



Universiteit  
Leiden  
The Netherlands

## **Heterogeneous data analysis for annotation of microRNAs and novel genome assembly**

Zhang, Y.

### **Citation**

Zhang, Y. (2011, November 24). *Heterogeneous data analysis for annotation of microRNAs and novel genome assembly*. Retrieved from <https://hdl.handle.net/1887/18145>

Version: Not Applicable (or Unknown)

License: [Leiden University Non-exclusive license](#)

Downloaded from: <https://hdl.handle.net/1887/18145>

**Note:** To cite this publication please use the final published version (if applicable).

# Chapter 4

## **Comparison and Integration of Target Prediction Algorithms for microRNA Studies**

*Based on*

*Yanju Zhang and Fons J. Verbeek. (2010). Comparison and Integration of Target Prediction Algorithms for microRNA Studies. Journal of Integrative Bioinformatics, 7(3):127.*

### **Summary**

microRNAs are short RNA fragments that have the capacity of regulating hundreds of target gene expression. Currently, due to lack of high-throughput experimental methods for miRNA target identification, a collection of computational target prediction approaches have been developed. However, these approaches deal with different features or factors are weighted differently resulting in diverse range of predictions. The prediction accuracy remains uncertain. In this chapter, three commonly used target prediction algorithms are evaluated and further integrated using algorithm combination, ranking aggregation and Bayesian Network classification. Our results revealed that each individual prediction algorithm displays its advantages as was shown on different test data sets. Among different integration strategies, the application of Bayesian Network classifier on the features calculated from multiple prediction methods significantly improved target prediction accuracy.

## 1 Introduction

microRNAs (miRNAs) are a class of novel post-transcriptional gene expression regulators which are involved in a variety of developmental, physiological or disease-associated cellular processes [1]. They bind to their targets, messenger RNAs (mRNAs). This binding marks the targets for degradation or translation inhibition [3, 2]. In the past years, around 500 miRNA genes have been discovered in human. Functional annotations are, however, available only for a small fraction of these miRNAs [5]. This fact leaves the mechanism of miRNA-mediated gene regulation largely unknown.

One crucial aspect of the functional annotation of miRNAs is the identification of the miRNA targets with which they directly interact [15]. Due to the limitations of the current techniques, high-throughput target validation via biological experiments is not practical. Given these circumstances, a lot of algorithms for computational target prediction have been developed.

Each algorithm has a key focus. On the basis of this the prediction algorithm can be categorized into three groups: i.e. sequence-based, energy-based and machine learning-based groups. In the first group, the degree of sequence complementarity is considered as primary. This principle is used in e.g. miRanda [3] and TargetScanS [9]. miRanda first calculates sequence complementarity score with a weighting scheme; TargetScanS mainly takes the complementarity of the seed region, i.e. nucleotides 2-7, of miRNAs into account. Algorithms in the second group utilize thermodynamics as the main criterion. RNAhybrid [19] belongs to this group. It predicts the hybridization sites that are energetically most favourable as the binding sites. In the third group, algorithms such as NBmiRTar [26], miTarget [8] and *Zhang et al.* [27] collect different types of features and utilize machine learning techniques to find the feature patterns shared by true miRNA-target interactions.

Recently, the number of miRNA target prediction algorithms has been significantly increased. They do facilitate target identification, however, so far none of them could capture all true targets. Moreover, these computational approaches differ in algorithmic style; i.e., they use various features or factors are weighted differently. The lack of systematic verification and justification on the algorithms leaves the prediction accuracy and consistency unclear. To that end, generating a common criterion and test sets to analyze their

prediction performance and then integrating these algorithms to improve prediction accuracy will be very beneficial.

*Lin et al.* [11] mentioned that data integration can, in general, be approached from two routes; the “low-level” which deals with multi-factorial raw data directly and the “high-level” which combines multiple same type results from different studies.

In this study, we evaluate the performance of different target prediction algorithms and use integration methods to improve prediction accuracy. Both high-level integration approaches, e.g. algorithm combinations and ranking aggregation and low-level integration approach, e.g. the application of Bayesian Network classification, are performed. All of the methods are tested against miRNA-target interactions that are experimentally supported and several compiled negative control data sets. Our methods revealed that the system performance measured by the product of sensitivity and specificity provides a good criterion for algorithm comparisons. Algorithms categorized in the same group have similar prediction patterns. Algorithms categorized in different groups demonstrate their own advantages on different data sets. We inspected on the characteristics of miRNA-target site interactions and discovered that miRNAs have binding preference at the end of their target. We utilized three different integration strategies and demonstrated that the Bayesian Network classification results in best prediction accuracy.

## 2 Materials and methods

### 2.1 Materials

In the past years, the number of validated miRNA targets has been increased. Tarbase is a comprehensive repository recording a collection of experimentally supported miRNA targets in animal species, plants and viruses [15, 20]. The latest version, Tarbase 5.0, extracts data from a total of 203 scientific papers resulting in 1333 experimentally supported miRNA target gene interactions. For each interaction, it also provides direct evidences, such as reporter gene assay, and/or indirect experiment evidences such as microarray.

In this study we focused on human miRNAs as, to date, these are the best studied and also a large number of experimentally validated targets is available. To that end, a collection of 1093 experimentally confirmed human miRNA-target interactions from Tarbase is down-

| True set                                                 | False sets                                                |
|----------------------------------------------------------|-----------------------------------------------------------|
| True: miRNAs and 3'UTR of true target interactions (157) | False_ori: miRNAs and 3'UTR of false targets (28)         |
|                                                          | Shuffled: miRNAs and shuffled 3'UTR of true targets (157) |
|                                                          | Coding: miRNAs and coding region of true targets (157)    |

**Table 1: Data sets**

loaded. Only the direct interactions, which have the strongest experimental evidence, have been selected.

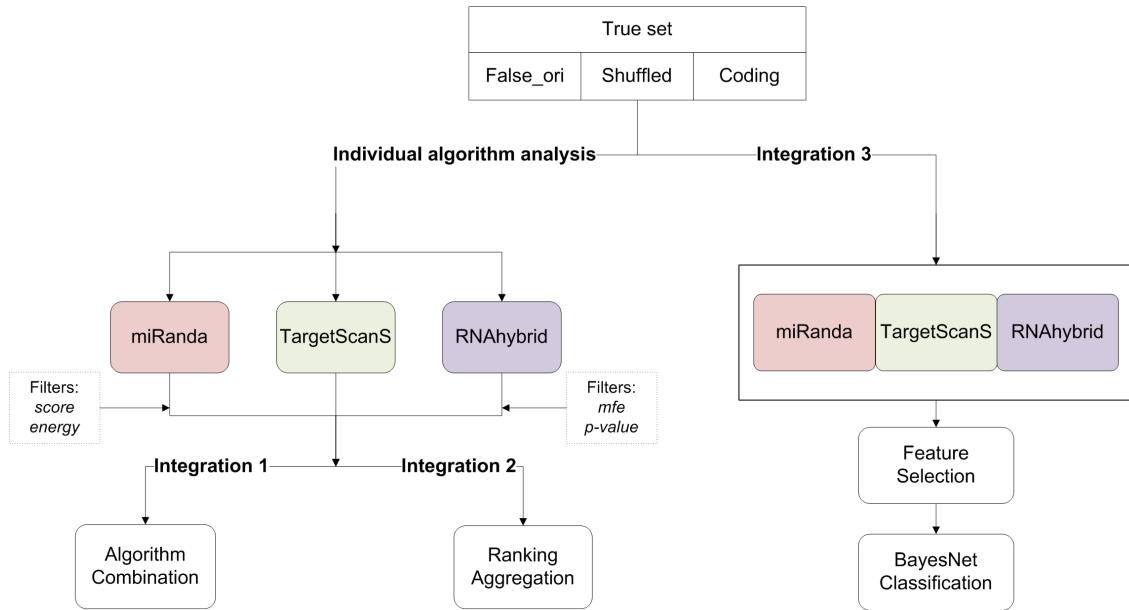
**True and False\_ori sets.** In further inspection of initial data, we found 5 interactions as ambiguous since they are reported as both true and false based on different forms of evidence. After removing redundant, ambiguous and sequence unavailable entries, finally, 157 and 28 miRNA-target gene interactions are kept to serve as true and false examples, respectively. We refer to the original false targets as false\_ori set. A small number of false samples is insufficient for data mining algorithms, and therefore, we compiled two more false sets. All data sets are listed in Table 1.

**Shuffled set.** Using 3'UTR sequences of 157 true targets as templates, we randomly shuffled the order of the nucleotides in these sequences, i.e. the frequencies of the nucleotides A, C, G and U are the same as in the original true target sequences; the order of the nucleotides, however, is random. In our experiments, this data set is registered as the shuffled set. In the analysis stage, we shuffle the sequences 20 times resulting in 20 sets of the random strings which are analyzed individually and over which the averages are computed.

**Coding set.** It has been established that miRNAs tend to bind their targets at the 3' UTR region. Given this feature, coding sequences are not supposed to contain binding sites. Therefore they are potential false target sequences. For our experiments, we used the coding sequences of the 157 true targets as a negative set.

## 2.2 Definition of predicted miRNA-target interactions

Tarbase provides experimentally validated miRNA-target gene interactions. However, computational algorithms predict putative binding sites, also referred to as predicted miRNA-target site interaction where miRNAs and their target mRNAs interact. In order to connect



**Figure 1: Schema. True dataset together with three false sets are processed through individual and integrated algorithm analysis.**

experimental and computational results, we define predicted miRNA-target gene interactions as follows. A miRNA-target gene interaction is predicted only if there is at least one binding site where this miRNA interacts with at least one of the transcripts of this gene. The scale of the interactions increases in a sequence of miRNA-target site interactions, miRNA-target mRNA interactions and miRNA-target gene interactions.

## 2.3 Methods

In this study, we analyse the efficiency of three target prediction algorithms i.e. miRanda, TargetScanS and RNAhybrid and of a selection of integration strategies on these algorithms using multiple data sets. The motivation for choosing these three as the objects for comparison and integration is that they are the most frequently used target prediction algorithms and that they are open source which allow us to execute them locally and adapt them to different data sets and extract new self-defined features. In miRNA-target prediction, conservation is used but not always fully understood when applied over a multitude of distant species. Moreover, calculation of binding site conservation involves multiple sequence alignment over the multitude of species. This considerably contributes to computational load; in order to reduce this load, the conservation filter in each algorithm was

not applied in our experiments.

The computational procedure of our method is illustrated in Fig. 1. In the following paragraphs, we will briefly explain the components in the diagram followed by the experiment set-ups and how these algorithms are integrated.

**MiRanda.** miRanda [3] is one of the earliest developed large-scale target prediction algorithms for vertebrates. The standard version of miRanda selects target genes based on three properties: sequence complementarity using a position-weighted local alignment algorithm, free energies of RNA-RNA duplexes using the Vienna RNA fold package [25], and conservation of targets in related genomes. These features are weighed in a decreasing order. In this application, only the first two filtering layers, i.e. sequence and energy scores are applied to restrict the predictions.

**TargetScanS.** TargetScanS [9] is the new and simplified version of TargetScan [10] and it has a stronger emphasis on the seed region. In the standard version, the predicted target-sites require first a 6-nucleotide (nt) match to the seed region of miRNA, i.e., nucleotides 2-7; second, a binding site conservation in 5 genomes (human, mouse, rat, dog and chicken). Each binding site is associated with a site-type, which is either "1a" or "8mer" or "m8". In the application of local TargetScanS, only seed complementarity is required.

**RNAhybrid.** RNAhybrid [19] finds the energetically most favourable hybridization sites between miRNAs and their target mRNAs using integrated powerful statistical models. It takes candidate target sequences and a set of miRNAs and looks for energetically favourable binding sites. In our practice, we first apply the RNACalibrate tool to estimate distribution parameters, and then use the RNAhybrid tool to find the minimum free energy hybridization. The RNAeffective tool which calculates the effective numbers across species is not performed.

**Ranking aggregation.** Ranking aggregation is a strategy for optimization problems. In theory, it combines several individual ranked lists to produce a super list which will be as close as possible to all individual lists simultaneously. In our application, we use RankAggreg [17], an R package for weighted rank aggregation. It was illustrated by *Lin et al.* [11] that the utility of ranking aggregation leads to satisfactory simulation results when combining miRNA target lists from different algorithms. Further to these findings, we use

ranking aggregation as one integration option and test its performance. Our experimental set-up is using the tau distance function to measure distance and the Cross-Entropy Monte Carlo method [16] for aggregation.

**Feature selection and Bayesian Network.** Feature selection is the technique of selecting a subset of relevant features for building learning models. A Bayesian Network is a probabilistic model for classification. It is represented as a directed acyclic graph in which nodes represent attributes and edges represent conditional dependencies. The probability of any variable of a joint distribution can be calculated from conditional probabilities using the chain rules in probability theory [7]. This strategy is implemented in the Weka software environment [24]. CfsSubsetEval [6] and BayesNet [24] are applied for the purpose of feature subset selection and target classification. The error is estimated by 10-fold cross-validation.

### 2.3.1 Individual analysis

Running miRanda and RNAhybrid locally, one needs to decide cut-offs for several features. This is not the case for TargetScanS. The default settings of miRanda and RNAhybrid are to ensure detecting targets as much as possible. They also lead to many false positive predictions. In this case, we tune the parameters to find the best trade-off of true positive and false positive predictions. It is known that miRanda associates each predicted target site with a score which represents sequence complementarity degree between miRNA and its target as well as a free energy which measures the thermodynamics of the duplex. RNAhybrid predicts the targets with a minimum free energy (*mfe*) value and a *p*-value which represents the binding significance. The optimum cut-offs are those which achieve the performance (PERF) with highest combination of sensitivity (SENS) and specificity (SPEC), as defined by equations:

$$SENS = \frac{TP}{TP + FN} \quad (1)$$

$$SPEC = 1 - \frac{FP}{FP + TN} \quad (2)$$

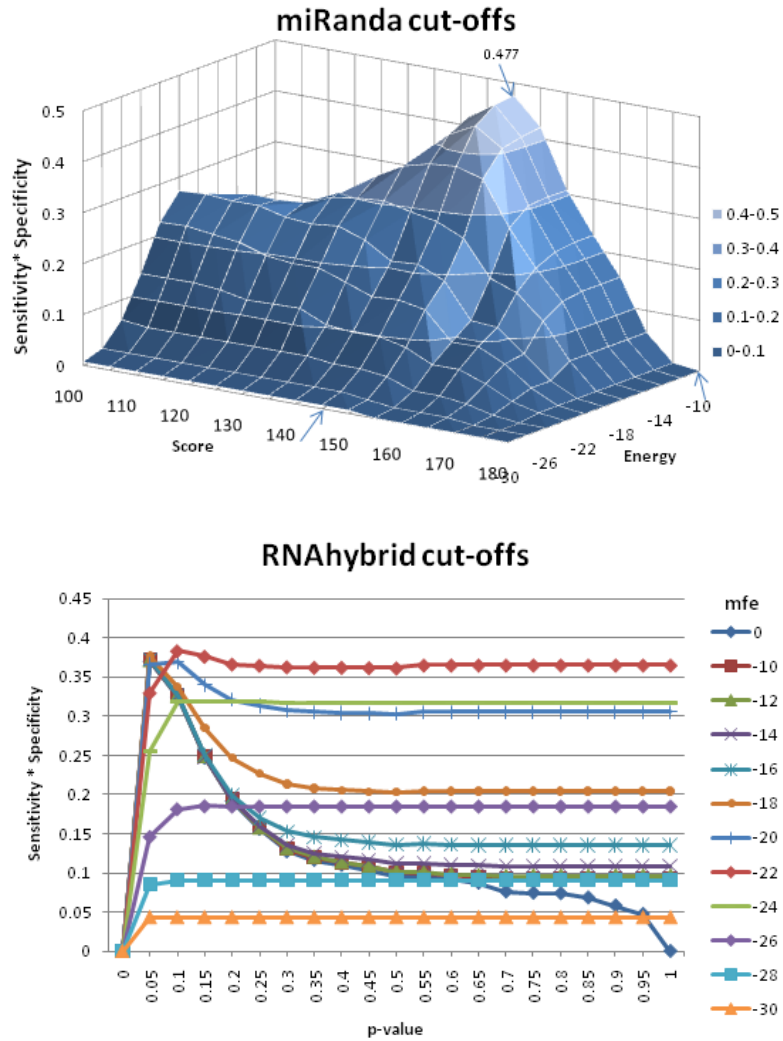
$$PERF = SENS \times SPEC \quad (3)$$

where TP, FN, TN and FP represent true positive, false negative, true negative and false positive respectively. Sensitivity is also referred as to the true positive rate (TPR) which is defined as the ratio of experimentally supported miRNA-target gene interactions predicted by an algorithm. Specificity is equal to 1- false positive rate (FPR) which is defined as the ratio of false miRNA-target gene interactions detected by an algorithm as being true. We define performance as the product of sensitivity and specificity as written in equation 3. This performance is used to optimize the parameters and serves as a common reference in comparing the different integration strategies.

For model comparison, several performance measures have been described. In machine learning, the area under ROC Curve (AUC) [4] is often applied. This number, however, does not give a clue for parameter optimization. Alternatively, accuracy (ACC) [22] or F1\_score [23] could be used and give similar results to our performance measure as they are derived from sensitivity and specificity as well. Our motivation to use the performance as given in equation 3 is that it reflects the requirement to the system to achieve both a high sensitivity and specificity. The value of performance is therefore logical and intuitive. And the performance differences are amplified at high values of two variables when comparing to the linear calculations.

### 2.3.2 Integration

After analysis of individual algorithms, three integration strategies are performed. The first is combining three individual approaches using various unions, intersections and majority vote. The second integration method is ranking aggregation which combines several ordered predicted target lists to generate a super list. In our practice, targets are first ranked according to the major feature of each prediction algorithms which are sequence score in miRanda, site-type in TargetScanS and *mfe* in RNAhybrid. After that, three ranked lists are integrated to a final list via a Cross-Entropy Monte Carlo method. The third integration approach is the application of a Bayesian network classifier. In our approach, the Bayesian network classifier is applied to the features measured by the individual target prediction approaches in order to discriminate two classes of targets. For each miRNA-target gene interaction a maximum and a minimum value of feature sets are registered. These features are (1) from miRanda: complementarity score, free energy, length of 3' UTR, relative binding position, number of hits; (2) from TargetScanS: site-type, length



**Figure 2: Filter optimizations for miRanda and RNAhybrid.** The combination of sequence score and free energy which achieves the best performance is set as thresholds for miRanda (left). The best combination of *mfe* and *p*-value is set as thresholds for RNAhybrid (right).

of 3' UTR, relative binding position, number of hits; (3) from RNAhybrid: *mfe*, *p*-value, length of 3' UTR, relative binding position. Considering both the maximum and the minimum values, there are 26 features in total. Subsequently, these features are then selected by feature selection and further classified by a Bayesian network.

|             | Cut-offs                                            | <i>SENS</i> | <i>SPEC</i> |          |        | $\overline{PERF}$ |
|-------------|-----------------------------------------------------|-------------|-------------|----------|--------|-------------------|
|             |                                                     | True        | False_ori   | Shuffled | Coding |                   |
| miRanda     | score $\geq$ 145<br>energy $\leq$ -10               | 0.662       | 0.429       | 0.72     | 0.822  | 0.414             |
| TargetScanS | -                                                   | 0.815       | 0.286       | 0.66     | 0.79   | 0.438             |
| RNAhybrid   | <i>mfe</i> $\leq$ -22<br><i>p</i> -value $\leq$ 0.1 | 0.578       | 0.454       | 0.654    | 0.629  | 0.313             |

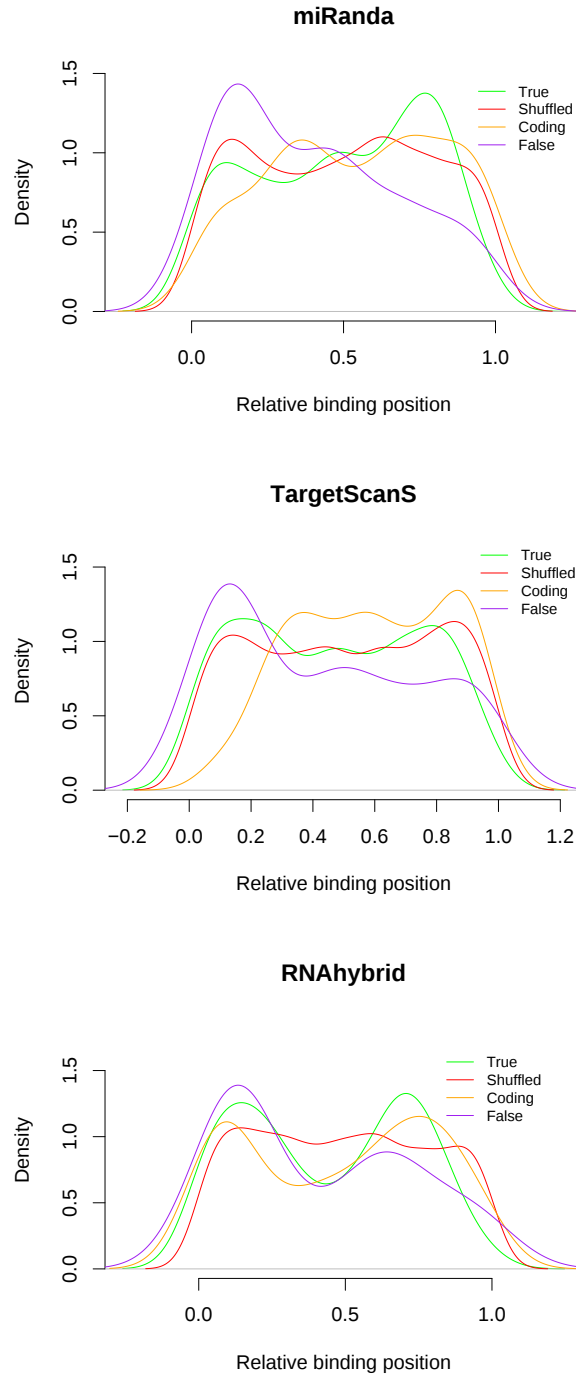
**Table 2: Performance of individual algorithms.** The last column shows the average performance over the different sets. In order to assure that the results can be compared to that in Table 3,4,5, the shuffled set is listed but not used in the calculation of the average performance in this table.

### 3 Results

Before comparing prediction accuracy, feature cut-offs of miRanda and RNAhybrid are optimized. In order to achieve this, miRanda sequence complementarity score is tested from 100 to 180 with step=5 and energy is set from -10 kcal/mol to -30 kcal/mol with step=-2 kcal/mol; In RNAhybrid, minimum free energy is tuned from -10 kcal/mol to -30 kcal/mol with step=-2 kcal/mol and the *p*-value is tuned from 0 to 1 with step=0.05. Fig. 2 shows the optimization results for miRanda and RNAhybrid tested on true and shuffled sets. On the left, it is a landscape plot of sequence score, energy and system performance for miRanda. As can be seen, score=145 and energy=-10 kcal/mol lead to best performance represented by sensitivity \* specificity =0.477. On the right, the graph shows the relationships between *mfe*, *p*-value and system performance. Free energy is represented by each line. X-axis shows the *p*-value changing from 0 to 1. Performance is depicted in y-axis. Further inspecting the graph, we found that when *mfe*=-22 kcal/mol, the system has very good performance overall. The system reaches the highest performance when *mfe*=-22 kcal/mol and *p*-value=0.1.

#### 3.1 Individual performance

After feature optimization, a peak performance for miRanda and RNAhybrid could be accomplished. Subsequently, three algorithms are compared. The average performance for each of the individual is summarized in Table 2. We found that TargetScanS has the high-



**Figure 3: Characteristics of relative binding position in miRanda (left), TargetScanS (middle) and RNAhybrid (right). The density plot of relative target-binding position of true, false.ori, shuffled and coding sets are depicted in green, purple, red and orange respectively.**

est sensitivity; miRanda has the highest specificity when testing on shuffled and coding sequences; RNAhybrid has the highest specificity when testing on validated false target set. Moreover, we observed that miRanda and TargetScanS have similar patterns on different data sets. The specificity on coding set drop around 10% and 40-50% when comparing to that of shuffled and false\_ori set, respectively. RNAhybrid, however, did not follow this pattern. A possible reason for this is that miRanda and TargetScanS are sequence-based algorithms which respond similarly on different types of sequences; whereas RNAhybrid is energy-based. In general, all three exhibit either a relative low specificity or/and sensitivity indicating that their prediction accuracy cannot yet be considered satisfactory.

In addition, we found one interesting target-binding site feature that is consistently displayed in three methods. Fig. 3 shows the distribution of relative binding position of each data set predicted by each method. The densities are estimated with a Gaussian kernel using R stats package [18]. Relative position is calculated as the position of a binding site divided by the length of target sequence. In the true data set, the predicted target sites in miRanda have location bias at the end; they have slightly a higher density at the two ends of the sequences in TargetScanS. This two ends binding preference is more obvious from RNAhybrid. In contrast, the target binding sites of false\_ori set appeared more often at the beginning. While, the shuffled set shows nearly the uniform distribution. In summary, we conclude that the potential true target sites are enriched at the end of the binding sequences. The reason is probably that the binding sites are close to polyA tails which are the known factor effecting translation efficiency [5, 14].

### 3.2 Integration 1: combination

Our first strategy for integration is combining individual algorithms through various unions and intersections. The average performance of each combination over different sets is displayed in Table 3. It can be seen that majority vote is the best combination strategy, since it has the highest prediction performance. It is also higher than that of each individual algorithm. In the intersection part, we observed that targets predicted by miRanda and TargetScanS have higher degree of overlap than the other intersections. This is also because both of them weigh sequence complementarity as a main factor in their algorithms. We also suggest that, for the study of finding the networks involved by all the targets of a miRNA, using the union of these three is an option. This solution will cover a high

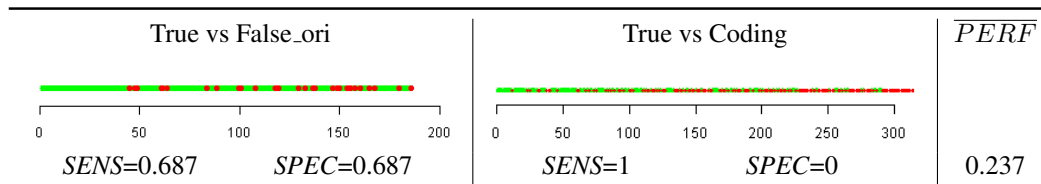
|                                 | <i>SENS</i> | <i>SPEC</i> |        | $\overline{PERF}$ |
|---------------------------------|-------------|-------------|--------|-------------------|
|                                 | True        | False_ori   | Coding |                   |
| <b>Unions</b>                   |             |             |        |                   |
| miRanda, TargetScanS, RNAhybrid | 0.879       | 0.179       | 0.573  | 0.331             |
| <b>Intersections</b>            |             |             |        |                   |
| miRanda, TargetScanS, RNAhybrid | 0.452       | 0.607       | 0.879  | 0.336             |
| miRanda, TargetScanS            | 0.643       | 0.464       | 0.866  | 0.428             |
| miRanda, RNAhybrid              | 0.471       | 0.571       | 0.841  | 0.333             |
| TargetScanS, RNAhybrid          | 0.516       | 0.536       | 0.841  | 0.335             |
| <b>Majority Vote</b>            |             |             |        |                   |
| miRanda, TargetScanS, RNAhybrid | 0.79        | 0.357       | 0.79   | 0.453             |

**Table 3: Performance of various unions and intersections of the individual algorithms.**

range of true targets, as a trade-off, it will also cover a large number of false targets. This high false prediction rate can be reduced by further functional annotation analysis, e.g. targets can be further screened according to annotations with pathway, disease and gene ontologies.

### 3.3 Integration 2: ranking aggregation

Three ranked target lists from miRanda, TargetScanS and RNAhybrid are generated by sorting sequence score, binding site-type and energy respectively. For the miRNA-target gene interaction with multiple binding sites, the best values are selected to represent the whole interaction, i.e. highest sequence score, stringent binding site and lowest minimum free energy. After that, three lists are integrated to one via the RankAggreg package. The symbolic results are displayed in Table 4. Ranking for top to the end are displayed in a



**Table 4: Performance of ranking aggregation and symbolic plots of ranked lists. In the plot, true and negative targets are displayed in green and red respectively. Axis shows the ranking index.**

| Features  |         |             |                         | False_ori   |             | Coding      |             | BayesNet          |
|-----------|---------|-------------|-------------------------|-------------|-------------|-------------|-------------|-------------------|
| RNAhybrid | miRanda | TargetScanS |                         | <i>SENS</i> | <i>SPEC</i> | <i>SENS</i> | <i>SPEC</i> | $\overline{PERF}$ |
| 0         | 0       | 1           | Feature Selection       | 1           | 0           | 0.815       | 0.79        | 0.322             |
| 0         | 1       | 0           |                         | 1           | 0           | 0.809       | 0.611       | 0.247             |
| 1         | 0       | 0           | &                       | 0.917       | 0.643       | 0.828       | 0.49        | 0.498             |
| 0         | 1       | 1           | BayesNet Classification | 1           | 0           | 0.815       | 0.822       | 0.335             |
| 1         | 0       | 1           |                         | 0.987       | 0.607       | 0.815       | 0.809       | 0.629             |
| 1         | 1       | 0           | →                       | 0.987       | 0.607       | 0.834       | 0.739       | 0.608             |
| 1         | 1       | 1           |                         | 0.987       | 0.607       | 0.815       | 0.815       | 0.632             |

**Table 5: Performance of Bayesian Network classification on different features. In the Features column, 1 represents that the features from this algorithm are selected for the machine learning.**

direction from left to right. Green represents true targets; red represents false targets of coding (on the right) and false\_ori sets (on the left) respectively. The average performance value is 0.237 indicating that ranking aggregation cannot precisely detect the true targets. A conceivable explanation is that the majority of true miRNA target does not always have very high sequence complementarity or has low free energy scores. Therefore, they are not always found in the top ranking list when using the key factor exclusively as the ranking criterion.

### 3.4 Integration 3: Bayesian Network classification

Feature sets are first processed through a feature selection procedure and then classified by a Bayesian Network. Their average performances are listed in Table 5. It shows that discriminating true and false targets based on the features from all different algorithms achieves the best performance. Furthermore, the classification performance on the features from RNAhybrid together with either miRanda or TargetScanS is also relatively high. This indicates that features from miRanda and TargetScanS are highly correlated and therefore could be redundant. Evaluation using this machine learning approach puts RNAhybrid as the best algorithm of the three individuals. As a comparison to the integration method 1 and 2, the Bayesian Network classification method based on the features from all three algorithms results the best overall performance and therefore can be con-

sidered as an optimal integration strategy.

## 4 Conclusions and Discussion

The increasing interest in miRNA regulatory function triggered the development of many computational approaches for miRNA target prediction. However, the large amount of approaches and the low degree of prediction overlap between them might leave the users confused. In this study, we demonstrated that the performance of current target prediction algorithms is by no means perfect. However, a proper integration of these prediction algorithms can significantly improve the prediction accuracy.

One of our contributions to the study of miRNA is to measure the performance of miRNA target prediction algorithms using both the true-positive and false-positive rate. Measuring target prediction performance has been recently addressed in few literature reviews. Most of these reviews compared target prediction approaches either from algorithmic point of view [1, 13], or using the estimated false positive rates [12] or using small numbers of experimentally validated miRNA targets [21]. However, using only false positive or true positive rates is not sufficient to indicate the prediction performance.

In our method, we generated the negative sets in a different way compared to previous studies. Current research focuses on finding the true targets, and consequently, only a small number of false targets are identified as by-products. This complicates calculation of false positive rates. The most common way to generate negative data set is sequence-shuffling. Besides that, we also used coding sequences as potential negative classes since most binding sites are not located in this region. Interestingly, error rates approximated on different type of negative sets have similar patterns. False positive rates on coding set are smaller than those on random set in general; while false positive rates on 3' UTR of real false target set are larger than those on random set. This indicates that all three prediction algorithms predict relatively more binding sites at 3'UTR.

The challenge of integration is to combine available data in a proper and efficient manner. In this chapter, we present three ways to integrate miRNA prediction algorithms. Algorithm combination and ranking aggregation are high-level integration methods, and application of a Bayesian Network classifier to the features measured by multiple prediction methods is a novel low-level integration method. Testing on common data sets, in-

tegration through Bayesian Network significantly improves prediction performance. This proves that, although high-level integration methods are easy and direct to apply, they lose information as not all data is passed to the integration stages. Moreover, low-level analysis which models raw data from different sources is complicated but, on the other hand, higher accuracy can be achieved. We also chose the proper classifier. *Yousef et al.* used Naive Bayes to classify targets [26]. The Naive Bayes classifier is based on the assumption of strong independence between the features. In our case, we found the Bayesian Network classifier did outperform the Naive Bayes since some of the features are not independent. In the future, for the functionally unknown miRNAs, of which the targets are unclear, we suggest the application of Bayesian Network classifier for target prediction.

From our computational analysis, we discovered one significant feature of miRNA target interaction. We observed that miRNAs have potential binding preference at the end of the target sequences. In paper [5], it is claimed that miRNAs have location bias at the beginning and at the end of 3' UTR. We found similar patterns. However, after further inspection of these patterns, we also observed many false targets at the beginning of 3' UTR.

Although Tarbase is a valuable resource for machine learning algorithms, the number of validated true targets and especially validated false targets is too small. We expect that with more validated targets available, the prediction accuracy of our proposed integration methods using Bayesian Network classification will increase. More improvement can be achieved by including more relative features such as binding site conservation. One of our further research will direct towards categorization of miRNA-target interaction to subtypes: once the target is validated, it is interesting to understand and establish whether it is the target for degradation or translation inhibition.

## Acknowledgements

We thank Amalia Kallergi, Joris Slob and Kuan Yan for their critical discussions on methodology and results. This research has been partially supported by the BioRange program of the Netherlands Bioinformatics Centre (NBIC, BSIK grant).

## References

- [1] Christian Barbato, Ivan Arisi, Marcos Frizzo, Rossella Brandi, Letizia Da Sacco, and Andrea Masotti. Computational challenges in mirna target predictions: to be or not to be a true target? *Journal of biomedicine & biotechnology*, 2009.
- [2] David P. Bartel. Micrnas: Genomics, biogenesis, mechanism, and function. *Cell*, 116(2):281–297, Jan 2004. [http://dx.doi.org/10.1016/S0092-8674\(04\)00045-5](http://dx.doi.org/10.1016/S0092-8674(04)00045-5).
- [3] A. J. Enright, B. John, U. Gaul, T. Tuschl, C. Sander, and D. S. Marks. Micrna targets in drosophila. *Genome Biol*, 5(1), 2003. 1465-6914.
- [4] Tom Fawcett. An introduction to roc analysis. *Pattern Recogn. Lett.*, 27(8):861–874, June 2006.
- [5] Dimos Gaidatzis, Erik van Nimwegen, Jean Hausser, and Mihaela Zavolan. Inference of mirna targets using evolutionary conservation and pathway analysis. *BMC bioinformatics*, 8:69, Mar 2007.
- [6] M. Hall. *Correlation-based Feature Selection for Machine Learning*. PhD thesis, 1998.
- [7] D. Heckerman, D. Geiger, and D. M. Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20(3):197–243, 1995.
- [8] Sung Kyu Kim, Jin Wu Nam, Je Keun Rhee, Wha Jin Lee, and Byoung Tak Zhang. mitarget: micrna target-gene prediction using a support vector machine. *BMC Bioinformatics*, 7, Oct 2006. 1471-2105.
- [9] Benjamin Lewis, Christopher Burge, and David Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are micrna targets. *Cell*, 120(1):15–20, Jan 2005.
- [10] Benjamin P. Lewis, I. Hung Shih, Matthew W. Jones-Rhoades, David P. Bartel, and Christopher B. Burge. Prediction of mammalian micrna targets. *Cell*, 115(7):787–798, Dec 2003. [http://dx.doi.org/10.1016/S0092-8674\(03\)01018-3](http://dx.doi.org/10.1016/S0092-8674(03)01018-3).
- [11] Shili Lin and Jie Ding. Integration of ranked lists via cross entropy monte carlo with applications to mrna and micrna studies. *Biometrics*, May 2009.
- [12] G. Martin. Prediction and validation of micrna targets in animal genomes. *J Biosci*, 32(6):1049–1052, Oct 2007.
- [13] P. Maziere and A. J. Enright. Prediction of micrna targets. *Drug Discov Today*, 12(11-12):452–458, Jun 2007. 1359-6446.
- [14] B. Mazumder, V. Seshadri, and P. L. Fox. Translational control by the 3'-utr: the ends specify the means. *Trends in biochemical sciences*, 28(2):91–98, Feb 2003.
- [15] 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):D155–D158, 2008. <http://dx.doi.org/10.1093/nar/gkn809>.
- [16] V. Pihur and S. Datta. Weighted rank aggregation of cluster validation measures: a monte carlo cross-entropy approach. *Bioinformatics*, 23(13):1607–1615, Jul 2007.
- [17] Vasyl Pihur, Susmita Datta, and Somnath Datta. Rankagg, an r package for weighted rank aggregation. *BMC Bioinformatics*, 10(1):62, 2009.
- [18] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.
- [19] M. Rehmsmeier, P. Steffen, M. Hochsmann, and R. Giegerich. Fast and effective prediction of microrna/target duplexes. *RNA*, 10(10):1507–1517, 2004. 1355-8382.
- [20] Praveen Sethupathy, Benoit Corda, and Artemis G. Hatzigeorgiou. Tarbase: A comprehensive database of experimentally supported animal microrna targets. *RNA*, 12(2):192–197, Feb 2006. <http://dx.doi.org/10.1261/rna.2239606>.
- [21] 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, 2006. 1548-7091.
- [22] Wikipedia. Accuracy and precision—Wikipedia, the free encyclopedia. [Online].
- [23] Wikipedia. F1 score — Wikipedia, the free encyclopedia. [Online].
- [24] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 1999.
- [25] S. Wuchty, Fontana W., Hofacker I. L., and Schuster P. Complete suboptimal folding of rna and the stability of secondary structures. *Biopolymers*, 49:145–165, 1999.
- [26] Malik Yousef, Segun Jung, Andrew V. Kossenkov, Louise C. Showe, and Michael K. Showe. Naive bayes for microrna target predictions machine learning for microrna targets. *Bioinformatics*, Oct 2007.
- [27] Yanju Zhang, Jeroen S. de Bruin, and Fons J. Verbeek. mirna target prediction through mining of mirna relationships. *Bioinformatics and BioEngineering*, pages 1–6, Jul 2008.

