Chapter 3

miRNA Target Prediction through Mining of miRNA Relationships

Based on

Yanju Zhang, Jeroen S. de Bruin, Fons J. Verbeek. (2010). Specificity enhancement in microRNA target prediction through knowledge discovery. Chapter 20. In: Machine Learning, Eds. Yagang Zhang, ISBN: 978–953–307–033–9, INTECH

Summary

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 chapter 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 predicted by other existing methods. Our method contributes to the improvement of target identification by predicting targets with high specificity and without conservation limitation.

1 Introduction

In this chapter we explore and investigate a range of methods in pursue of improving target prediction of microRNA. The currently available prediction methods produce a large output set that also includes a rather high amount of false positives. Additional strategies for target prediction are necessary and we elaborate on one particular group of microRNAs; i.e. those that might bind to the same target. We intend to transfer our approach to other groups of microRNAs as well as the broader application to the important model species.

microRNAs (miRNAs) are a novel class of post-transcriptional gene expression regulators discovered in the genome of plants, animals and viruses. The mature miRNAs are about 22 nucleotides long. They bind to their target messengerRNA (mRNA) and therefore induce translational repression or degradation of target mRNAs [6, 1]. Recent studies have elucidated that these short molecules are highly conserved between species indicating their fundamental roles conserved in evolutionary selection. They are implicated in developmental timing regulation [26], apoptosis [3] and cell proliferation [19]. Some of them even act as potential tumor suppressors [14], potential oncogenes [13] and might be important targets for drugs [23].

The identification of large number of miRNAs existing in different species has increased the interest in unraveling the mechanism of this regulator. It has been proven that more than one miRNA regulates one target and vice versa [6]. Therefore understanding this novel network of regulatory control is highly dependent on identification of miRNA targets. Due to the costly, labor-intensive nature of experimental techniques required, currently, there is no large-scale experimental target validation available leaving the biological function of the majority completely unknown [5]. These limitations of the wet experiments lead to the development of computational prediction methods.

It has been established that the physical RNA interaction requires sequence complementarity and thermodynamic stability. Unlike plant miRNAs, which bind to their targets through near-perfect sequence complementarity, the interaction between animal miRNAs and their targets is more flexible. Partial complementarity is frequently found [6] and this flexibility complicates computation. Lots of effort has been put into characterizing functional miRNA-target pairing. The most frequently used prediction algorithms are

miRanda, TargetScan/TargetScanS, RNAhybrid, DIANA-microT, picTar, and miTarget.

MiRanda [6] is one of the earliest developed large-scale target prediction algorithm which was first designed for Drosophila then adapted for human and other vertebrates. It consists of three steps: First, a dynamic programming local alignment is carried out between miRNAs and 3' UTR of potential targets using a scoring matrix. After filtering by threshold score, the resulting binding sites are evaluated thermodynamically using the Vienna RNA fold package [35]. Finally, the miRNA pairs that are conserved across species are kept.

TargetScan/TargetScanS [22, 21] have a stronger emphasize on the seed region. In the standard version of TargetScan, the predicted target-sites first require a 7-nucleotide (nt) match to the seed region of miRNA, i.e., nucleotides 2-8; second, conservation in 4 genomes (human, mouse, rat and puffer fish), and third, thermodynamic stability. TargetScanS is the new and simplified version of TargetScan. It extends the cross-species comparison to 5 genomes (human, mouse, rat, dog and chicken) and requires a seed match of only 6-nt long (nucleotides 2-7). Through the requirement of more stringent species conservation it leads to more accurate predictions even without conducting free energy calculations.

RNAhybrid [25] was the first method which integrated powerful statistical models for large-scale target prediction. Basically, this method finds the energetically most favorable hybridization sites of a small RNA in a large RNA string. It takes candidate target sequences and a set of miRNAs and looks for energetically favorable binding sites. Statistical significance is evaluated with an extreme value statistics of length normalized minimum free energies for individual hits, a Poisson approximation of multiple hits, and the calculation of effective numbers of orthologous targets in comparative studies of multiple organisms. Results are filtered according to p-value thresholds.

DIANA-microT identified putative miRNA-target interaction using a modified dynamic programming algorithm with a sliding window of 38 nucleotides that calculated binding energies between two imperfectly paired RNAs. After filtering by an energy threshold, the candidates are examined by the rules derived from mutation experiments of a single let-7 binding site. Finally, those which were conserved between human and mouse were further considered for experimental verification [12, 28].

PicTar takes sets of co-expressed miRNAs and searches for combinations of miRNA binding sites in each 3' UTR [17]. And 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 [16].

Among the algorithms discussed previously, miRanda and TargetScan/TargetScanS belong to the sequence-based algorithms which evaluate miRNA-target complementarity first, then calculate the binding site thermodynamics to further prioritize; in contrast, DIANA-microT and RNAhybrid are based on algorithms that are rooted in thermodynamics, thus using thermodynamics as the initial indicator of potential miRNA binding site.

Until now, it remains unclear whether sequence or structure is the better predictor of a miRNA binding site [23]. All of the above mentioned methods produce a large set of predictions and include a relatively high false positive ratio; all in all this indicates that these methods are promising methods but still far away from perfect. The estimated false-positive rate (FPR) for PicTar, miRanda and TargetScan is about 30%, 24-39% and 22-31% respectively [2, 30, 22]. It has been reported that miTarget has a similar performance as TargetScan [16]. 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 [6, 8]. In general we also 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. For example, in human miR-17 and miR-20a both regulate the expression of E2F1; while miR-221 and miR-222 both bind to KIT. Subsequently, we considered that this observation would allow target identification from the analysis of functionally similar miRNAs.

Based on this idea, we present an approach which analyzes miRNA-miRNA relationships and utilizes them for target prediction. Our aim is to improve target prediction by using different features and discovering significant feature patterns through tuning and combining several machine learning techniques. To this respect, we applied feature selection, principle component analysis, classification, decision trees, and propositionalization-

based relational subgroup discovery to reveal the feature patterns between known miRNAs. During this procedure, different data setups were evaluated and the parameters were optimized. Furthermore, the derived rules were applied to functionally unknown miRNAs so as to see if new targets could be predicted. In the analysis of functionally similar miRNAs, we found that genomic distance, seed and overall sequence similarities 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 five functionally unknown miRNAs which were also detected by some of the existing methods. Our method is complementary to the existing prediction approaches. It contributes to the improvement of target identification by predicting targets with high specificity and without conservation limitation. Moreover, we discovered that knowledge discovery especially the propositionalization-based relational subgroup discovery, is suitable for this application domain since it can interpret patterns of similar function miRNAs with respect to the limited features available.

The remainder of this chapter is organized as follows. In Section 2, miRNA biology and databasing as well as the background of the machine learning techniques which are the components of our method are explained: i.e., miRNA biogenesis and function, related databases, feature selection, principle component analysis, classification, decision trees and propositionalization-based relational subgroup discovery. Section 3 specifies the proposed method including data preparation, algorithm configuration and parameter optimization. The results are summarized in Section 4. Finally, In Section 5, we discuss the strengths and the weaknesses of the applied machine learning techniques and feasibility of the derived miRNA target prediction rules.

2 Background

The first two subsections are devoted to the exploration of miRNA biology whereas the latter two subsections have a computational nature.

2.1 microRNA biogenesis and function

The mature miRNAs are ∼22 nucleotide single-stranded noncoding RNA molecules. They are derived from miRNA genes. First, miRNA gene is transcribed to primary miRNA transcripts (pri-miRNA), which is between a few hundred or a few thousand base pair long. Subsequently, this pri-miRNA is processed into hairpin precursors (pre-miRNA), which has a length of approximately 70 nucleotides, by the protein complex consisting of the nuclease Drosha and the double-stranded RNA binding protein Pasha. The pre-miRNA then is transported to cytoplasm and cut into small RNA duplexes of approximately 22 nucleotides by the endonuclease Dicer. Finally, either the sense strand or antisense strand can function as templates giving rise to mature miRNA. Upon binding to the active RISC complex, mature miRNAs interact with the target mRNA molecules through base pair complementarity, therefore inhibit translation or sometimes induce mRNA degradation [4].

It is suggested that miRNAs tend to bind 3' UTR (3' Untranslated Region) of their target mRNAs [20]. Further studies have discovered that position 2-8 of miRNAs, which is called 'seed' region, has been described as a key specificity determinant of binding, requires good or perfect complementarity [22, 21]. The process of biogenesis and function of miRNAs are illustrated in Fig. 3, Chapter 1. A detailed miRNA-target interaction is also showed with a highlighted seed region.

2.2 miRNA databases

miRBase: MiRBase is the primary online repository for published miRNA sequence data, annotation and predicted gene targets [9, 10]. It consists of three parts:

The miRBase Registry acts as an independent authority of miRNA gene nomenclature, assigning names prior to publication of novel miRNA sequences.

The miRBase Sequences is a searchable database for miRNA sequence data and annotation. The latest version (Release 13.0, March 2009) contains 9539 entries representing hairpin precursor miRNAs, expressing 9169 mature miRNA products, in 103 species including primates, rodents, birds, fish, worms, flies, plants and viruses.

The miRBase Targets is a comprehensive database of predicted miRNA target genes. The

core prediction algorithm currently is miRanda (version 5.0, Nov 2007). It searches over 2500 animal miRNAs against over 400 000 3' UTRs from 17 species for potential target sites. In human, the current version predicts 34788 targets for 851 human miRNAs.

Tarbase: Tarbase is a comprehensive repository of a manually curated collection of experimentally supported animal miRNA targets [29, 24]. It describes each supported target site by the miRNA which binds it, the target genes which includes this binding site, the direct and indirect experiments that were conducted to validate it, binding site complementarity and etc. The latest version (Tarbase 5.0, Jun 2008) records more than 1300 experimentally supported miRNA target interactions for human, mouse, rat, zebrafish, fruitfly, worm, plant, and virus. As machine learning methods become more popular, this database provides a valuable resource to train and test for machine learning based target prediction algorithms.

2.3 Pattern recognition

Pattern recognition is considered a sub-topic of machine learning. It concerns with classification of data either based on a priori knowledge or based on statistical information extracted from the patterns. The patterns to be classified are usually groups of measurements, features or observations, which define data points in an appropriate multidimensional space. Our pattern recognition proceeds in three different stages: feature reduction, classification and cross-validation.

Feature reduction: Feature reduction includes feature selection and extraction. Feature selection is the technique of selecting a subset of relevant features for building learning models. In contrast, feature extraction seeks a linear or nonlinear transformation of original variables to a smaller set. The reason why not all features are used is because of performance issues, but also to make results easier to understand and more general. Sequential backward selection is a feature selection algorithm. It starts with entire set, and then keeps removing one feature at a time so that the entire subset so far performs the best. Principle component analysis (PCA) is an unsupervised linear feature extraction algorithm. It derives new variables in decreasing order of importance that are a linear combinations of the original variables, uncorrelated and retain as much variation as possible [33].

Classification: Classification is the process of assigning labels on data records based on their features. Typically, the process starts with a training dataset that has examples already classified. These records are presented to the classifier, which trains itself to predict the right outcome based on that set. After that, a testing set of unclassified data is presented to the classifier, which classifies all the entries based on its training. Finally, the classification is being inspected. The better the classifier, the more good classifications it has made. Linear discriminant classifier (LDC) and quadratic discriminant classifiers (QDC) are two frequently used classifiers which separate measurements of two or more classes of objects or events by a linear or a quadric surface respectively.

Cross-validation: Cross-validation is the process of repeatedly partitioning a dataset in a training set and a testing set. When the dataset is partitioned in *n* parts we call that *n*-fold cross-validation. After partitioning the set in *n* parts, the classifier is trained with *n-1* parts, and tested on the remaining part. This process is repeated n times, each time a different part functions as the training part. The n results from the folds then can be averaged to produce a single estimation of error.

2.4 Knowledge discovery

Knowledge discovery is the process which searches large volumes of data for patterns in order to find understandable knowledge about the data. In our knowledge discovery strategy, decision tree and relational subgroup discovery are applied.

Decision tree: The decision tree [34] 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, and then it is split into different groups. Ideally, this process will be repeated until all the leaves are pure.

Relational subgroup discovery: Subgroup discovery belongs to descriptive induction [36] which discover patterns described in the form of individual rules. Relational subgroup discovery (RSD) is the algorithm which utilizes relational datasets as input, generates subgroups whose class-distributions differ substantially from the complete dataset with respect to the property of interest [18]. The principle of RSD can be simplified as follows; first, a feature is constructed through first-order feature construction and the features

covering empty datasets are retracted. Second, rules are induced using weighted relative accuracy heuristics and weighted covering algorithm. Finally, the induced rules are evaluated by employing the combined probabilistic classifications of all subgroups and the area under the receiver operating characteristics (ROC) curve [7]. The key improvement of RSD is the application of weighted relative accuracy heuristics and weighted covering algorithm, i.e.

$$WRAcc(H \leftarrow B) = p(B) \cdot (p(H \mid B) - p(H)) \quad (1)$$

The weighted relative accuracy heuristics is defined as equation 1. 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 [36].

In the weighted covering algorithm, the covered positive examples are not deleted from the current training set which is the case for the classical covering algorithm. Instead, in each run of the covering loop, the examples are given decreasing weights while the number of iterations is increasing. In doing so, it is possible to discover more substantial significant subgroups and thereby achieving to find interesting subgroup properties of the entire population.

3 Experimental setups, methods and materials

3.1 Data collection

In the interest of including maximally useful data, human miRNAs are chosen as the research focus. The latest version of TarBase (TarBase-V5 released at 06/2008) includes 1093 experimentally confirmed human miRNA-target interactions. Among them, 243 are supposed by direct experiment such as in vitro reporter gene (Luciferase) assay, while the rest are validated by an indirect experimental support such as microarrays. Considering the fact that the indirect experiments could induce the candidates which are in the

downstream of the miRNA involved pathways, it is uncertain whether these can virtually interact with miRNA or not. Thus they are excluded and only the miRNAs-target interactions with direct experiment support are used in this study.

We observed that some miRNAs are validated as binding the same target. According to this observation, we pair the miRNAs as positive if they bind the same target, and randomly couple the rest as the negative data set. In total, there are 93 positive pairs. After checking the consistency of the name of miRNAs and removing the redundant data (for example, miR-26 and miR-26-1 refer to the same miRNA), 73 pairs are kept and thus another 73 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.

3.2 Feature collection

In the study of miRNA-target interaction, it has been established that this physical binding requires sequence complementarity and thermodynamic stability. Here some of miRNA-target interaction features are transformed to the study of functionally similar miRNA pairs.

We predefine four features: overall sequence ($\sim$22 nt) similarity, seed (position 2-8) similarity, non-seed (position 9-end) similarity and genomic distance. Seed has been proven to be an important region in miRNA-target interaction which display an almost perfect match to the target sequence [15], thus we suggest that seed similarity between miRNAs is a potentially important feature. Additionally, including non-seed and sequence similarity features enables us to investigate the property behaviors of these two regions. Genomic distance is not a well investigated feature which is defined as base pair distance between two genes. 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 [38]. Here we induce this idea to the study of miRNAs relationships based on the genomic distance.

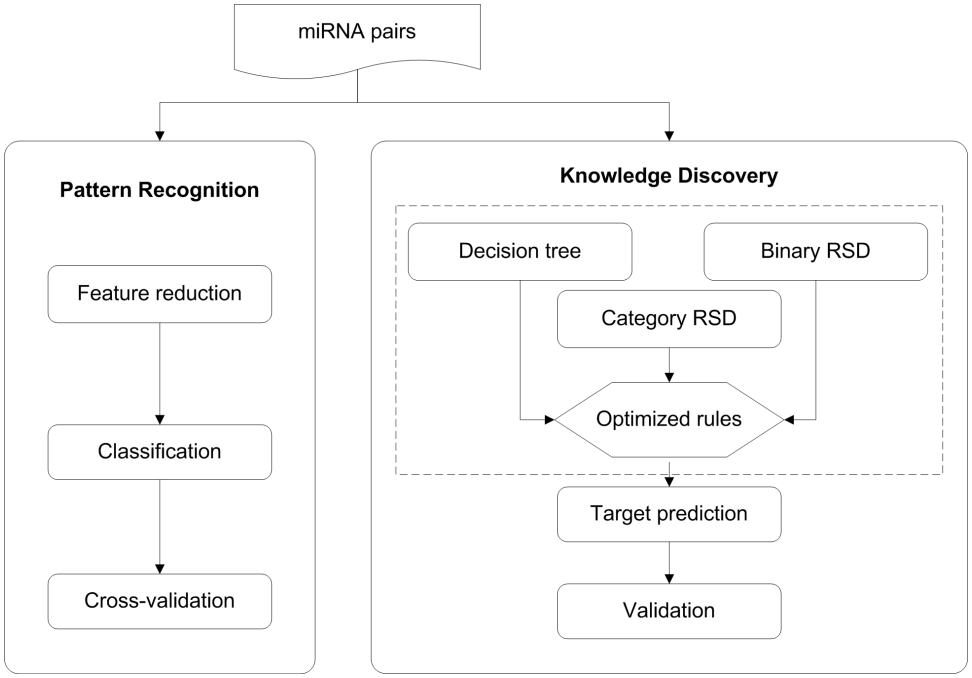


Figure 1: Workflow. miRNA pairs are analyzed by both pattern recognition and knowledge discovery strategies.

In the data preparation, sequence similarity is calculated using the EBI pairwise global sequence alignment tool: i.e. Needle [27]. Genomic sequence and location are retrieved from the miRBase Sequence Database. 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.

3.3 Workflow

As showed in Fig. 1, we use two strategies to discover miRNA-miRNA relationships. In pattern recognition strategy, different classifiers are applied to preprocessed dataset in order to discriminate positive and negative miRNA pairs. Then the performance of each classifier is evaluated by cross-validation. In knowledge discovery, rules are first discovered from three methods with respect to decision tree and relational subgroup discovery techniques. Through combining the results, the optimized rules describing functionally alike miRNAs are generated which are used for final targets prediction and validation.

Pattern recognition: In this strategy, the first step is feature reduction. Features are selected by sequential backward elimination algorithm and extracted by principle com-

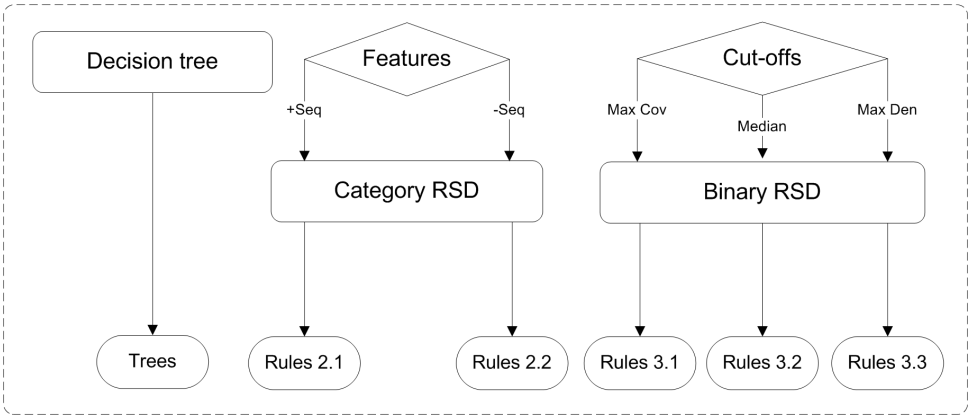
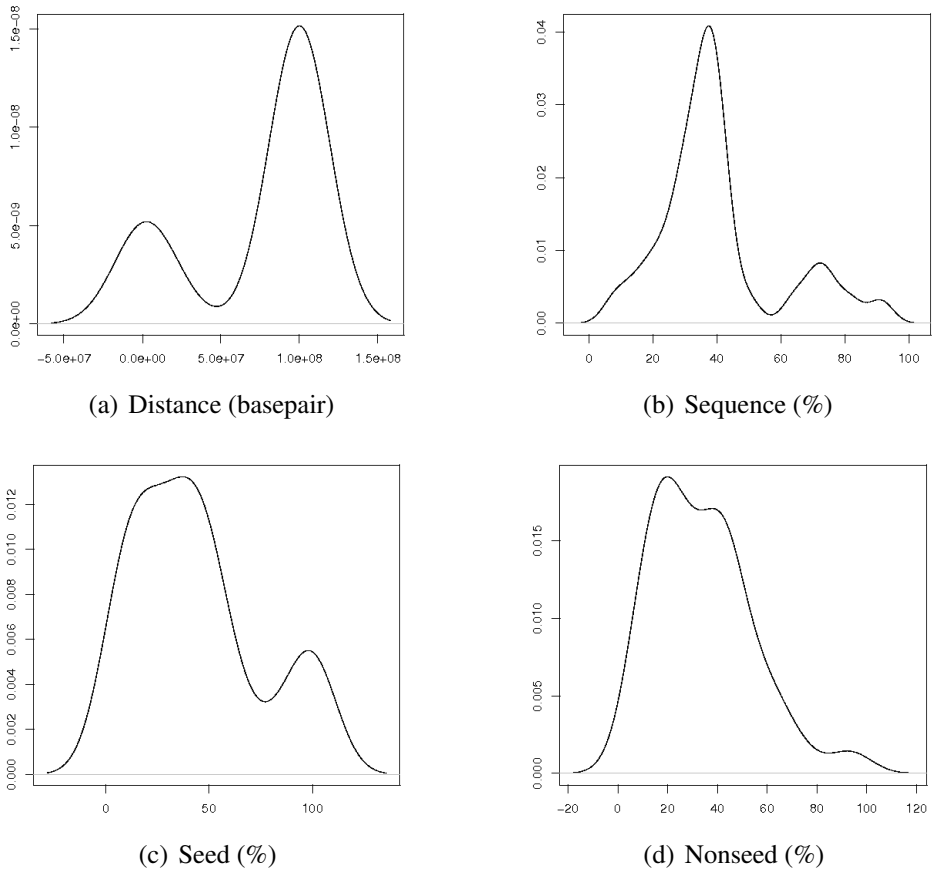


Figure 2: Detailed experimental design in rule generation stage. Three methods are applied which are Decision tree, Category RSD and Binary RSD. In Category RSD, datasets are first categorized into groups. Subsequently, data with two feature sets, which are with and without overall sequence similarity, are used as the input to RSD algorithm. In Binary RSD, feature values are binariezed using decision tree. Due to the fact that data are sampled 10 times, the cut-offs are then established using max coverage (Max Cov), median and max density (Max Den). Finally, RSD is applied to all 3 conditions in order to find out the feature cut-offs, which lead to the most significant rule sets.

ponent analysis. As it is known that sequential forward selection adds new features to a feature set one at a time until the final feature set is reached [33]. It is simple and fast. The reason it is not applied in our experiment is due to the limitation that the selected features could not be deleted from the feature set once they have been added. This could lead to local optimum. After dimension reduction, classification is performed by both linear and quadratic classifiers. Finally, the performance is examined by 5-fold cross-validation with 10 repetitions. This part was implemented with PRtools [32] a plugin for the MatLab platform.

Knowledge discovery: As contrasted to the pattern recognition which classifies miRNA pairs by complicated statistical models, knowledge discovery describes data patterns which allow us gain knowledge about the data. This could promote our understanding of functionally similar miRNAs. Furthermore, integration of this knowledge could finally promote target prediction. In this strategy, there are three phases: rule generation illustrated in the framework (dashed) of Fig. 1, target prediction and validation. In the first step, rules are discovered using decision trees and relational subgroup discovery. With the aim to discover the most significant rules, different data structures and feature thresholds are evaluated and compared. Details are explained in the following sections and an overview of this methodology is shown in Fig. 2.



(a) Distance (basepair)

(b) Sequence (%)

(c) Seed (%)

(d) Nonseed (%)

Figure 3: 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.

Decision tree learning is utilized as a first step in order to build a classifier discriminating two classes of miRNA pairs. In our experiments, we used the decision tree from the Weka software platform [34]. The features were tested using the J48 classifier and evaluated by 10 fold cross-validation.

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 our predefined features. In our experiments, we used the propositionalization based relational subgroup discovery algorithm [36]. We prefer rules that contain only the positive pairs and portray high coverage. Consequently, the repetitive rules are selected, if their E-value is greater than 0.01 and at the same time the significance is above 10.

Both the Category RSD and the Binary RSD reveal feature patterns by utilizing the rela-

tional subgroup discovery algorithm. The main difference is that the former analyzes the data in a categorized format, whereas in later algorithm the data is transformed to a binary form.

As a pilot experiment for RSD, data is first categorized as follows: 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 on a different chromosome). Two relational input tables, which are 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 non-seed parts.

Through the observation of density graphs of the features, as depicted in Fig. 3, 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 non-seed sequence similarities. Therefore, in this method, we apply a decision tree algorithm to discriminate 4 feature values into binary values. Each feature is calculated individually and only the root classifier value in the tree is used for establishing the cut-off. After that, binary tables are generated according to three criteria:

- Maximum coverage where the value covers the most positive pairs. Max coverage (distance, sequence, seed, non-seed) = 8947013 b, 56.5%, 71.4%, 53.3%

- Median. Median (distance, sequence, seed, non-seed) = 3679 b, 65.2%, 71.4%, 60.65%

- Maximum density which is the region with the highest positive pair density. Max density (distance, sequence, seed, non-seed) = 3679 b, 69.6%, 75%, 64.7%

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

4 Results

4.1 Classification

After application of sequential backward feature selection, features including genomic distance, seed similarity and non-seed similarity are selected as the top 3 informative features. Sequence similarity is the least informative feature because it is highly correlated to seed and non-seed similarities. Scatter plots of two classes of miRNA pairs in the selected feature space are depicted in Fig. 4. As can be seen in the four sub-graphs of Fig. 4, the majority of positive and negative miRNA pairs are overlapping which is an indication for the complexity of the classification. The distribution of negative class is more compact. We observed that the majority of this class located in the area of non-seed$<$60%, seed$<$70% and distance is infinite. Furthermore, we noticed that for those functionally similar miRNAs, seed similarity vary from 0 to 100%. This implies that miRNAs with the same or different seed sequence can bind the same targets. This is due to the fact that miRNAs can bind to the same targets at the same binding site which leads to high similarity and at different binding site resulting low similarity. The evaluation of the classifier performance shows that the average error and standard deviation for the quadratic classifier are 0.29739 and 0.01082, and for the linear classifier are 0.30987 and 0.0131.

In Fig. 5 the dataset is plotted in 2-dimensional PCA space in combination with the linear and quadratic classifiers. In this projected 2D space, the average error and standard deviation for the quadratic classifier are 0.3029 and 0.00721, and for the linear classifier are 0.31657 and 0.00871.

With around 30% of classification errors, this means two classes are difficult to separate using features currently available. Furthermore, although the classifiers provide a statistical explanation and meaning, no biological insight is gained from them in order to be able to interpret the miRNA mechanism(s).

4.2 Rule discovery

In the decision tree analysis, several 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

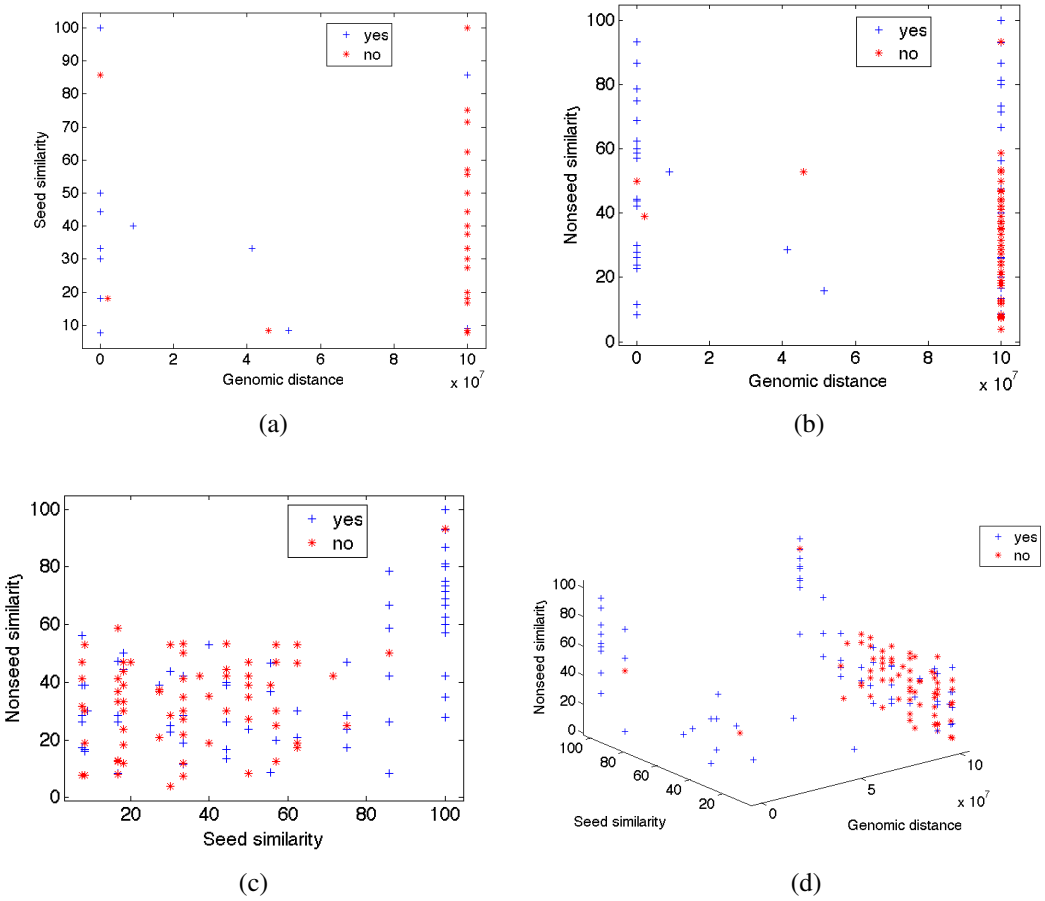


Figure 4: Scatter plots of two classes of miRNA pairs in the selected feature spaces. Positive pairs are denoted using a token of plus (blue); negatives are demonstrated by asterisk (red).

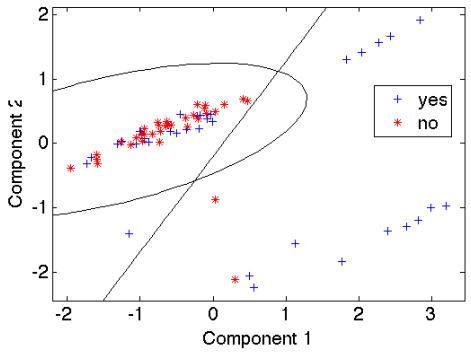


Figure 5: Scatter plot of two classes of miRNA pairs in a 2D PCA space together with a linear discriminant and a quadratic discriminant classifier showed by a line and an arc respectively.

Table 1: Category RSD results. Rules generated from two data structures: considering overall sequence, seed, non-seed similarities as well as distance (a) and only seed, non-seed similarities and distance (b).

(a)

Label	-Overall sequence : YES Rules 2.1	Significance
A	Seed>80%	26.7
A.1	Dis=undef & Seed>80%	14.3
B	Dis≤1 kb	14.1
A.2	Seed>80% & Nonseed=(60%,80%]	12.6
C	Dis=(1 kb,10 kb]	11

(b)

Label	+Overall sequence : YES Rules 2.2	Significance
A	Seed>80%	26.7
A.1	Dis=undef & Seed>80%	14.3
B	Dis≤1 kb	14.1
A.2	Seed>80% & Nonseed=(60%,80%]	11.2
C	Dis=(1 kb,10 kb]	11

mainly associated with distance, sequence and seed similarity properties, while non-seed 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 1 where the rules in Table 1(b) take overall sequence into account but those in Table 1(a) do not. 'YES'-rules describe functionally similar miRNAs characterized by our predefined features. 'Significance' denotes the average significance over 10 replications. Further inspection of Table 1 shows that both rule sets consist of 3 main groups with features being Seed>80%, Dis≤1 kb and Dis=(1 kb,10 kb] labeled by A, B, C respectively. The remainder is the subset of these groups. Considering overall sequence in the rule generation results only the fourth rule (A.2) in Table 1(a) and 1(b) to be different. These results indicate that genomic location and seed similarity between miRNAs are probably dominant features when deciding which miRNAs bind to the same target. Sequence information may be relevant but it is

Table 2: Binary RSD results. Rules generated from 3 sets of parameters are shown in a sequence of Max coverage (a), Median (b) and Max density (c).

(a)

Label	Max coverage: YES Rules 3.1	Significance
A.1	Seed>71.4% & Seq>56.5%	30
A	Seed>71.4%	27.2
A.2	Nonseed>53.3% & Seed>71.4% & Seq>56.5%	21.6
B	Dis≤8947013 b	19.8
A.3	Nonseed>53.3% & Seed>71.4%	18.2
A.4	Dis>8947013 b & Seed>71.4% & Seq>56.5%	13.5
A.5	Dis>8947013 b & Seed>71.4%	12.3

(b)

Label	Median: YES Rules 3.2	Significance
A	Seed>71.4%	27.2
A.1	Seed>71.4% & Seq>65.2%	23.3
B	Dis≤3679 b	23.3
B.1	Dis≤3679 b & Nonseed≤60.65%	15.9
A.2	Dis>3679 b & Seed>71.4%	14.9
A.3	Nonseed>60.65% & Seed>71.4% & Seq>65.2%	13.7
A.4	Nonseed>60.65% & Seed>71.4%	13.7
C.1	Nonseed>60.65% & Seq>65.2%	13.7
C	Seq>65.2%	12.2

(c)

Label	Max density: YES Rules 3.3	Significance
A	Seed>75%	26.7
B	Dis≤3679 b	23.3
A.1	Seed>75% & Seq>69.6%	20.8
C	Seq>69.6%	20.8
B.1	Dis≤3679 b & Nonseed≤64.7%	18
B.2/C.1	Dis≤3679 b & Seq≤69.6%	14.1
A.2	Dis>3679 b & Seed>75%	11.5
A.3/C.2	Nonseed>64.7% & Seed>75% & Seq>69.6%	11
C.3	Nonseed>64.7% & Seq>69.6%	11

not as strong as seed and distance features.

Table 2 shows the rules generated by Binary RSD, thereby using three cutoff criteria: Max coverage (a), Median (b) and Max density (c). As can be seen, three rule sets have similar structures but different feature cutoffs which lead to different significance. The main feature groups derived using max coverage, median and max density criteria respectively are Seed>71.4% (A) and Dis≤8947013 b (B) in rule set 3.1; Seed>71.4% (A), Dis≤3679 b (B) and Seq>65.2% (C) in rule set 3.2; and Seed>75% (A), Dis≤3679 b (B) and Seq>69.6% (C) in rule set 3.3. Others are the subsets of these groups.

Furthermore, the rules with similar features but different feature values are compared. The decision on final cut-off is based on the value which results in the highest significance. Therefore the final optimized rules are:

Rule 1: IF distance between two miRNAs ≤3679 b,
Rule 2: IF seed similarity between two miRNAs > 71.4%,
Rule 3: IF sequence similarity between two miRNAs > 69.6%
THEN they bind the same target.

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 coverage as a cutoff criterion, we obtained that all the rules have the max significance lower than 8. This test therefore demonstrates that the rules derived from the original data are more significant compared to the random situation.

4.3 Target prediction

We apply the above rules searching for miRNAs which serve similar functions as the known miRNAs. Rule 1, 2 and 3 discovered 75, 655 and 150 miRNA pairs respectively in each subgroup which highly extends our previous findings [37] based on the similar methodology. Among them, 23 miRNA predicted targets which are covered by all of the 3 rules are selected for further validation. Since this group has relative small pairs which are easy to validate. Furthermore, as they involve more constraints, it is considered to be more reliable.

Further observation of these 23 miRNA pairs, we found that it consists of 3 confirmed pairs in which both miRNAs from each pair are well studied, 15 pairs with both members

Table 3: 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	Our prediction miRNA2	Target	Targets predicted by TargetScan		MiRanda		Pictar		miTarget		RNAhybrid-mfe kcal/mol	
(m1)	(m2)		m1	m2	m1	m2	m1	m2	m1	m2	m1	m2
hsa-miR-15a	*hsa-miR-16*	*BCL2*	√	√	×	×	√	√	×	√	-24.3	-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-17	**hsa-miR-18a**	*E2F1*	√	×	√	×	√	×	√	√	-26.8	-26.8
		AIB1	-	-	-	-	-	-	√	√	-26.3	-26.6
hsa-miR-106a	**hsa-miR-18b**	*RB1*	√	×	×	×	√	×	×	×	-23.2	-28.3
hsa-miR-106a	**hsa-miR-20b**	*RB1*	√	√	×	×	√	√	×	×	-23.2	-27.2
hsa-miR-132	**hsa-miR-212**	*RICS*	×	×	-	-	√	√	-	-	-	-
hsa-miR-141	**hsa-miR-200c**	*Clock*	×	×	√	√	×	×	√	×	-22.1	-20.1

from the same family which are supposed to have the same targets, and 5 new pairs which have one well-studied miRNA and one functional unknown partner. Therefore, we induce the targets for these 5 unknown miRNAs hsa-miR-18a/ 18b/ 20b /212 /200c from their known partner. Their predicted targets are listed in Table 3.

Informatic validation is performed to check the prediction consistency with the existing methods. Table 3 shows validation for the 3 confirmed and 5 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. For example the table can be read as following: 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, Retinoblastoma 1 (RB1) for hsa-miR-20b are predicted by TargetScan and Pictar; Circadian Locomoter Output Cycles Kaput (Clock) for hsa-miR-200c is 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.

5 Conclusions and discussion

Machine learning is widely used in commercial businesses which produce vast amounts of data. The life-sciences, molecular oriented research in particular, is a rapidly growing field which has gained a lot of attention lately especially now that the genomes of the major research model species have been sequenced and are publicly available. With the development of more and more large-scale and advanced techniques in biology, the need to discover hidden information triggered the application of machine learning in the field of the life-sciences. But these applications bear a risk, since, first of all, because most biological mechanisms are not yet fully understood, and second, some techniques produce too little experimental data due to the limitations of these techniques, thereby making machine learning unreliable. In this chapter, we explained how we integrated different machine learning algorithms and tuned and optimized experimental setups to a growing but not yet mature research field, miRNA target prediction. The innovation of this approach is not only integration and optimization of machine learning algorithms, but also the prediction through new features in miRNA relationship instead of widely studied features of miRNA-target interaction. Existing methods for analysis have shown to be insufficient in identifying targets from this perspective.

As illustrated in the methods and results sections, pattern recognition generates models enabling class descriptions. In this case, a rather high misclassification error around 30% is surfacing. In contrast, subgroup discovery aims at discovering statistically unusual patterns of interesting classes [36]. It discovers three main groups describing only the positive miRNA pairs.

One of the disadvantages of pattern recognition method is that the model is not biologically interpretable. Consisting of linear or quadratic transformations of features, the classifiers tell nothing about the mechanisms of miRNA-target binding. However decision tree and relational subgroup discovery are descriptive induction algorithms which discover patterns in the form of rules. With these discovered rules, we gain knowledge about miRNA-target interaction which can be used to predict more targets.

We compared two main algorithmic approaches used in knowledge discovery. Given the circumstances that not all the targets and useful features are known in advance, the classification of miRNA data using decision trees is not recommended. However, the

relational subgroup discovery, an advanced subgroup discovery algorithm, has shown to be suitable 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, seed and sequence similarities 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 [31]. 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 [31]. These observations are similar with the feature embedded in Rule 1 (cf. Section 4.2). Besides the fact that clustered miRNAs can be transcribed together, we further showed that miRNAs that are in close proximity to each other can bind to the same target so as to serve as the regulators for the same goal. In this study, we showed that the genomic location also contributes to miRNA target identification.

As for seed similarity, Rule 2 (cf. Section 4.2) describes that the miRNAs with seed similarity above 71.4% 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 a nearly perfect match with the target [15]. Moreover, TargetScanS 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 can be considered reasonable.

Overall sequence similarity is also a predictor but not as decisive as seed and genomic distance. This means that not only the seed region is important; sometimes two miRNAs with generally similar sequences can also bind to the same target. This is consistent with the finding that some miRNA-target interaction bindings have a mismatch or wobble in the 5' seed region but compensate through excellent complementarity at the 3' end, which leads to high average sequence complementarity [23].

In order to support our findings, we validated the results using five existing algorithms presented in Table 3. 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. Most of the candidates are predicted by at least one of these methods. Both miTarget and our method are based on machine learning techniques; miTarget uses a support vector machine and considers sequence and structure features of miRNA-target duplexes whereas we focus the integration of several machine learning algorithms 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 [6]. In the application of RNAhybrid tool, a predefined threshold of the normalized minimum free energy (mfe) is lacking, we therefore decided to list the original values. We found that most of our predicted miRNA-target duplexes are more stable illustrated by the lower minimum free energy relative to 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, seed and sequence similarities, more features need to be included in order to find more and better patterns shared by functionally alike miRNAs. *Grimson et al.* uncovered five general features of target site context beyond seed pairing that boost site efficacy [11]. In future research we will explore the site context in the miRNA relationship analysis. Additionally, we also consider taking into account miRNA co-expression patterns.

In summary, we conclude that genomic distance, seed and sequence similarities are the determinants for describing the relationships of functionally similar miRNAs. Our method is complementary to the approaches that are currently used. It 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. The methodology can be transferred to a broad range of other species as well.

Acknowledgements

We would like to thank Dr. Erno Vreugdenhil for discussing some biological implications of the results and Peter van de Putten for suggestions on the use of WEKA. This research has been partially supported by the BioRange program of the Netherlands BioInformatics Centre (BSIK grant).

References

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

[2] I. Bentwich. Prediction and validation of microrna 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] C. Z. Chen. MicroRNAs as oncogenes and tumor suppressors. *N Engl J Med*, 353(17):1768–1771, October 2005.

[5] A. J. Enright and S. Griffiths-Jones. mirbase: a database of microrna sequences, targets and nomenclature. In Krishnarao Appasani, Sidney Altman, and Victor R. Ambros, editors, *MicroRNAs:From Basic Science to Disease Biology*, pages 157–171. Cambridge University Press, 2007.

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

[7] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, June 2006.

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

[9] 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.

[10] Sam Griffiths-Jones. The microRNA registry. *Nucleic acids research*, 32(Database issue), January 2004.

[11] Andrew Grimson, Kyle Kai-How K. Farh, Wendy K K. Johnston, Philip Garrett-Engele, Lee P P. Lim, and David P P. Bartel. Microrna targeting specificity in mammals: Determinants beyond seed pairing. *Mol Cell*, 27(1):91–105, July 2007.

[12] D. Grun and N. Rajewsky. Computational prediction of microrna targets in vertebrates, fruitflies and nematodes. In Krishnarao Appasani, Sidney Altman, and Victor R. Ambros, editors, *MicroRNAs:From Basic Science to Disease Biology*, pages 172–186. Cambridge University Press, 2007.

[13] Lin He, Michael M. Thomson, Michael T. Hemann, Eva Hernando-Monge, David Mu, Summer Goodson, Scott Powers, Carlos Cordon-Cardo, Scott W. Lowe, Gregory J. Hannon, and Scott M. Hammond. A microrna polycistron as a potential human oncogene. *Nature*, 435(7043):828–833, 2005.

[14] 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.

[15] Fedor V. Karginov, Cecilia Conaco, Zhenyu Xuan, Bryan H. Schmidt, Joel S. Parker, Gail Mandel, and Gregory J. Hannon. A biochemical approach to identifying microrna targets. *Proceedings of the National Academy of Sciences*, pages 19291–19296, November 2007.

[16] Sung-Kyu Kim, Jin-Wu Nam, Je-Keun Rhee, Wha-Jin Lee, and Byoung-Tak Zhang. mitarget: microrna target-gene prediction using a support vector machine. *BMC Bioinformatics*, 7:411+, September 2006.

[17] Azra Krek, Dominic Grün, Matthew N. Poy, Rachel Wolf, Lauren Rosenberg, Eric J. Epstein, Philip Macmenamin, Isabelle d. da Piedade, Kristin C. Gunsalus, Markus Stoffel, and Nikolaus Rajewsky. Combinatorial microrna target predictions. *Nature Genetics*, 37(5):495–500, April 2005.

[18] Nada Lavrac, Filip Zelezny, and Peter A/ Flach. Rsd: Relational subgroup discovery through first-order feature construction. In *Proceedings of the 12th International Conference on Inductive Logic Programming*, pages 149–165. Springer-Verlag, July 2003.

[19] CH 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.

[20] R. C. Lee, R. L. Feinbaum, and V. Ambros. The c. elegans heterochronic gene lin-4 encodes small rnas with antisense complementarity to lin-14. *Cell*, 75(5):843–854, December 1993.

[21] 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.

[22] Benjamin P. Lewis, I-Hung Shih, Matthew W. Jones-Rhoades, David P. Bartel, and Christopher B. Burge. Prediction of mammalian microrna targets. *Cell*, 115(7):787–798, December 2003.

[23] Pierre Mazière and Anton J. Enright. Prediction of microRNA targets. *Drug discovery today*, 12(11-12):452–458, June 2007.

[24] Giorgos L. Papadopoulos, Martin Reczko, Victor A. Simossis, Praveen Sethupathy, and Artemis G. Hatzigeorgiou. The database of experimentally supported targets: a functional update of TarBase. *Nucleic acids research*, 37(Database issue):gkn809+, January 2009.

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

[26] Brenda J. Reinhart, Frank J. Slack, Michael Basson, Amy E. Pasquinelli, Jill C. Bettinger, Ann E. Rougvie, Robert H. Horvitz, and Gary Ruvkun. The 21-nucleotide let-7 rna regulates developmental timing in caenorhabditis elegans. *Nature*, 403(6772):901–906, February 2000.

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

[28] P. Sethupathy, M. Megraw, and A. G. Hatzigeorgiou. Computational approaches to elucidate mirna biology. In Krishnarao Appasani, Sidney Altman, and Victor R. Ambros, editors, *MicroRNAs:From Basic Science to Disease Biology*, pages 187–198. Cambridge University Press, 2007.

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

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

[31] Andrea Tanzer and Peter F. Stadler. Molecular evolution of a MicroRNA cluster. *Journal of Molecular Biology*, 339(2):327–335, May 2004.

[32] Ferdinand van der Heijden, Robert Duin, Dick de Ridder, and David M. J. Tax. *Classification, Parameter Estimation and State Estimation: An Engineering Approach Using MATLAB*. Wiley, 1 edition, November 2004.

[33] Andrew R. Webb. *Statistical Pattern Recognition, 2nd Edition*. John Wiley & Sons, October 2002.

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

[35] S. Wuchty, W. Fontana, I. L. Hofacker, and P. Schuster. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, 49(2):145–165, February 1999.

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

[37] Yanju Zhang, J. S. de Bruin, and Fons J. Verbeek. mirna target prediction through mining of mirna relationships. *BioInformatics and BioEngineering*, pages 1–6, 2008.

[38] Yanju Zhang, Joost M. Woltering, and Fons 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.