity $>$ 75%,
THEN they bind the same target.

We apply the above rules to find all miRNAs which serve similar functions as the known miRNAs. Rule 1, 2 and 3 discovered 46, 233 and 24 miRNA pairs respectively in each subgroup. Among these rules, subgroup 3 has been selected to be further examined since it has relative small pairs which are easy to validate. Furthermore, as Rule 3 involves more constraints, it is considered to be more reliable than the other two.

We observed that these 24 miRNA pairs discovered by Rule 3 consist of 6 confirmed pairs in which both miRNAs from each pair are well studied, 11 pairs with both members from the same family which are supposed to have the same targets, and 7 new pairs which have one well-studied miRNA and one functional unknown partner. Therefore, we induce the targets for these 7 unknown miRNAs hsa-miR-18a/ 27a/ 18b/ 20b /301b /212 /200c from their known partner. Their predicted targets are listed in TABLE III.

Informatic validation is performed to check the prediction consistency with the existing methods. TABLE III shows validation for the 6 confirmed and 7 predicted miRNA pairs. The miRNAs with confirmed targets are indicated in italic, while the miRNAs in boldface are the unknown ones for which the targets are predicted from their known partners. All of their targets are validated by examining whether they are predicted by TargetScan, miRanda, Pictar, miTarget and RNAhybrid. The table can be read as following: for example, whether the target (BCL2) is predicted by the existing methods (TargetScan) for m1 (hsa-miR-15a) or m2 (hsa-miR-16). Consequently, we discover that among our prediction, V-maf musculoaponeurotic fibrosarcoma oncogene homolog B (MAFB) for hsa-mir-301b and Retinoblastoma 1 (RB1) for hsa-miR-20b are also predicted by TargetScan and Pictar; Circadian Locomoter Output Cycles Kaput (Clock) for hsa-miR-200c and Chemokine C-X-C motif ligand 12 (CXCL12) for hsa-miR-27a are captured by miRanda; Rho GTPase-activating protein (RICS) for hsa-miR-212 is detected by Pictar; E2F transcription factor 1 (E2F1) and AIB1 for hsa-miR-18a are identified by miTarget.

IV. CONCLUSIONS AND DISCUSSION

To date, the interaction between miRNAs and their targets is not fully understood. In an effort to identify miRNA targets, some of the existing methods induce a large number of predictions and have a high estimated false-positive rate. These complicate biological validation. In this paper, we presented an approach which discovers miRNAs relationships through rule mining and utilized them for target prediction. Existing analysis methods are insufficient in identifying targets from this perspective.

Given the circumstances that not all the targets and useful features are known in advance, the classification of miRNA data using decision trees failed due to the relatively small number of samples. Whereas the relational subgroup discovery, an advanced subgroup discovery algorithm, is suitable

TABLE III

INFORMATIC VALIDATION OF CONFIRMED AND PREDICTED MIRNA PAIRS. MIRNA1 AND MIRNA2 ARE THE PARTNERS IN ONE PAIR. TARGET COLUMN SHOWS THE VALIDATED TARGETS FOR THE KNOWN MIRNAS (IN ITALIC) AND THE PREDICTED TARGETS FOR THE UNKNOWN MIRNAS (IN BOLDFACE). M1 AND M2 COLUMNS DENOTE WHETHER THE TARGETS ARE PREDICTED BY THE EXISTING METHODS FOR MIRNA1 (M1) AND MIRNA2 (M2) RESPECTIVELY.

	miRNA1 (m1)	miRNA2 (m2)	Target	Targets predicted by TargetScan m1	TargetScan m2	MiRanda m1	MiRanda m2	Pictar m1	Pictar m2	miTarget m1	miTarget m2	RNAhybrid-mfe (kcal/mol) m1	RNAhybrid-mfe (kcal/mol) m2
Confirmed	*hsa-miR-15a*	*hsa-miR-16*	*BCL2*	√	√	×	×	√	√	×	√	-24.3	-24.1
	hsa-miR-15b	*hsa-miR-16*	*BCL2*	√	√	×	×	√	√	×	√	-26.2	-24.1
	hsa-miR-17	*hsa-miR-20a*	*E2F1*	√	√	√	×	√	√	√	√	-26.8	-24.6
	hsa-miR-221	*hsa-miR-222*	*KIT*	√	√	×	×	×	×	√	√	-24.9	-26.4
	hsa-miR-23b	*hsa-miR-27b*	*Notch1*	×	×	×	×	×	×	-	-	-	-
	hsa-miR-372	*hsa-miR-373*	*LATS2*	√	√	×	×	√	√	×	×	-17.4	-23.2
New	*hsa-miR-17*	**hsa-miR-18a**	*E2F1*	√	×	√	×	√	×	√	√	-26.8	-26.8
			AIB1	-	-	-	-	-	-	√	√	-26.3	-26.6
	hsa-miR-23a	**hsa-miR-27a**	*FLJ13158*	-	-	-	-	-	-	-	-	-	-
			CXCL12	√	×	×	√	√	×	√	×	-24.2	-24.4
	hsa-miR-106a	**hsa-miR-18b**	*RB1*	√	×	×	×	√	×	×	×	-23.2	-28.3
	hsa-miR-106a	**hsa-miR-20b**	*RB1*	√	√	×	×	√	√	×	×	-23.2	-27.2
	hsa-miR-130b	**hsa-miR-301b**	*MAFB*	√	√	×	×	√	√	√	×	-26.6	-21.9
			MCSF	-	-	-	-	-	-	-	-	-	-
	hsa-miR-132	**hsa-miR-212**	*RICS*	×	×	-	-	√	√	-	-	-	-
	hsa-miR-141	**hsa-miR-200c**	*Clock*	×	×	√	√	×	×	√	×	-22.1	-20.1

for this application domain since it can discover the rules for subgroups of similar function miRNAs with respect to our predefined features. During the rule mining, we also noticed that feature threshold optimization is a crucial procedure which helps revealing the significant rules.

We have established that distance and seed similarity are determinants. The question is whether it makes sense from the biological point of view? It has been reported that many miRNAs appear in clusters on a single polycistronic transcript [23]. They are transcribed together in a long primary transcript, yielding one or more hairpin precursors and finally are cut to multi-mature miRNAs. *Tanzer et al.* reported that the human mir-17 cluster contains six precursor miRNA (mir-17/ 18/ 19a/ 20/ 19b-1/ 92-1) within a region of about 1kb on chromosome 13 [23]. These observations are consistent with the feature embedded in Rule 1. Besides the fact that clustered miRNAs can be transcribed together, we further showed that miRNAs that are in each others proximity can bind to the same targets so as to serve as the regulators for the same purpose.

As for seed similarity, Rule 2 describes that the miRNAs with seed similarity above 75% share the same targets. This means only a perfect match or one mismatch in the seed is allowed in the process of binding the same targets. This is consistent with the idea that seed is a specific region, in particular it requires nearly perfect match with the target [10]. Moreover, TargetScanS (the new version of TargetScan) [14] also only requires a 6-nt seed match comprising nucleotides 2-7 of the miRNA. Thus, the rule requiring at least 6 out of 7 nucleotides to be similar in seed region is reasonable.

In this research field, there is, currently, already an agreement on the importance of seed region, but little attention has been given to the genomic vicinity. In this study, we proved that the genomic location also contributes for miRNA target identification.

At the end of our analysis, we assigned targets for seven miRNAs according to Rule 3. For instance, we suggest hsa-miR-20b has the same targets as hsa-miR-106a which targets tumor suppressors Retinoblastoma 1 (RB1). It has been studied that hsa-miR-106a and hsa-miR-20a are associated to colon, pancreas and prostate cancers [24]. And hsa-miR-20b is another isoform of hsa-miR-20. Therefore hsa-miR-20b has high probability of binding to RB1.

In order to support our findings, we validated the results using five existing algorithms presented in TABLE III. Not all of the predicted targets are identified by TargetScan, miRanda, Pictar, miTarget and RNAhybrid, whereas this is the same case for the known targets. Besides FLJ13158 and MCSF whose 3' UTR sequence information is not available in the database, the rest of the candidates are predicted by at least one of these methods. Both miTarget and our method are based on machine learning techniques; miTarget considers sequence and structure features of miRNA-target duplexes whereas we focus on the genomic location and sequence features between miRNAs. Moreover, we noticed that miRanda has a relatively low performance for target prediction in human. This may be due to the fact that miRanda was initially developed to predict miRNA targets in *Drosophila melanogaster*, and later adapted to vertebrate genomes [5]. In the application of RNAhybrid tool, predefined threshold of the normalized minimum free energy (mfe) is lacking, we therefore decided to list the original data resulting values. We found that most of our predicted miRNA-target duplexes are more stable illustrated by the relatively lower minimum free energy than the known ones.

In addition to these encouraging results, we also noticed that only groups of miRNA relationships are discovered by

our method. Some miRNAs which are located far apart and whose seed similarity is low still have the same target. This indicated that besides genomic distance and seed similarity, more features need to be included in order to find more and better pattens shared by functionally alike miRNAs. *Grimson et al.* uncovered five general features of target site context beyond seed pairing that boost site efficacy [7]. In future research we will explore the site context in the miRNA relationship analysis. Additionally, we also consider to take into account miRNA co-expression patterns.

In summary, we conclude that genomic distance and seed similarity are the determinants for describing the relationships of functionally similar miRNAs. Our method contributes to the improvement of target identification by predicting targets with high specificity. Moreover, it does not require conservation information for classification, so it is free from the limitations of some of the existing methods. In future research, with more biologically validated targets and features available, more rules can be generated from a large dataset, and consequently more targets can be identified to the functionally unknown miRNAs.

V. ACKNOWLEDGEMENTS

We would like to thank Erno Vreugdenhil for his discussion on the biological implications of the results and Peter van de Putten for his technical suggestion on the use of WEKA. We are also grateful to Laura Bertens and Amalia Kallergi for their suggestions in writing the manuscript.

REFERENCES

[1] D. P. Bartel. Micrornas: Genomics, biogenesis, mechanism, and function. *Cell*, 116(2):281–297, January 2004.

[2] I. Bentwich. Prediction and validation of micrornas and their targets. *FEBS Lett*, 579(26):5904–5910, October 2005.

[3] J. Brennecke, D. R. Hipfner, A. Stark, R. B. Russell, and S. M. Cohen. bantam encodes a developmentally regulated microrna that controls cell proliferation and regulates the proapoptotic gene hid in drosophila. *Cell*, 113(1):25–36, April 2003.

[4] J. Brennecke, A. Stark, R. B. B. Russell, and S. M. M. Cohen. Principles of microrna-target recognition. *PLoS Biol*, 3(3), February 2005.

[5] A. J. Enright, B. John, U. Gaul, T. Tuschl, C. Sander, and D. S. Marks. Microrna targets in drosophila. *Genome Biol*, 5(1), 2003.

[6] S. Griffiths Jones, R. J. Grocock, S. van Dongen, A. Bateman, and A. J. Enright. mirbase: microrna sequences, targets and gene nomenclature. *Nucleic Acids Res*, 34(Database issue), January 2006.

[7] A. Grimson, K. K.-H. K. Farh, W. K. K. Johnston, P. Garrett-Engele, L. P. P. Lim, and D. P. P. Bartel. Microrna targeting specificity in mammals: Determinants beyond seed pairing. *Mol Cell*, 27(1):91–105, July 2007.

[8] L. He, M. M. Thomson, M. T. Hemann, E. Hernando-Monge, D. Mu, S. Goodson, S. Powers, C. Cordon-Cardo, S. W. Lowe, G. J. Hannon, and S. M. Hammond. A microrna polycistron as a potential human oncogene. *Nature*, 435(7043):828–833, 2005.

[9] S. M. Johnson, H. Grosshans, J. Shingara, M. Byrom, R. Jarvis, A. Cheng, E. Labourier, K. L. Reinert, D. Brown, and F. J. Slack. Ras is regulated by the let-7 microrna family. *Cell*, 120(5):635–647, March 2005.

[10] F. V. Karginov, C. Conaco, Z. Xuan, B. H. Schmidt, J. S. Parker, G. Mandel, and G. J. Hannon. A biochemical approach to identifying microrna targets. *Proceedings of the National Academy of Sciences*, pages 19291–19296, November 2007.

[11] S.-K. Kim, J.-W. Nam, J.-K. Rhee, W.-J. Lee, and B.-T. Zhang. mitarget: microrna target-gene prediction using a support vector machine. *BMC Bioinformatics*, 7:411+, September 2006.

[12] A. Krek, D. Grün, M. N. Poy, R. Wolf, L. Rosenberg, E. J. Epstein, P. Macmenamin, I. d. da Piedade, K. C. Gunsalus, M. Stoffel, and N. Rajewsky. Combinatorial microrna target predictions. *Nature Genetics*, 37(5):495–500, April 2005.

[13] C. Lecellier, P. Dunoyer, K. Arar, J. Lehmann-Che, S. Eyquem, C. Himber, A. Sab, and O. Voinnet. A cellular microrna mediates antiviral defense in human cells. *Science*, 308(5721):795–825, April 2005.

[14] B. P. Lewis, C. B. Burge, and D. P. Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microrna targets. *Cell*, 120(1):15–20, January 2005.

[15] B. P. Lewis, I.-H. Shih, M. W. Jones-Rhoades, D. P. Bartel, and C. B. Burge. Prediction of mammalian microrna targets. *Cell*, 115(7):787–798, December 2003.

[16] L. P. Lim, N. C. Lau, P. Garrett-Engele, A. Grimson, J. M. Schelter, J. Castle, D. P. Bartel, P. S. Linsley, and J. M. Johnson. Microarray analysis shows that some micrornas downregulate large numbers of target mrnas. *Nature*, aop(current), January 2005.

[17] K. P. S. C. Martin G, Schouest K. Prediction and validation of microrna targets in animal genomes. *J Biosci*, 32(6):1049–1052, September 2007.

[18] M. Rehmsmeier, P. Steffen, M. Hochsmann, and R. Giegerich. Fast and effective prediction of microrna/target duplexes. *RNA*, 10(10):1507–1517, October 2004.

[19] B. J. Reinhart, F. J. Slack, M. Basson, A. E. Pasquinelli, J. C. Bettinger, A. E. Rougvie, R. H. Horvitz, and G. Ruvkun. The 21-nucleotide let-7 rna regulates developmental timing in caenorhabditis elegans. *Nature*, 403(6772):901–906, February 2000.

[20] D. Sankoff and J. Kruskal. *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*. Center for the Study of Language and Inf, December 1999.

[21] P. Sethupathy, B. Corda, and A. G. Hatzigeorgiou. Tarbase: A comprehensive database of experimentally supported animal microrna targets. *RNA*, 12(2):192–197, February 2006.

[22] P. Sethupathy, M. Megraw, and A. G. Hatzigeorgiou. A guide through present computational approaches for the identification of mammalian microrna targets. *Nature Methods*, 3(11):881–886, October 2006.

[23] A. Tanzer and P. F. Stadler. Molecular evolution of a microrna cluster. *J Mol Biol*, 339(2):327–335, May 2004.

[24] S. Volinia, G. A. A. Calin, C.-G. G. Liu, S. Ambs, A. Cimmino, F. Petrocca, R. Visone, M. Iorio, C. Roldo, M. Ferracin, R. L. L. Prueitt, N. Yanaihara, G. Lanza, A. Scarpa, A. Vecchione, M. Negrini, C. C. C. Harris, and C. M. M. Croce. A microrna expression signature of human solid tumors defines cancer gene targets. *Proc Natl Acad Sci U S A*, February 2006.

[25] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, October 1999.

[26] F. Zelezny and N. Lavrac. Propositionalization-based relational subgroup discovery with rsd. *Machine Learning*, 62(1-2):33–63, February 2006.

[27] Y. Zhang, J. M. Woltering, and F. J. Verbeek. Screen of microrna targets in zebrafish using heterogeneous data sources: A case study for dre-mir-10 and dre-mir-196. *International Journal of Mathematical, Physical and Engineering Sciences*, 2(1):10–18, November 2007.