BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 22 2007, pages 2987–2992
doi:10.1093/bioinformatics/btm484

Sequence analysis

Naïve Bayes for microRNA target predictions—machine learning for microRNA targets

Malik Yousef, Segun Jung†, Andrew V. Kossenkov, Louise C. Showe and Michael K. Showe*

The Wistar Institute, Philadelphia, PA 19104, USA

Received on July 16, 2007; revised on September 11, 2007; accepted on September 14, 2007

Advance Access publication October 8, 2007

Associate Editor: Limsoon Wong

ABSTRACT

Motivation: Most computational methodologies for miRNA:mRNA target gene prediction use the seed segment of the miRNA and require cross-species sequence conservation in this region of the mRNA target. Methods that do not rely on conservation generate numbers of predictions, which are too large to validate. We describe a target prediction method (NBmiRTar) that does not require sequence conservation, using instead, machine learning by a naïve Bayes classifier. It generates a model from sequence and miRNA:mRNA duplex information from validated targets and artificially generated negative examples. Both the 'seed' and 'out-seed' segments of the miRNA:mRNA duplex are used for target identification.

Results: The application of machine-learning techniques to the features we have used is a useful and general approach for microRNA target gene prediction. Our technique produces fewer false positive predictions and fewer target candidates to be tested. It exhibits higher sensitivity and specificity than algorithms that rely on conserved genomic regions to decrease false positive predictions.

Availability: The NBmiRTar program is available at http://wotan.wistar.upenn.edu/NBmiRTar/

Contact: yousef@wistar.org

Supplementary information: http://wotan.wistar.upenn.edu/NBmiRTar/

1 INTRODUCTION

MicroRNAs (miRNAs) are short (~22 nt) RNA molecules that either mark their target mRNAs for degradation or suppress their translation by binding to the 3′-untranslated region (3′UTR). Previous studies have suggested that the miRNA seed segment (Fig. 1) that includes 6–8 nt at the 5′ end of the mature miRNA sequence is very important in the selection of the target site. Thus most of the computational tools developed to identify mRNA target sequences depend heavily on the complementarity with the 'seed' sequence in the target. However, Brennecke *et al*. (2005) have recently suggested that the 3′ out-seed segment of the miRNA:mRNA duplex sequence can compensate for imperfect base pairing within the seed segment. To our knowledge, only one recent computational approach (Yan *et al*., 2007) has considered the contributions of the out-seed miRNA segment in target identification.

Several methods for the prediction of miRNA targets have emerged recently. These methods mainly use sequence complementarity, thermodynamic stability calculations and evolutionary conservation among species to determine the likelihood of a productive miRNA:mRNA duplex formation (Bartel, 2004; Lai, 2004). Using sequence conservation reduces false positive predictions but some less-conserved target sites may be missed. The dilemma that is posed is how to avoid rejection of less highly conserved target sites while reducing the very large numbers of predictions that will be found when seed region conservation in the target is not required.

In order to reduce the false positive predictions inherent in methods that heavily weight specific target sequence conservation, Lewis *et al*. (2005) developed TargetScanS. TargetScanS scores target sites based on the conservation of the target sequences between five genomes (human, mouse, rat, dog and chicken). SaeTrom *et al*. (2005) have developed TargetBoost, a machine-learning algorithm for miRNA target prediction using only sequence information to create weighted sequence motifs that capture the binding characteristics between miRNAs and their targets. The authors suggest that TargetBoost is stable and identifies more of the already verified true targets than do other existing algorithms.

Sung-Kyu *et al*. (2005), have reported the development of an algorithm using SVM with a RBF kernel, an approach with some similarities to the one we describe here. Our approach differs mainly from that of Sung-Kyu et al. in that it includes the contributions of a number of additional features we feel are important for more accurate target prediction. In addition, while artificial and random negative classes are generated by both approaches, the negative class we have generated for the current study differs from theirs. The best reported results of Sung-Kyu *et al*. (2005) were 0.921 as sensitivity and 0.833 as specificity. More recent work reported by Yan *et al*. (2007), also using a machine-learning approach, employs features extracted from the seed and out-seed segments. However, these features are different from those we have used in NBmiRTar. The best

*To whom correspondence should be addressed.
†Present address: The Sackler Institute of Biomedical Sciences, NYU, USA.

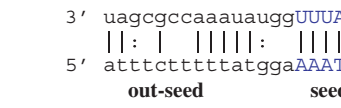

Fig. 1. Duplex partitioned into two parts, the seed and the out-seed, for miRNA hsa-miR-579 and its target LRIG3. The seed part is shown in capital letters.

results that were obtained by Yan *et al.* (2007) using an optimized ensemble classifier has an accuracy of 82.95%. The results were generated using only 48 positive human and 16 negative examples, a relatively small training set.

Recently Thadani and Tammi (2006) launched MicroTar, a statistical computational tool for prediction of miRNA targets from RNA duplexes, which does not use homology for prediction. MicroTar mainly relies on a quite novel approach to estimate the duplex energy. However, the reported sensitivity (60%) is significantly lower than the sensitivity achieved using other published algorithms.

Sethupathy *et al.* (2006b) conducted a survey and a comparison of the five most used tools for mammalian target prediction and indicated that 30% of the experimentally validated target sites are non-conserved, supporting the need for the development of different computational approaches to capture these target sites. Rajewsky (2006) in a similar review discusses the importance of selecting the so-called control set, which we refer to is as the negative class in our terminology and an issue we have attempted to address. Lai (2004) in an additional review noted that there is almost no overlap among the predicted targets identified by the various methods and suggests that each tool captures a subset of the entire target class as a function of the specific features they have incorporated into their prediction models. Furthermore, the large number of predictions that each of these tools is producing suggests that the heavy reliance on homology or comparative sequence analysis is not sufficient to generate accurate predictions with a high sensitivity.

We present here, a machine-learning approach for predicting miRNA target site based on the naïve Bayes (NB) classifier. In our approach, we include features extracted from both the seed and 'out seed' sequences, the duplex structures and a number of additional sequence features. We believe that the inclusion of these features (enumerated below) contributes significantly to improving the performance of the NBmiRTar classifier. We also suggest that the negative class we have generated for training is more appropriate than those generated for previous studies. One direct use of our classifier is as a filter for the output of the miRanda tool (Enright *et al.*, 2003; John *et al.*, 2004). We demonstrate that this filtering step decreases the false positive prediction by miRanda significantly. The NBmiRTar algorithm demonstrates both high specificity, and a high sensitivity.

2 MATERIALS AND METHODS

2.1 Data

A collection of 225 confirmed miRNA targets (human, mouse, fruit fly worm and zebrafish) and 38 confirmed false target predictions were downloaded from the TarBase (Sethupathy *et al.*, 2006a) web site to serve as positive and negative examples, respectively, for training the classifier. We will refer to this negative set as NEG_0. Since we anticipated that the current set of 38 confirmed negative examples would not be sufficient for the learning algorithm to function efficiently (see Supplementary Material Table A), additional artificial negative examples were generated as described below.

2.2 Generation of the artificial negative examples

We used the 3000 artificial mature miRNA (all 30 nt long) from SaeTrom *et al.* (2005) to generate an artificial negative class. These artificial mature miRNA consist of a random string of nucleosides appearing with frequencies of A, C, G and U with a probability of 0.34, 0.19, 0.18 and 0.29, respectively, that are not consistent with the base frequencies in true miRNAs. MiRanda was then used to generate target predictions for the 3000 artificial MiRNAs from the 29 785 human 3′UTR sequences in MiRanda. All these target results are assumed to be false positive predictions since the query search did not include true miRNAs. The minimum free energy (MFE) and the miRanda score threshold (SC) are two important parameters that one can set to increase the stringency of the predictions and thus decrease the selection of weaker false positive predictions (Hsu *et al.*, 2006). In this case, the artificial negative set was produced by setting the MFE at 25 kcal/mol and SC at 180, which are both stringent values. However, using all 3000 artificial negative examples yielded a large and unmanageable set of predictions. One hundred negative examples were then chosen at random from the 3000 and used to re-query miRanda. MiRanda now produced 133 316 false targets to form the pool of negative examples we have used. We refer to this pool as NEG_1.

2.3 Designing duplex structure and sequence features

Most studies agree that the seed segment is the most important factor in target selection anticipating a perfect or near perfect binding between the mature miRNA and its target. In most cases, the out-seed (3′) segment of the miRNA is not considered to influence target binding. However, an important early observation by Vella *et al.* (2004) indicated that the out-seed segment including various bulges can also have an important role in the miRNA–target interaction. It appears that a compensatory relation between the two segments exists (Brennecke *et al.*, 2005) so that a poorer homology between the miRNA and target seed regions can be compensated for by a stronger interaction in the out-seed. They also showed that out-seed with complementarity, is alone not sufficient to make a functional duplex; some contribution from the seed is required. The observations of Vella *et al.* (2004) as well as more recent studies of Brennecke *et al.* (2005) and Grimson *et al.* (2007) support the importance of the 'out-seed' region in target identification. The data and conclusions reported in these two papers have contributed to the feature design that we have used in this study, which includes both seed and out-seed regions of the miRNA sequence. Our features are chosen on the following assumptions: (1) The complementarity of 7–8 bases in the seed region are sufficient for good duplex formation. (2) A seed segment with weak complementarity can be compensated for by the out-seed sequence (3′ end of the miRNA) to make a functional duplex. (3) Good complementarity in the out-seed region alone is not sufficient for functional duplex formation.

To date there is no computational tool that builds a model capable of capturing even all known validated targets so the identification of new targets is problematic. We use machine learning based on miRNA features. Machine learning enables one to generate automatic rules based on observation of the appropriate examples by the learning machine. However, the selection and design of the features that will be considered in the learning process are very important and the parameters that distinguish the positive from negative classes need to

be carefully chosen. For example, if only features of the seed segment are considered then the learning machine will not be able to model cases where the target sites include a compensatory out-seed contribution. Consequently, we have partitioned the duplex into two parts, the seed (5′ 8 nt of the miRNA) and out-seed (3′ remainder) as described in Figure 1. For each of these parts the following features are extracted to give 57 structural features: (1) The number of paired bases (bp), (2) The number of bulges (inserts on one strand between paired bases), (3) The number of loops (unpaired bases opposite each other between paired bases), (4) The number of asymmetric loops (loops with unequal numbers of unpaired bases on the two strands), (5) Eight features, each representing the number of bulges of lengths 1–7 and those with lengths greater than 7, (6) Eight features, each representing the number of symmetric loops with lengths 1–7 and those with lengths >7, (7) Eight features each representing the number of asymmetric loops with lengths 1–7 and those with lengths >7 and (8) The distance from the start of the seed (the 3′ end) to the first paired base of the 5′ start of the out-seed part is an additional feature that is extracted. In addition, nucleotide sequence '*words*' with lengths 4, 5, 6, 7, 8, 9 are extracted from the *seed* sequence. These '*motif*' features are not fixed in number and influence the dimension of the vector space. Although no more than 15 unique words come from any one seed, each miRNA contributes a different set of words that depends on its sequence. The dimension of the feature vector is determined, at a later point, to be 57 plus the number of unique '*words*' that are obtained. A similar method of feature extraction was successful for predicting miRNA genes (Yousef *et al.*, 2006).

2.4 NBmiRTar schema

The NBmiRTar tool is illustrated in Figure 2. The filters applied to the predictions before and after the data is processed through the NB classifier are listed to the right of the diagram. The main diagram shows that the NB classifier reprocesses the miRanda output using the features we have described to revise and rescore the prediction. The filtering layers are applied to further restrict and reduce the predictions. The parts of the schema are described below.

2.5 Naïve Bayes classifier

Naïve Bayes is a classification model obtained by applying a relatively simple method to a training dataset (Mitchell, 1997). A NB classifier calculates the probability that a given instance (example) belongs to a certain class. It makes the simplifying assumption that the features constituting the instance are conditionally independent given the class. Given an example X, described by its feature vector $(x_1,\ldots,x_n)$, we are looking for a class C that maximizes the likelihood: $P(X|C) = P(x_1,\ldots,x_n|C)$. The (naïve) assumption of conditional independence among the features, given the class, allows us to express this conditional probability $P(X|C)$ as a product of simpler probabilities: $P(X|C) = \prod_{i=1}^{n} P(x_i|C)$.

We used the Rainbow program (McCallum, 1996) to train the NB classifier. To combine the numeric features identified in the miRNA–target duplex with the sequence features ('*words*') in the target candidate sequence, a dictionary of all the unique '*words*' was generated and the frequency of each '*word*' in the sequence is used.

2.6 MiRanda and energy scores filters

As shown in Figure 2, the NB classifier is applied to the output of the miRanda program (John *et al.*, 2004). MiRanda associates to each prediction, a score that describes the maximal local complementarity alignments. For each single-residue-matched pair a specific score is given, for example, +5 for G:C and A:T pairs and +2 for G:U wobble pairs. The final miRanda score is computed as the sum of the single-residue-pair match scores over the duplex structure. The MFE of the duplex is determined using the Vienna package (Hofacker, 2003). We have required that only the miRanda predictions with MFE ≤12 kcal/mol will be used by the NB classifier for further classification and scoring. We have also chosen 90 as the miRanda score filter to be applied to the output of the NB classifier. For further discussion and information about the miRanda score and the MFE, see Hsu (2006).

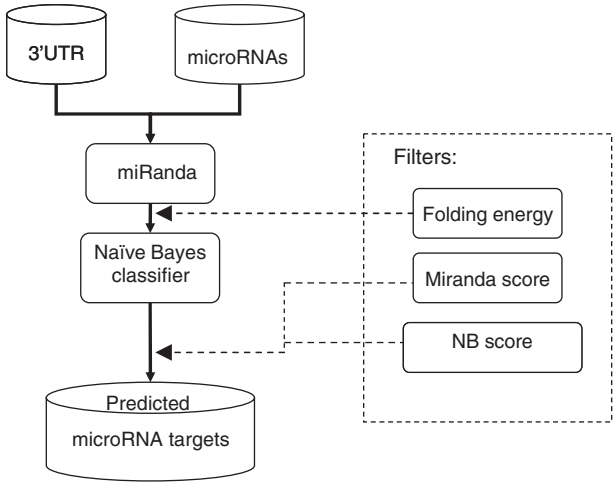

Fig. 2. The computational procedure for implementing NBmiRTar tool for miRNA target prediction.

2.7 Naïve Bayes score filter

We have found that our NB classifier has high accuracy (high sensitivity and specificity) at finding known miRNA:mRNA target genes. However, we were interested in further reducing the number of false positive predictions, since the data to be examined at this point could include very large numbers of examples. Even with a small percentage of false positives, tens of thousands of predictions could be generated, making it difficult to validate these predictions in the laboratory. Thus, further analysis was applied to determine the appropriate threshold for further eliminating false positive predictions. The NB classifier assigns a score to each miRNA: mRNA candidate and classifies it into one of the two predefined classes: the positive class (target) and the negative class (non-target). Figure 3a shows the distribution of the classifier's scores for the true positive (TP) and the false positive (FP) predictions. A threshold of 0.9 reduces the false predictions by ~64% while only losing 4.5% of the true miRNA target when applied to the output of the classifier. A stricter threshold might be chosen to further reduce the number of predictions.

3 RESULTS

3.1 Training and evaluation of performance

We trained the classifier by multiple rounds of 10-fold cross-validation. In each round, a randomly selected 90% of each class was used for training and the remaining samples tested to determine accuracy. This was repeated 100 times and the average fraction of true positive (sensitivity) and true negative (specificity) predictions were determined.

To estimate classification performance, we have evaluated the accuracy of NBmiRTar in identifying the 38 validated negative examples (NEG_0) and the artificial negative class

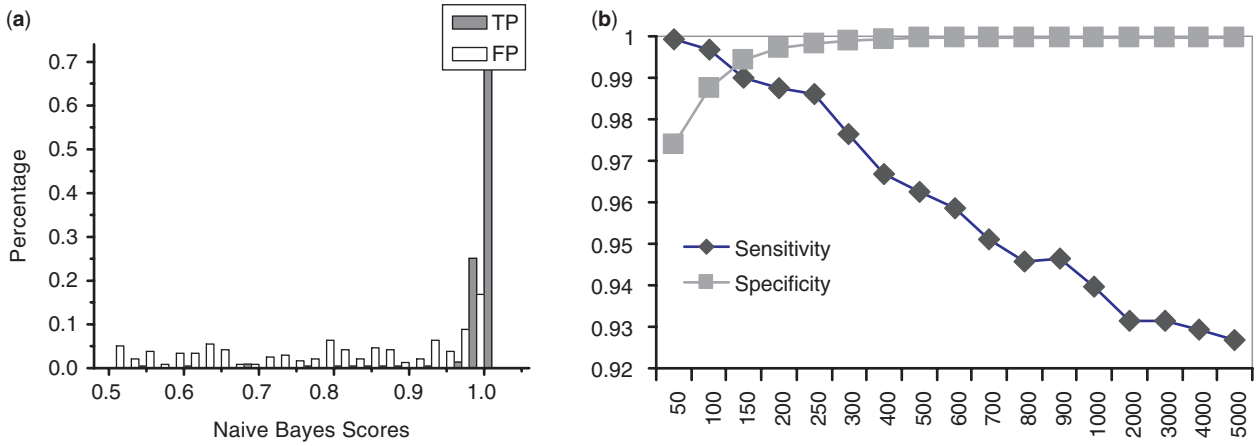

Fig. 3. (**a**) Distribution of Naive Bayes scores over the miRNA target class and the negative class. This figure shows the distribution of the NBSs for the true positive (fraction TP) and the false positive (fraction FP) rates. (**b**) Accuracy of prediction as a function of size of negative class. Sensitivity is the true positive (number called positive divided by number of positive examples and specificity is the true negatives (number called negative divided by number of negative examples).

(NEG_1) generated from miRanda. Our primary focus was to build a classifier with as high specificity as possible in target identification in order to reduce the false positive miRNA target predictions to levels that would be amenable to laboratory validation procedures.

3.2 Results for the artificial negative (NEG_1)

Results from cross-validation during training using NEG_0 are about 0.93 as sensitivity and 0.7 as specificity (see Supplementary Material). It is clear that the current set of 38 validated negative examples is not large enough to represent the negative class as we had anticipated, and therefore more negative data is required.

We then used the NB classifier with the 133 316 negative examples (NEG_1) generated as described in the Methods section. The NB classifier was trained multiple times. In each training epoch, a set of 200 known mRNA:miRNAs (90% of the positive data) was randomly selected from the 225 and used as positive examples. We varied the number of negative examples from the NEG_1 dataset across different sets of experiments, randomly choosing a set of 50, 100, 150, 200, 250, 300, 400, 500, 600, 700, 800, 900, 1000, 2000, 3000, 4000 or 5000 from the pool of 133, 316 negative examples and training on 90% of this set. The testing was performed with the remaining 10% from the positive class and the remaining NEG_1 examples from the selected set. This evaluation procedure was repeated 100 times. The results are shown in Figure 3b. For example, using 900 negative examples yields 0.99 as specificity and 0.94 as sensitivity.

3.3 Significant features

We analyzed the most significant features (Table 1) with highest average mutual information (MI, Shannon and Weaver, 1949) in order to better understand the miRNA:mRNA interactions. We used the 900 negative example training set as a compromise between sensitivity and specificity (Fig. 3b). Each class (positive or negative) has its own significant features that consist of the 57 duplex structure features plus the unique 'words' generated as sequence features: 11 987 from the 225 positive examples and 143 134 from the 900 negative examples. Since, the negative class was selected from the false positive target predictions from miRanda, the features that define the negative class share some similarities with the features that define the positive class. The most informative features distinguish between the two classes as differences in the mean values for each of the features. For instance the relative values for features number 1 and 2 suggest that the positive class includes more bulges of 1 nt length in the seed part than the negative class (mean of 0.35 versus 0.003), while based on features number 4, 14 and 16 it would be concluded that the number of bulges in the out-seeds of the negative class is more than for the positive class. Based on features 5, 8 and 10, the positive class has more asymmetric loops in its out-seed segment than the negative class. Based on features 3 and 27, the positive class has fewer symmetric loops in the out-seed than the negative class.

3.4 Results with human known targets

We compared the number of target predictions generated by miRanda, the NB classifier and the NB classifier with the NB-filter when tested for their ability to capture 13 verified human targets for the 10 known miRNAs. The 3′UTR human dataset was downloaded from miRanda along with the sequences for each of 10 miRNA to serve as the input to NBmiRTar tool. The results for each individual miRNA and the number of predictions generated by each tool are shown in Table 2. The record of recovery of the true target for each input miRNA is shown for each of the three tool analyses. The accuracy of the whole test is calculated based on the recovery of the specific correct targets.

The NBmiRTar tool has a reduction of 75% [1 – (834 082/3 331 410)] of miRanda predictions with a recovery rate of 77%

Table 1. The top 30 significant features obtained by MI

	Feature name	Positive Mean	Negative Mean
1	Number of bulges in seed	0.35	0.004
2	Number of bulges in seed with length 1	0.328	0.0035
3	Number of symmetric loops in out-seed with length 1	1.186	2.718
4	Number of bulges in out-seed	0.764	1.655
5	Number of asymmetric loops in out-seed with length greater than 7	0.15	0.0095
6	Acucca	0.0755	0
7	Acuc	0.128	0.0085
8	Number of asymmetric loops in seed	0.0711	0
9	Acucc	0.111	0.006
10	Number of asymmetric loops in out-seed	1.04	0.4835
11	Accuu	0.084	0.002
12	Gcuuu	0.057	0
13	Gcuuua	0.057	0
14	Number of bulges in out-seed with length 1	0.413	0.7955
15	Acuccau	0.053	0
16	Number of bulges in out-seed with length 2	0.115	0.401
17	Accu	0.088	0.0065
18	Uguga	0.048	0
19	Ugugau	0.048	0
20	Cucca	0.048	0
21	Accuuc	0.048	0
22	Gcuu	0.057	0.001
23	Cagg	0	0.106
24	Caggg	0	0.103
25	Acauucc	0.044	0
26	Cagggc	0	0.1025
27	Number of symmetric loops in out-seed with length 2	0.457	0.8665
28	Acau	0.048	0.0005
29	Acauu	0.048	0.0005
30	Accuucu	0.04	0

The mean value of each feature across the positive and negative class examples are listed in the columns at the left.

(10/13) of the confirmed targets and that the same recovery rate is obtained when the NB-filter is applied (threshold 0.9) but with a further reduction in the miRanda prediction to 81%. The output from the NB-filter of 620 757 predictions obtained from the 10 miRNAs gives an average of about two target sites per miRNA in each 3′UTR.

We also ran the same test with a single human miRNA, mir-15. In this case, miRanda produced 88 376 predictions that we subsequently reduced to 3479 predictions after applying the NB-filter and miRanda Score-filter. This is 4% of the original predictions from miRanda

4 DISCUSSION

Most of the existing tools for miRNA predict very large numbers of miRNA target predictions making biological validation very difficult. Although those that do use conserved sequence features produce smaller numbers of predictions, they have the disadvantage of being able to only predict targets sites with highly conserved sequences. We have described a machine-learning approach to miRNA target prediction that does not rely on conservation and is still able to significantly reduce the number of target predictions while retaining an acceptable sensitivity. The ability of NBmiRTar to maintain the levels of sensitivity we have demonstrated suggests: (1) The extracted features we have used have captured important features that define the miRNA:Target class requiring contributions from both the seed and the out-seed segments, (2) The negative class that we selected from the miRanda output for 3000 artificial non-miRNAs appears to accurately represent the negative class.

Although we have considerably reduced the number of target predictions while retaining sensitivity, there is room for significant improvement. One possible way to accomplish a further reduction in prediction numbers is to use the approach we have applied here in developing new tools that can be used in tandem with available tools such as miRanda, PicTar, etc. to rescore and or re-filter the predictions.

Choosing an appropriate negative class for training a classifier to recognize miRNA targets is necessarily arbitrary. We used the negative class described by SaeTrom *et al.* (2005) that consists of random nucleotides with frequencies not consistent with true miRNA. An alternative negative class could use 22mers selected at random from the human genome. When we compared targets predicted from such a class to a comparable number of human miRNA, target predictions, the results were similar to those in Figure 3a. The predictions of NBmiRTar were four times as many targets predicted for the real human miRNA as the random sequences (23 versus 5.8). Applying the NB-filter with cutoff 0.99, the ratio of predictions from real miRNAs to random sequences was 16:1.8. We conclude that the ability of NBmiRTar to identify targets is not strongly dependent on the exact definition of the negative target class since two different negative classes we have tried give similar results.

A recent study conducted by Yan *et al.* (2007) proposed a similar technique to the one we have reported with best-reported accuracy of 83% as compared to our accuracy of 94%. These results were generated using only 48 positive (human) and 16 negative examples and the specificity of the performance is not reported.

It is not well understood how miRNAs recognize and regulate their target genes. The extraction of features in our current study attempts to include the less obvious elements of the interaction between the miRNA and its target by considering every structural feature that may be considered in the formation of the duplex as well as the sequence features. It is interesting that the two most important features dictate imperfect matches in the seed sequence. This might allow flexibility for a single miRNA to have multiple targets.

In the present model, miRanda predictions serve as the input to NBmiRTar but certainly output predictions from other tools could also be used. We selected the miRanda output as input for the present studies because of its high success rate in identifying validated targets (Chen *et al.*, 2005;

Table 2. Predicted human miRNAs targets by miRanda and NBmiRTar

miRNA	Number of confirmed targets	MiRanda predictions	Recovery by miRanda	NBmiRTar	Recovery by NBmiRTar	NB-filter 0.9
1	1	401 592	1/1	87 843	0/1	60 108
2	2	64 984	2/2	34 380	2/2	27 239
3	2	321 312	2/2	80 632	2/2	60 967
4	2	49 556	2/2	24 090	1/2	19 013
5	1	563 477	1/1	259 423	1/1	202 339
6	1	294 255	1/1	84 153	1	61 725
7	1	596 411	1/1	92 337	1	62 118
8	1	381 933	1/1	54 636	0	34 138
9	1	329 770	1/1	42 736	1	45 447
10	1	328 120	1/1	73 852	1	47 663
Sum	13	3 331 410	13	834 082	10	620 757

The last column represents the number of predictions when the NB-filter with 0.9 threshold is used.

Enright *et al*., 2003; John *et al*., 2004; Leaman *et al*., 2005). In addition, the availability of the source code allowed us to embed it into our computational procedure. We expect as more validated target predictions emerge and the number of positive examples increases, the accuracy of our predictions will also increase. As suggested by Brennecke *et al*. (2005), the class of the interactions of miRNA:mRNA could be assigned to different families based on different rules that might be applied for the interactions.

We have launched a web-server (v1.0 Beta) (available at http://wotan.wistar.upenn.edu/NBmiRTar/**)** that allows the user to obtain his prediction by inputting the miRNA(s) and 3′UTR sequences with option of applying any of the three filters (miRanda score filter, folding free energy and Naive Bayes score filter) we have described in the Methods section.

ACKNOWLEDGEMENTS

This project is funded in part by U01 CA85060 and the Pennsylvania Department of Health (PA DOH Commonwealth Universal Research Enhancement Program), and Tobacco Settlement grants ME01-740 and SAP 4100020718 (L.C.S.), NSF RCN 0090286 (M.K.S). We would like to thank Shere Billouin for preparing the manuscript.

Conflict of Interest: none declared.

REFERENCES

Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281.

Brennecke,J. *et al*. (2005) Principles of microRNA-target recognition. *PLoS Biol*., **3**, e85.

Chen,P.Y. *et al*. (2005) The developmental miRNA profiles of zebrafish as determined by small RNA cloning. *Genes Dev*., **19**, 1288–1293.

Enright,A. *et al*. (2003) MicroRNA targets in Drosophila. *Genome Biol*., **5**, R1.

Grimson,A. *et al*. (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol. Cell*, **27**, 91–105.

Hofacker,I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res*., **31**, 3429–3431.

Hsu,P.W. (2006) miRNAMAP: genomic maps of microRNA genes and their target genes in mammalian genomes. *Nucleic Acids Res*., **34**, D135.

Hsu,P.W.C. *et al*. (2006) miRNAMap: genomic maps of microRNA genes and their target genes in mammalian genomes. *Nucleic Acids Res*., **34**, D135–139.

John,B. *et al*. (2004) Human microRNA targets. *PLoS Biol*., **2**, e363.

Lai,E. (2004) Predicting and validating microRNA targets. *Genome Biol*., **5**, 115.

Leaman,D. *et al*. (2005) Antisense-mediated depletion reveals essential and specific functions of microRNAs in Drosophila development. *Cell*, **121**, 1097.

Lewis,B.P. *et al*. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, **120**, 15.

McCallum,A.K. (1996) Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering text retrieval, classification and clustering. http://www.cs.cmu.edu/mccallum/bow

Mitchell,T. (1997) *Machine Learning*. McGraw Hill, New York.

Rajewsky,N. (2006) MicroRNA target predictions in animals. *Nat. Genet*., **38**, S8–S13.

SaeTrom,O.L.A. *et al*. (2005) Weighted sequence motifs as an improved seeding step in microRNA target prediction algorithms. *RNA*, **11**, 995–1003.

Sethupathy,P. *et al*. (2006a) TarBase: a comprehensive database of experimentally supported animal microRNA targets. *RNA*, **12**, 192–197.

Sethupathy,P. *et al*. (2006b) A guide through present computational approaches for the identification of mammalian microRNA targets. *Nat. Methods*, **3**, 881.

Shannon,C. and Weaver,E. (1949) *The Mathematical Theory of Communication*. University of Illinois Press, Chicago and Urbana.

Sung-Kyu,K. *et al*. (2005) A kernel method for microRNA target prediction using sensible data and position-based features. *Computational Intelligence in Bioinformatics and Computational Biology*. In *Proceedings of the 2005 IEEE Symposium 1*.

Thadani,R. and Tammi,M. (2006) MicroTar: predicting microRNA targets from RNA duplexes. *BMC Bioinformatics*, **7**, S20.

Vella,M.C. *et al*. (2004) Architecture of a validated microRNA:target interaction. *Chem. Biol*., **11**, 1619.

Yan,X. *et al*. (2007) Improving the prediction of human microRNA target genes by using ensemble algorithm. *FEBS Lett*., **581**, 1587.

Yousef,M. *et al*. (2006) Combining multi-species genomic data for microRNA identification using a Naive Bayes classifier. *Bioinformatics*, **22**, 1325–1334.