2008-01-01

# Animal microRNA Target Prediction By Incorporating Diverse Sequence-Specific Determinants

Yun Zheng and Weixiong Zhang

More recent evidence has shown that access of animal microRNAs (miRNAs) to their complementary sites in target mRNAs is determined by more sequence-specific determinants than the seed regions in the 5' end of miRNAs. Although these factors have been shown to be related to the repressive power of miRNAs and used, in separate programs, to predict the efficacy of miRNA complementary sites, it remains unclear whether these factors can help to improve miRNA target prediction. We develop a new miRNA target prediction algorithm, called Hitsensor, by incorporating more sequence-specific features that determine complementarities between miRNAs and their targets, in... **Read complete abstract on page 2.**

# Animal microRNA Target Prediction By Incorporating Diverse Sequence-Specific Determinants

Yun Zheng and Weixiong Zhang

Complete Abstract:

More recent evidence has shown that access of animal microRNAs (miRNAs) to their complementary sites in target mRNAs is determined by more sequence-specific determinants than the seed regions in the 5' end of miRNAs. Although these factors have been shown to be related to the repressive power of miRNAs and used, in separate programs, to predict the efficacy of miRNA complementary sites, it remains unclear whether these factors can help to improve miRNA target prediction. We develop a new miRNA target prediction algorithm, called Hitsensor, by incorporating more sequence-specific features that determine complementarities between miRNAs and their targets, in addition to the canonical seed regions in the 5' ends of miRNAs. We evaluate the performance of our algorithm on 720 known animal miRNA:target pairs in four species, Homo sapiens, Mus musculus, Drosophila melanogaster and Caenorhabditis elegans. Our prediction results show that Hitsensor outperforms five popular existing algorithms, indicating that our unique scheme to quantify the determinants of complementary sites is effective in improving the performance of a miRNA target prediction algorithm. Unlike most existing algorithms, our method does not use conservation information and can find many unconserved miRNA:target pairs.

2008-20

# Animal microRNA Target Prediction By Incorporating Diverse Sequence-Specific Determinants

Authors: Yun Zheng and Weixiong Zhang

Corresponding Author: Weixiong Zhang

Abstract: More recent evidence has shown that access of animal microRNAs (miRNAs) to their complementary sites in target mRNAs is determined by more sequence-specific determinants than the seed regions in the 5' end of miRNAs. Although these factors have been shown to be related to the repressive power of miRNAs and used, in separate programs, to predict the efficacy of miRNA complementary sites, it remains unclear whether these factors can help to improve miRNA target prediction. We develop a new miRNA target prediction algorithm, called Hitsensor, by incorporating more sequence-specific features that determine complementarities between miRNAs and their targets, in addition to the canonical seed regions in the 5' ends of miRNAs. We evaluate the performance of our algorithm on 720 known animal miRNA:target pairs in four species, Homo sapiens, Mus musculus, Drosophila melanogaster and Caenorhabditis elegans. Our prediction results show that Hitsensor outperforms five popular existing algorithms, indicating that our unique scheme to quantify the determinants of complementary sites is effective in improving the performance of a miRNA target prediction algorithm. Unlike most existing algorithms, our method does not use conservation information and can find many unconserved miRNA:target pairs.

Type of Report: Other

# Animal microRNA Target Prediction By Incorporating Diverse Sequence-Specific Determinants

Yun Zheng [a], Weixiong Zhang [a,b,*],

[a] *Department of Computer Science and Engineering, Washington University in St. Louis*

*1 Brookings Drive, St. Louis, MO 63130, USA*

[b] *Department of Genetics, Washington University School of Medicine*

*1 Brookings Drive, St. Louis, MO 63130, USA*

## Abstract

**More recent evidence has shown that access of animal microRNAs (miRNAs) to their complementary sites in target mRNAs is determined by more sequence-specific determinants than the seed regions in the 5' end of miRNAs. Although these factors have been shown to be related to the repressive power of miRNAs and used, in separate programs, to predict the efficacy of miRNA complementary sites, it remains unclear whether these factors can help to improve miRNA target prediction. We develop a new miRNA target prediction algorithm, called *Hitsensor*, by incorporating more sequence-specific features that determine complementarities between miRNAs and their targets, in addition to the canonical seed regions in the 5' ends of miRNAs. We evaluate the performance of our algorithm on 720 known animal miRNA:target pairs in four species, *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster* and *Caenorhabditis elegans*. Our prediction results show that Hitsensor outperforms five popular existing algorithms, indicating that our unique scheme to quantify the determinants of complementary sites is effective in improving the performance of a miRNA target prediction algorithm. Unlike most existing algorithms, our method does not use conservation information and can find many unconserved miRNA:target pairs.**

*Key words:* microRNA, sequence analysis, microRNA target prediction, sequence specific

determinants

running title: Animal microRNA Target Prediction

## Introduction

MicroRNAs are non-coding RNAs that regulate the expression of protein-coding genes at post-transcriptional level [1]. They function by base-paring to their target mRNAs, subsequently leading to translational repression [1,2], mRNA cleavage [3–5] or miRNA-induced degradation [6–8]. Due to the complexity in experimental validation of miRNA targets, several computational miRNA target prediction methods have been developed, including TargetScan [9] (later updated to TargetScanS [10]), Miranda [11,12], PicTar [13], methods in [14,15], RNAHybrid [16], rna22 [17], PITA [18] for animals, and methods in [19–21], miRU [22] for plants. Many of these methods were reviewed in [23].

Most predicted and reported complementary sites of animal miRNAs are located in the 3' untranslated region (3' UTR) of target mRNAs [9–16]. The imperfect complementarity between miRNAs and their targets in animals makes target prediction much harder than in plants. Many existing methods for animals [9–16] extensively make use of the seed region, which is from the 2 to 8 nucleotides from the 5' end of a mature miRNA, in their prediction.

However, a substantial number of miRNA:target pairs do not have good seed regions. Brennecke *et al.* [24] found that there are mainly two categories of miRNA complementary sites, 5' dominant sites and 3' compensatory sites. The first category constitutes most animal miRNA complementary sites [13,10,24]. For this category, 7mer and 8mer 5' seed matches are sufficient to function with 3' paring below a random noise level [24]. On the other hand, 3'

* Corresponding author.
*Email addresses:* zhengy@cse.wustl.edu (Yun Zheng), zhang@cse.wustl.edu (Weixiong Zhang).
*URL:* http://www.cse.wustl.edu/∼zhang/ (Weixiong Zhang).

compensatory sites have insufficient 5' seed matches and require strong 3' pairing in order to be functional [24]. One example is the *let-7* binding sites in *lin-41* [25]. Thus, a strong preference to seed region by the existing methods may miss 3' compensatory sites. For example, TargetScan cannot find 3' compensatory sites [23].

Most existing methods [9–16] also use evolutional conservation, which is effective for finding conserved targets. On the other hand, conservation information does not help to identify species specific targets.

More recent evidence indicated that there exist other determining factors besides the seed regions in miRNA complementary sites. As well documented, most miRNAs start with uridine; correspondingly, their binding sites end with adenosine. Even for some miRNAs that do not begin with uridine, the position complementary to the first nucleotide of miRNA is preferentially adenosine [26]. Lewis *et al.* [10] found that seed complementary sites are often flanked by adenosines. Nielsen *et al.* [26] noticed the preference of adenosine or uridine for the site complementary to the ninth nucleotide from the 5' end of a miRNA. They also found that an increased AU content in the 3' of the seed region is correlated with an increased mRNA down-regulation effect. Jing *et al.* [27] and Grimson *et al.* [28] further noticed that many effective sites preferentially reside within regions that are locally AU rich. As suggested by [24], 3' compensatory sites can function because there are extensive parings in those regions. Moreover, Grimson *et al.* [28] quantified a compensatory pairing region of 12-17 nucleotides from the 5' end of a miRNA. In addition, Grimson *et al.* [28] also found that closely spaced sites in the 3' UTR of a target mRNA often synergistically promote the repression of the target, and effective complementary sites often locate after the 15-th nucleotide from the stop codon of the mRNA and in the first and last quarters of the 3' UTR. All these results indicated that local AU-content, 12-17 nt pairing, closely-paced sites, site positions, along with seed pairing, are important determinants to enhance miRNA-induced repression.

Motivated by the evidence mentioned above, we hope that incorporating these determinants can further improve miRNA target prediction. One of our aim is to investigate whether these determinants are useful to improve the performance of a target prediction algorithm. In particular, we propose a novel miRNA target prediction algorithm, called Hitsensor, to exploit and combine various sequence determinants. In the Hitsensor algorithm, we introduce a set of rules to quantify the contributions from the seed region, 12-17nt region, local AU-content, close sites and site positions. Although some existing algorithms, such as Miranda [11], also give additional rewards to seed region, our approach uses a new rewarding scheme to emphasize the continuously matched seed. Briefly, the Hitsensor algorithm does not use conservation information in its prediction. It starts from a sequence alignment with the Smith-Waterman algorithm [29] for miRNA and its target mRNA, calculates the scores of the 5 determinants for each alignment site, and then adds these individual determinant contributions to the alignment score to get the total score of a miRNA complementary site. Finally, sites with total scores larger than a pre-specified threshold are outputted.

Grimson *et al.* [28] proposed a context score to predict the site efficacy with these determinants. However, the goal of our method and context score is different. The model proposed by Grimson *et al.* [28] was used to predict the efficacy of a miRNA complementary site in repressing the target, especially at mRNA levels. Their regression model requires mRNA expression information. In contrast, our method focuses on predicting true miRNA:target relations, meanwhile, it attempts to reduce false miRNA:target relations by making use of additional information from these determinants. In another study, Wang and Naqa [30] employed mRNA expression profiles to select important features for prediction of miRNA:target pairs. Our method is easier to use than those in [28,30] because our method uses no information other than sequences of miRNAs and targets. Furthermore, the methods in [28,30] could miss some targets if they are down-regulated by miRNAs at protein level.

We adopt two methods to evaluate the performance of the Hitsensor algorithm, the receiver operating characteristic (ROC) curve and the signal-to-noise ratio (S2N). We use a data set of 96 verified functional and 83 non-functional miRNA:target pairs of *Drosophila melanogaster* to quantify the contributions of individual determinants. The Hitsensor algorithm reaches an area under the ROC curve (AUC) of 0.794 and an S2N of 7.62, which are the highest among all compared algorithms, including PITA [18], PicTar [13] and Miranda [11], on this data set.

We then select 541 verified functional miRNA:target pairs across four species, *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster* and *Caenorhabditis elegans*, to validate the performance of the Hitsensor algorithm. Again, the Hitsensor algorithm produces the largest number of correct predictions, 293, among all algorithms compared. In comparison, the existing algorithms, PITA with and without flanking sequences [18], TargetScanS [10], PicTar [13] and Miranda [11] have, respectively, 231, 262, 188, 138 and 123 correct predictions on these selected data sets.

## Materials and Methods

### Data Sets

As summarized in Table 1, we extrated 720 *experimentally verified* miRNA:target pairs for four species from [18], the TarBase [31] and [32]. Kertesz *et al.* [18] summarized a data sets with 190 *Drosophila melanogaster* miRNA:target pairs, 102 functional and 88 non-functional. Because the target genes of 6 and 5 pairs from 102 functional and 88 non-functional sets, respectively, have no 3' UTR in the FlyBase (http://flybase.bio.indiana.edu/), we only use the remaining 96 functional and 83 non-funcational pairs, i.e., dme96P and dme83N in Table 1, which are used as training data set to find optimal quantifications of the 5 determinants.

In addition, the TarBase contains another 16 functional miRNA:target pairs of *Drosophila* not in dme96P, which form dme16P in Table 1. The cel, hsa and mmu data sets are for worm

Table 1

The experimentally verified miRNA:target pairs used in training and testing.

| | No. | Functionality | Reference |
|---|---|---|---|
| *training* | | | |
| dme96P | 96 | functional | [18] |
| dme83N | 83 | non-functional | [18] |
| *subtotal* | 179 | | |
| *testing* | | | |
| dme16P | 16 | functional | TarBase[31] |
| cel | 14 | functional | TarBase[31] |
| hsa | 440 | functional | TarBase[31] |
| mmu | 49 | functional | TarBase[31] |
| unc-hsa | 22 | functional | [32] |
| *subtotal* | 541 | | |
| Total | 720 | | |

*Caenorhabditis elegans*, human *Homo sapiens* and mouse *Mus musculus* and downloaded from the TarBase. After removing some miRNA:target pairs of worm, human and mouse in the TarBase because either their miRNA or target sequences are not available, we have 14, 440 and 49 pairs in cel, hsa and mmu data sets. The unc-hsa data set in Table 1 consists of 22 of the 23 unconserved human miRNA:target pairs in [32], because we did not find 3' UTR for 1 of the 23 pairs in [32]. The detailed list of these 720 miRNA:target pairs are given in Supplementary Table S1.

The sequences of miRNAs of these 720 pairs were downloaded from the miRBase (release 10) [33]. The sequences of mRNA targets were from NCBI RefSeq database (ftp://ftp.ncbi.nih.gov/refseq/H_sapiens/H_sapiens/RNA for hsa, ftp://ftp.ncbi.nih.gov/refseq/M_musculus/Contigs/RNA for mmu) and NCBI CoreNucleotide database and the FlyBase for dme96P, dme84N, dme16P and cel.

**Algorithms Compared**

We will compare Hitsensor with 5 benchmark methods, i.e., PITA with (PITAf) and without (PITAn) flanking sequences [18], TargetScanS [10], PicTar [13] and Miranda [11]. The features used by the algorithms compared are summarized in Table 2 and discussed in detail in the next section. All these algorithms make use of the seed region, although in different ways. Hitsensor and Miranda give additional rewards to Watson-Crick pairs in seed regions with different schemes (to be discussed in next section). PITAn, PITAf and TargetScanS directly find perfectly seed regions [18,10]. PicTar prefers perfect seed matches but also allows imperfect seed matches [13,32]. Hitsensor is the only algorithm that uses the 12-17nt region. Hitsensor, PITAf and TargetScanS employ the flanking regions of seeds [32,18]. Hitsensor uses close site determinants and optionally uses site position determinants. PITA, i.e. both PITAf and PITAn, is the only algorithm that considers the free energy of 3' UTR before miRNA binding ($\Delta G_{open}$) by employing the energy gain after and before a miRNA binds its target, i.e., $\Delta G_{duplex} - \Delta G_{open}$ [18]. Miranda, TargetScanS and PicTar compute the free energy of miRNA:target duplex, $\Delta G_{duplex}$, with different methods [32]. Finally, conservation information is used by TargetScanS, PicTar, and optionally by Miranda [32]; therefore, these algorithms are conservation based.

The results of TargetScanS were downloaded from the TargetScan website (http://www.target

Table 2

The features used by the 6algorithms compared. 'opt.' means optional. $\Delta G_{open}$ is energy cost of unpairing the 3' UTR of target. $\Delta G_{duplex}$ is the free energy of miRNA:target duplex.

| | HITS | MIRA | PITAn | PITAf | TSS | PicTar |
|---|---|---|---|---|---|---|
| Seed | √ | √ | √ | √ | √ | √ |
| 12-17nt | √ | | | | | |
| seed flank | √ | | | √ | √ | |
| close site | √ | | | | | |
| site position | opt. | | | | | |
| $\Delta G_{open}$ | | | √ | √ | | |
| $\Delta G_{duplex}$ | | √ | √ | √ | √ | √ |
| conservation | | opt. | | | √ | √ |

scan.org/), for both conserved and nonconserved miRNA families. The results of PicTar were downloaded from annotation databases of dm2, hg17, mm7 and ce2 of the UCSC genome browser (http://hgdownload.cse.ucsc.edu/). The results of PITA were downloaded from (http://genie.weizmann.ac.il/pubs/mir07/mir07_data.html) for targets catalog with and without 3nt upstream and 15nt downstream flanking sequences. We used a local version of the Miranda algorithm (version 1.9), available at the Miranda website (http://www.microrna.org/miranda_new.html), to obtain its results.

Because some verified complementary sites, such as miR-431 complementary sites on RTL1/Rtl1 [34], are located in coding regions of targets, we applied Hitsensor and Miranda (without conservation information) separately to 3' UTRs and coding sequences (CDS) to examine miRNA complementary sites in CDS.

## Sequence-Specific Determinants

We use the example in Figure 1 to show how to use a set of parameters, called reward bases, to quantify the five sequence-specific determinants of miRNA complementary sites, i.e., seed region, 12-17nt region, local AU-content, close sites and site positions. Different values can be given to the reward bases to adjust the contributions of different determinants. In our implementation, we have assigned optimal default values, 8, 4, 44, 12 and 0 to reward bases of the seed region, 12-17nt region, local AU-content, close sites and site position determinant, respectively. We will discuss how to obtain these values of the reward bases in Results.

### *Seed Determinant*

Continuously matched seed regions are critical for repressing target mRNA or inducing target mRNA degradation [24,10,28]. To capture the importance of continuous matches in seed regions, we design a new score scheme that rewards functional, continuously matched seed regions with larger scores than uncontinuously matched counterparts, which often occur by chance. Formally, we give a reward to the seed region based on Equation (1),

$$SeedScore = R \times \sum_{i=1}^{8}(\alpha_i - \beta_i \times 2), \tag{1}$$

where $R$ is the reward base of seed determinant, $\alpha_i$ is the number of continuous Watson-Crick matches from the 5' end of a miRNA and is reset to 0 when a mismatch or a G:U pair occurs, and $\beta_i$ is the number of continuous mismatches or G:U pairs from the 5' end of a miRNA and is renewed to 0 when a Watson-Crick pair appears. $\alpha_i$ and $\beta_i$ in Equation (1) serve as a reward to continuous matches and a penalty to mismatches and G:U pairs, respectively.

In addition, because 8-mer perfect seeds are more effective to repress targets than 7-mer ones [28,26], we also adopt the following empirical rules: if there is a continuously paired 8-mer seed, an additional reward of $3R$ will be given; if there exists a continuously paired

7-mer seed with a G:U pair or mismatch at the first nucleotide, an additional reward of $2.5R$ and $2R$ will be added; if there is at least 7 continuously paired nucleotides and the ninth nucleotide is a Watson-Crick pair, an additional reward of $R$ will be given. Finally, if there are totally more than 2 mismatches or G:U pairs, we give an additional penalty of $R \times (n_m + n_{G:U})$, where $n_m$ and $n_{G:U}$ are the number of mismatches and the number of G:U pairs from the first to the eighth nucleotide of the miRNA, respectively.

For example, in site 2 (S2) of Figure 1, hsa-miR-101 is continuously paired to 3' UTR of EZH2 from the first to ninth nucleotide. Thus, this site receives a seed score of 160, i.e., $(1 + 2 + \ldots + 8) \times 4 = 144$ based on Equation (1), plus 12 for a continuous 8-mer pair and 4 for a paired ninth nucleotide. As another example, if there was a mismatch at the fifth nucleotide of S2, then $\alpha_5$ to $\alpha_8$ would become 0 to 3 (see Supplemental Figure S1(a)). Therefore, the seed score would be $(1 + \ldots + 4) \times 4 - 8 + (1 + 2 + 3) \times 4 = 56$, which is 104 less than a continuously matched 8-mer seed, where -8 is the penalty to a mismatch at position 4. In contrast, if the reward is determined by the number of Watson-Crick pairs, as used by Miranda [11] (see Figure S1(b)), the difference between the two cases is only $4 \times 8 - (4 \times 7 - 4) = 8$.

*12-17nt Region Determinant*

The continuously matched 12-17nt region is important and compensatory to imperfect seed region [24], and enhances miRNA binding [28]. Therefore, similar to the *SeedScore* in Equation (1), we reward the 12-17nt region with Equation (2).

$$TwelveSeventeenScore = U \times \sum_{j=1}^{6} (\alpha_j - \beta_j \times 2), \tag{2}$$

where $U$ is the reward base for the 12-17nt region determinant, $\alpha_j$ and $\beta_j$ have the same values as $\alpha_i$ and $\beta_i$ in Equation (1) except starting from 12nt of a miRNA. Similar to the seed region, we also give an additional penalty of $U \times (n_m + n_{G:U})$ where $n_m$ and $n_{G:U}$ are the number of mismatches and the number of G:U pairs from 12 to 17nt, respectively, if there are more than 2 mismatches or G:U pairs in the 12-17nt region.

A complementary site with sufficient matches in the seed region can function with little support from the pairing from the 3' end of the miRNA [24]. Therefore, if there exists at least one basic 6-mer (2-7nt) seed match, we will not give a penalty to the 12-17nt region, i.e., $penalty = 0$. On the contrary, if a complementary site does not contain a 6-mer seed match and 12 to 17nt form a 6-mer continuous Watson-Crick match, we will give an additional reward of $6U$ to 12-17nt determinant, and set the $SeedScore$ in Equation (1) to 0 if it is negative.

For the example in Figure 1, site S2 has an 8-mer matched seed, thus *penalty* to 12-17nt region is zero. There are totally four Watson-Crick paired nucleotides with 2 of them continuously matched, thus the total reward is 4 + 4 + 8 + 4 = 20.

*Local AU-Content Determinant*

We calculate the score of local AU-content with Equation (3).

$$AUScore = (\sum_{i=1}^{30} \frac{1}{i} \times IsAU_{up}(i) + \sum_{j=1}^{30} \frac{1}{j} \times IsAU_{down}(j)) \times B, \tag{3}$$

where $IsAU_{up}(i)$, a variable indicating whether a position on mRNA beginning from the opposite of 9nt of miRNA is A or U(T), will be 1 if the nucleotide at position $i$ is A or U(T), or 0 otherwise. Because local AU preference normally appears with continuous seed match [10,28], we allocate different reward base $B$ and $0.25B$ to sites with and without perfectly matched 6-mer seeds (2-7nt) to further differentiate functional sites with perfect seeds to those with imperfect seeds normally due to random chance. Because AU preference immediately beside seed region is important and decreases fast when the distance from the seed increases [10,28], the weights of these A and U, $1/i$ and $1/j$, are decreasing when the distance between them and seed, $i$ and $j$, increases. As shown in Figure 1, the weights of local A and U around seed are reflected by the height of the bars above the corresponding nucleotides. Thus, the sum operations in Equation (3) capture the effects of A and U in the flanking region of the seed. For the example in Figure 1, because the site has a matched 8-mer seed, $B$ is 8, and the score of local AU-content is 51.9, following Equation (3).

*Close Sites Determinant*

If a miRNA has more than one complementary sites on a target, these sites may synergistically repress the target, when they have an intersite distance between 19 to 34nt [28]. Thus, we first find all sites with at least a 6-mer matched seed or a total score from other determinants greater than that of an 8-mer matched seed plus 8 additional paired nucleotides, and then calculate the distances between these sites. If the distance between two sites is within 19 to 34nt, we give a close site score of $D$. In Figure 1, sites S2 and S1 have a close site score of $D = 12$, because S2 and S1 are 32nt apart and S1 has a 7-mer matched seed.

*Position Determinant*

We give a position score of $Q$ if a complementary site is located in the first or last quarter of a 3' UTR, and an additional reward of $0.5Q$ if the 3' UTR is longer than 1300nt. This is because complementary sites in the first and last quarters of 3' UTRs longer than 1300nt are more effective [28]. However, if a complementary site is located within the first 15nt of the first quarter of a 3' UTR, we will not give reward to it, because such a site is weaker than those in other regions of the 3' UTR [28]. The position determinant is only applicable to the miRNA complementary sites in 3' UTRs of target mRNAs. For the example in Figure 1, no position score is given to site S2, which is located in the second quarter of the 3' UTR.

## The Hitsensor Algorithm

Hitsensor first uses a modified Smith-Waterman (SW) algorithm [29] to find regions with sufficient matches between miRNAs and their targets. Instead of performing alignments with matched nucleotides, *e.g.*, A-A and C-C, Hitsensor finds complementary nucleotides, *i.e.*, G-C, A-U and G-U "wobble" pairs that have rewards of +6, +4 and +2, respectively, in alignment. The affine gap penalty, i.e., the penalty increasing linearly with the length of gap after initial gap opening penalty, is used for gap opening (-8) and gap extension (-4). The algorithm gives a penalty of -3 to known mismatch nucleotides and a penalty of -1 to mismatches to unspecified nucleotides (i.e., "N") in mRNAs. The algorithm will first recursively search for miRNA complementary sites on the whole target mRNA sequence. If a site has a positive alignment score, the algorithm will keep it for further analysis.

After obtaining a list of sites, Hitsensor will continue to evaluate the sequence-specific determinants for all sites and set the scores for the determinants. The final score of a complementary site is then the sum of the scores of all determinants and alignment score from the Smith-Waterman algorithm. For example, the final score for S2 in Figure 1 is 299.9 which is the sum of the scores of different determinants and the alignment score. If the final score of a given pair is greater than a user-specified threshold, Hitsensor will output this site. Finally, the max score of all sites for a given miRNA:target pair is used as the representative score of the pair to reflect the best possible binding of the pair. This information is useful because even though many miRNA:target pairs carry a single complementary site [32], a large number of them have multiple complementary sites. And when multiple sites exist, the most accessible site should be more likely to be bound than the other sites since a site with a larger final score should be more accessible than one with a smaller final score.

In some extreme cases, we found that some miRNA:pairs with perfect seed matches, such as dme-miR-79 vs bap, have optimal SW alignments with imperfectly matched seed regions.

Consequently, these sites will have low final scores based on our score scheme. To correct this drawback due to application of the SW alignment, Hitsensor will check whether the target has regions that perfectly match to 2-7nt of the miRNA if it fails to find complementary sites after evaluating all determinants. If such regions are found, Hitsensor will cut the flanking sequences, upstream 29nt and downstream 1nt, of these regions, re-evaluate the determinants and output these sites if they satisfy the specified threshold.

We have implemented the Hitsensor algorithm with the Java programming language. The software package and documents are available at the supplementary website of the journal.

**Evaluation Methods**

*The receiver operating characteristic (ROC) curve*

The ROC curve shows the sensitivity vs false positive ratios (fpr, i.e., 1 - specificity) under different score thresholds. The area under the curve (AUC) measures the ability of the algorithm to correctly classify functional and non-functional miRNA:target pairs. On an ROC curve, the point nearest to the upper left corner provides the optimal algorithm setting, where the algorithm reaches the optimal balance between sensitivity and specificity (i.e., 1 - fpr).

*Signal-To-Noise Ratio*

The signal-to-noise (S2N) ratio is often used to evaluate the performance of target prediction algorithms [13,9]. We use the scores of verified functional miRNA:target pairs as the scores of positive samples and the scores of verified non-functional miRNAs as values of negative samples to generate the signal-to-noise ratio.

Table 3

The number of positive predictions of the compared algorithms. The *subtotal* row lists the total number of correct predictions on all testing data sets. The last row shows the threshold scores to obtain the results. Algorithm names are the same as those in Figure 2. The best prediction performances, i.e., the largest numbers for data sets with functional pairs and the smallest number for dme83N with non-functional pairs, are shown in bold face.

| | 3UTR | | | | | | CDS | | 3UTR+CDS | |
|---|---|---|---|---|---|---|---|---|---|---|
| | HITS | MIRA | PITAn | PITAf | TSS | PicTar | HITS | MIRA | HITS | MIRA |
| dme96P | **72** | 40 | 69 | 69 | 31 | 61 | 13 | 21 | 77 | 48 |
| dme83N | 17 | 25 | 22 | 22 | **4** | 19 | 15 | 25 | 32 | 38 |
| dme16P | **11** | 1 | 5 | 2 | 2 | 10 | 0 | 0 | 11 | 1 |
| cel | 6 | 4 | 8 | **9** | 4 | 4 | 1 | 2 | 7 | 6 |
| hsa | **237** | 96 | 226 | 202 | 151 | 117 | 50 | 54 | 268 | 132 |
| mmu | 30 | 21 | 17 | 11 | **31** | 7 | 14 | 16 | 39 | 30 |
| unc-hsa | **9** | 1 | 6 | 7 | 0[a] | 0[a] | 3 | 5 | 10 | 6 |
| *subtotal* | **293** | 123 | 262 | 231 | 188 | 138 | 68 | 77 | 335 | 175 |
| *threshold* | *472* | *139* | *-6.8* | *-2.2* | NA | NA | *472* | *139* | *472* | *139* |

[a] results from [32].

# Results

## Improved Performance by Incorporating Diverse Sequence-Specific Determinants

### *Examining Effects of Different Determinants*

To find optimal quantifications of determinants, we exclusively changed the reward base for one of the five determinants, i.e., seed region ($R$), 12-17nt region ($U$), local AU-content ($B$), close sites ($D$) and site position ($Q$), from 0 to 20, and obtained the ROC curves of the training data set (dme96P+dme88N). The results are listed in Supplementary Figures S2(a) to (e), respectively. The AUC and S2N against various values of the five reward bases are given in Figures S3(a) and (b). As shown in Figures S2(a) and (b), the algorithm had very different performance, and reached best AUC and S2N when $R = 8$ and $U = 0$ in Figures S3(a) and (b). But after reviewing Figure S2(b), we found that the algorithm had the optimal tradeoff between sensitivity and specificity when $U = 4$. The increasing reward base of local AU-content, $B$, had a beneficial effect on the AUC and S2N of the algorithm, although less significantly than $R$ and $U$ (Figures S3(a) and (b)). After testing various $B$ values, we found that AUC reached the maximal value when $B$ was around 40 (Figure S3(c)). When $B = 44$, the Hitsensor algorithm had its best tradeoff between sensitivity and specificity (Figure S2(f)). Various reward bases of close site determinant, $D$, had little effect on the performance of the algorithm (Figures S2(d) and S3(a),(b)). We also found that increasing $Q$, the reward base of position determinant, could decrease AUC and S2N values (Figures S2(e) and S3(a),(b)). Therefore, we applied $R = 8$, $U = 4$, $B = 44$, $D = 12$ and $Q = 0$ to both the training and testing data sets. The obtained ROC curve, AUC and S2N of Hitsensor on training data sets, as well as those from other algorithms, are given in Figure 2, while the number of positive predictions, i.e., samples predicted as functional miRNA:target pairs, for all data sets at the optimal settings of the compared algorithms are listed in Table 3. The optimal thresholds of the compared algorithms are obtained with their ROC curves, as discussed in Methods. The complete lists of Hitsensor predictions when using 3' UTRs and CDS are given in Table S2 and S3, respectively.

### *miRNA Complementary Sites in 3' UTRs and CDS*

Although most verified animal miRNA complementary sites are located in 3' UTRs of targets [9–16], some mammalian coding genes also have miRNA complementary sites in their

CDS [34,35]. As shown in Table 3, both Hitsensor and Miranda predicted more miRNA complementary sites in 3' UTRs than in CDS. For instance, Hitsensor predicted 237 sites in 3' UTRs while only 50 sites in CDS. It is important to note that we found that some miRNAs can have complementary sites in both 3' UTRs and CDS. We found that among the 50 miRNAs that have complementary sites in CDS of human genes (on hsa data set), 19 also have complementary sites in 3' UTRs (Table S4). The regulatory effects of these 50 miRNAs on CDS can be well explained by the microarray gene expression profiles of the targets (see Table S4) in [7]. This suggests that these miRNA sites in CDS might play roles in the regulation of the targets. Furthermore, many miRNA complementary sites in CDS of RTL1/Rtl1, such as those of miR-136 and miR-341, have been directly verified with 5' RACE [34].

We also find that the miRNA complementary sites of two miRNA:target pairs, hsa-miR-125b:DDX19B/mmu-miR-125b-5p:Ddx19b and miR-431:RTL1/Rtl1, in CDS are conserved between human and mouse (see Table S3). It is interesting to point out that miR-125b:DDX19B was listed as an unconserved pair in [32] because there were no conserved complementary sites in 3' UTRs. However, our findings suggest that the regulatory relation of miR-125b and DDX19B is conserved between human and mouse through miR-125b complementary sites in CDS of DDX19B. As to be shown in Figure 3(a), the conservation of miR-431 complementary sites in CDS of RTL1/Rtl1 have been verified in [34]. In addition, a recent study also demonstrated that miR-148 targets coding region of human DNMT3b, which is conserved in mammals [35]

These findings suggest that the 3' UTRs of animals have evolved to accommodate most miRNA complementary sites, meanwhile coding regions still maintain a small portion of miRNA complementary sites.

## Comparisons With the Existing Methods

Hitsensor achieved the best overall performance in all algorithms compared on both training and testing data sets; the results are shown in Figure 2 and Table 3. On the training data sets, Hitsensor reached a sensitivity of 75% (72/96) and a specificity of 79.5% (1-17/83), which are 3% and 6% higher than those of PITA, repectively. As shown in Table 3, PITA had the best performance among all existing algorithms. This was also shown by Figure 2(a), where the closest point of all ROC curves to the up-left corner is on the ROC curve of Hitsensor. We attribute this to the 12-17nt determinant used by Hitsensor. As discussed early, Hitsensor could reach optimal tradeoff between sensitivity and specificity when the reward base of 12-17nt region, $U$, was 4 (see Figure S2(b)). Meanwhile, other algorithms compared did not use information from 12-17nt region, as shown in Table 2. If taking CDS of targets into account, Hitsensor could have a sensitivity of 80.2% and specificity of 70% on the training data sets (see Table 3).

On the testing sets, Hitsensor had an overall sensitivity of 54.2% (293/541), again the highest among all compared algorithms. When compared with the best sensitivity of the existing algorithms (from PITAn), Hitsensor had an improvement of 5.8%. Hitsensor found another 42 pairs, 7.8%, on all test data sets if both 3' UTRs and CDS were considered, as shown in Table 3. On individual data sets, Hitsensor performed the best in 4 out of the 7 data sets, shown in bold fonts in Table 3. On dme83N, Hitsensor produced 17 false positive predictions, which was only larger than that of TargetScanS. However, the sensitivities of TargetScanS were much lower than Hitsensor, except for the mmu data set.

Hitsensor reached an AUC value of 0.794 that is lightly higher than those of PITA, with and without flanking sequences, and much higher than that of Miranda (Figure 2(b)). As reported in [18], PITAf had an AUC of 0.79 on 190 samples, which were higher than those from method in [15], PicTar [13] and Miranda [11] (see Figure 2(b)). PITA had similar performance on our data sets with 179 samples and 190 samples originally reported by [18], which suggests that it is meaningful to compare our results with those methods in [18] (starred methods in Figure 2(b)). Again, Hitsensor had higher AUC value than those methods in [18] (see Figure 2(b)). Miranda performed better on the 190 samples in [18] than on our training data with 179 samples, which might be resulted from different versions of Miranda and/or different methods to calculate miRNA:target scores. Hitsensor also had higher S2N values when compared with PITA and Miranda, as in Figure 2(c). Wang and Naqa [30] also used the AUC to evaluate their method. Their models reached AUC values of 0.79 and 0.77 with and without the conservation information, respectively [30]. Hitsensor obtained a slightly better AUC value than that of Wang and Naqa's method [30] even though Hitsensor did not used mRNA expression information.

As shown in Table 3, PITA performed well by using free energy of target 3' UTRs and miRNA:target duplex (Table 2). In contrast, Hitsensor achieved an overall better performance than PITA without employing the thermodynamical information used by PITA, which is computationally expensive to compute. Because all algorithms used seed information, we attribute this improvement to two unique features that Hitsensor used, the 12-17nt region and the local AU-content (Table 2). As discussed early, 12-17nt region is effective to improve the tradeoff between sensitivity and specificity (Figure S2(b)). The reward to local AU-content determinant improved the AUC of Hitsensor (Figure S3(c)). In addition, the score of local AU-content is computationally cheaper to compute than the free energy of 3' UTR and miRNA:target duplex used by PITA.

We also compared the overlapped predictions of different algorithms for the dme96P and hsa data sets, and the results were shown in Table 4. For a given algorithm, the total number of overlapped predictions showed capability of this algorithm to find predictions from other algorithms compared. We thus listed the total number of overlapped predictions in the last column (for hsa data set) and last row (for dme96P data set). For instance, Hitsensor

Table 4

The overlapped predictions in 3' UTRs of different algorithms on the dme96P (below upper-left to lower-right diagonal) and hsa data sets (above upper-left to lower-right diagonal). The value in each cell means the overlapped predictions of the two algorithms from the row and column of the cell. The last row and column list the total number of commonly predicted pairs with other algorithms for the algorithm in this column and row on dme96P and hsa data sets, respectively. Algorithm names are the same as those in Figure 2.

| hsa<br>dme96P | HITS | MIRA | PITAn | PITAf | TSS | PicT | total |
|---|---|---|---|---|---|---|---|
| HITS | | 95 | 171 | 147 | 129 | 101 | 643 |
| MIRA | 19 | | 75 | 62 | 42 | 35 | 309 |
| PITAn | 55 | 37 | | 167 | 120 | 93 | 626 |
| PITAf | 57 | 37 | 62 | | 100 | 74 | 550 |
| TSS | 28 | 12 | 24 | 26 | | 109 | 500 |
| PicT | 56 | 30 | 47 | 49 | 28 | | 412 |
| total | 215 | 135 | 225 | 231 | 118 | 210 | |

respectively had 643 and 215 total common predictions for hsa and dme96P with the other 5 algorithms compared. As shown in Table 4, Hitsensor made much larger number of common predictions than Miranda, PITAf, TargetScanS and Pictar for the hsa data set. For the dme96P data set, Hitsensor, PITAn, PITAf and PicTar made comparable number of total common predictions, and the total common predictions of Miranda and TargetScanS were much smaller than the other four algorithms compared. These indicate that Hitsensor could successfully find major parts of correct positive predictions produced by other algorithms.

For example, Hitsensor found 171 out of the 226 (75.7%) predictions of the hsa data set from PITAn.

### Synergistic Complementary Sites

It has been observed that miRNAs can act synergistically in post-transcriptional regulation [36,28]. This has also been observed in our results, listed in Supplemental Table S5. We found that 12 miRNA:target pairs, which span over 11 miRNAs and 10 targets, in Table S5, have putative synergistic complementary sites of the same miRNA in the selected data sets.

We analyzed the complementary sites on RTL1 (of *Homo sapiens*)/Rtl1 (of *Mus musculus*) in Figure 3, where Hitsensor predicted a total of 6 new synergistic complementary sites (red and green sites), in addition to the 3 blue sites reported in [34]. The Hitsensor algorithm predicted two conserved synergistic miR-431 complementary sits on RTL1/Rtl1, as shown in Figure 3(a). Davis *et al.* [34] reported that 11 out of 12 clones correspond to the blue site. This suggests that at least some of the clones might be produced by the newly found red site in Figure 3(a). Figure 3(b) shows that, in addition to the site reported in [34], Hitsensor predicted two more complementary sites of mmu-miR-434-5p. Davis *et al.* [34] reported that only 5 out of 23 clones were shown to be the cleavage product of mmu-miR-434-5p at the position pointed by the arrow in Figure 3(c), and other clones were supposed to be random Rtl1 degradation products [34]. However, the predicted synergistic mmu-miR-434-5p complementary sites in Figure 3(c) suggest that the remaining clones are very likely to be cleavage products from the newly predicted red mmu-miR-434-5p sites. Furthermore, Hitsensor also predicted another pair of synergistic mmu-miR-434-5p sites, i.e., the green sites in Figure 3(c), which are 2nt downstream of the blue site. They only have 2 mismatched nucleotides and have an intersite distance of 26nt. They might also produce some of the remaining clones detected by Davis *et al.* [34].

Figure 3 shows that there exist two levels of cooperative miRNA-induced repression on Rtl1. First, at least three miRNAs, mmu-miR-431, mmu-miR-434-3p and mmu-miR-434-5p cooperatively repress Rtl1 by binding to their respective complementary sites on Rtl1. Furthermore, Davis *et al.* [34] reported that mmu-miR-127, mmu-miR-136, mmu-miR-433-3p and mmu-miR-433-5p are also involved in repressing Rtl1. Second, several copies of mmu-miR-431, mmu-miR-434-3p and mmu-miR-434-5p may bind to their respective synergistic complementary sites and collaboratively repress Rtl1.

## Discussion

We have studied the effects of different sequence-specific determinants on predicting miRNA target complementary sites and developed a new miRNA target prediction algorithm which we called Hitsensor. The Hitsensor algorithm has a superior performance over five benchmark miRNA target prediction methods that we compared on an extensive collection of experimentally validated data sets. We attribute the performance of Hitsensor to three major aspects.

First, we used various determinants in our methods, including the new scheme to quantify the conventional seed region used by the other algorithms. As discussed in Methods, our quantification method to seed region, as well as 12-17nt region, has given much higher rewards to continuously matched seed regions than uncontinuously matched counterparts, which might be produced by random chance. This has helped to distinguish functional miRNA:target pairs to randomly paired non-functional miRNA and mRNA. Another important factor contributing to the success of Hitsensor is local AU content around seed region. Functional miRNA complementary sites are often located in AU-rich regions in 3' UTRs of targets [28,27]. Appropriate reward to local AU content has helped to improve the AUC and optimal sensitivity vs specificity of Hitsensor, as shown in Figures S3(c) and S2(f).

Second, Hitsensor could predict species specific miRNA:target relations. In comparison, TargetScanS, PicTar and Miranda used conservation information in their prediction, which let them miss some species specific miRNA:target pairs, as shown by their predictions on unc-hsa data set in Table 3.

Finally, as shown in Table 3, we found that 13.5% (13/96) and 12.6% (68/541) functional pairs of training and testing data sets, respectively, have predicted complementary sites in CDS of targets. Some of the predicted complementary sites in CDS had been verified in [34]. As shown in Results, Hitsensor had better sensitivities when taking CDS into account. These results suggest that CDS of targets contain substantial percentage of miRNA complementary sites and should not be ignored when performing target prediction for animal miRNAs, although most miRNA complementary sites are located in 3' UTRs, as shown in Table 3 and in literature [9–16]. Hitsensor made 15 positive predictions for the dme83N data set when using CDS. Further experiments are necessary to verify whether these sites function or not, because only 3' UTRs of the targets were tested with reporter gene assays (see [18] and references therein).

As shown in Figure 2 and Table 3, PITA also performed well on the selected data sets. These suggest that difference between free energy of miRNA:target duplex and energe cost of unpairing the 3' UTR of target used by PITA is useful information in predicting animal miRNA targets. Hitsensor does not use the folding energy of miRNA:target duplex and 3' UTRs. However, there is a relationship between local AU-content and energy cost of unpairing the 3' UTR, because high local AU-content around seeds reduces the energy costs to make seeds accessible for miRNAs loaded in the RNA-induced silencing complex (RISC, see [1]). These imply that the seed and its flanking region are two critical factors that affect the performance of target prediction algorithms. Hitsensor performed better than PITA except for the cel data set (see Table 3) because it used additional information from 12-17nt determinant.

The Hitsensor algorithm is able to automatically predict putative synergistic complementary sites by incorporating the close site determinant.

## Acknowledgments

This research was supported in part by NSF grant IIS-0535257 and a grant from the Alzheimer's Association.

## References

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

[2] Esquela-Kerscher, A. and Slack, F. J. (2006) Oncomirs - microRNAs with a role in cancer. *Nature Reviews Cancer,* **6**, 259–269.

[3] Llave, C., Xie, Z., Kasschau, K. D., and Carrington, J. C. (2002) Cleavage of Scarecrow-like mRNA Targets Directed by a Class of Arabidopsis miRNA. *Science,* **297**(5589), 2053–2056.

[4] Tang, G., Reinhart, B. J., Bartel, D. P., and Zamore, P. D. (2003) A biochemical framework for RNA silencing in plants. *Genes Dev.,* **17**(1), 49–63.

[5] Yekta, S., Shih, I.-h., and Bartel, D. P. (2004) MicroRNA-Directed Cleavage of HOXB8 mRNA. *Science,* **304**(5670), 594–596.

[6] Bagga, S., Bracht, J., Hunter, S., Massirer, K., Holtz, J., Eachus, R., and Pasquinelli, A. E. (2005) Regulation by let-7 and lin-4 mirnas results in target mRNA degradation. *Cell,* **122**, 553–563.

[7] Lim, L. P., Lau, N. C., Garrett-Engele, P., Grimson, A., Schelter, J. M., Castle, J., Bartel, D. P., Linsley, P. S., and Johnson, J. M. (2005) Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature,* **433**, 769–773.

[8] Wu, L., Fan, J., and Belasco, J. G. (2006) MicroRNAs direct rapid deadenylation of mRNA. *PNAS,* **103**(11), 4034–4039.

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

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

[11] Enright, A., John, B., Gaul, U., Tuschl, T., Sander, C., , and Marks, D. (2003) microRNA target detection. *Genome Biology,* **5**, R1.

[12] John, B., Enright, A. J., Aravin, A., Tuschl, T., Sander, C., and Marks, D. S. (2004) Human microRNA targets. *PLoS Biol,* **2**(11).

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

[14] Rajewsky, N. and Socci, N. (2004) Computational identification of microRNA targets. *Genome Biology,* **5**(2), P5.

[15] Stark, A., Brennecke, J., Bushati, N., Russell, R. B., and Cohen, S. M. (2005) Animal microRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution. *Cell,* **5**(2), 1133-1146.

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

[17] Miranda, K. C., Huynh, T., Tay, Y., Ang, Y.-S., Tam, W.-L., Thomson, A. M., Lim, B., and Rigoutsos, I. (2006) A pattern-based method for the identification of microRNA binding sites and their corresponding heteroduplexes. *Cell,* **126**(6), 1203–1217.

[18] Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U., and Segal, E. (2007) The role of site accessibility in microRNA target recognition. *Nat Genet,* **39**(10), 1278–1284.

[19] Rhoades, M. W., Reinhart, B. J., Lim, L. P., Burge, C. B., Bartel, B., and Bartel, D. P. (2002) Prediction of plant microRNA targets. *Cell,* **110**(4), 513–520.

[20] Wang, X. J., Reyes, J. L., Chua, N. H., and Gaasterland, T. (2004) Prediction and identification of arabidopsis thaliana microRNAs and their mRNA targets. *Genome Biol,* **5**(9). R65.

[21] Jones-Rhoades, M. W. and Bartel, D. P. (2004) Computational identification of plant microRNAs and their targets, including a stress-induced mirna. *Mol Cell,* **14**(6), 787–799.

[22] Zhang, Y. (2005) miRU: an automated plant miRNA target prediction server. *Nucl. Acids Res.,* **33**(suppl_2), W701–704.

[23] Rajewsky, N. (2006) microRNA target predictions in animals. *Nat Genet,* **38 Suppl 1**(6s), S8–S13.

[24] Brennecke, J., Stark, A., Russell, R. B., and Cohen, S. M. (2005) Principles of microRNA-target recognition. *PLoS Biol,* **3**(3). e85.

[25] Vella, M. C., Choi, E.-Y., Lin, S.-Y., Reinert, K., and Slack, F. J. (2004) The C. elegans microRNA let-7 binds to imperfect let-7 complementary sites from the lin-41 3'UTR. *Genes Dev.,* **18**(2), 132–137.

[26] Nielsen, C. B., Shomron, N., Sandberg, R., Hornstein, E., Kitzman, J., and Burge, C. B. (2007) Determinants of targeting by endogenous and exogenous microRNAs and siRNAs. *RNA,* **13**(11), 1894–1910.

[27] Jing, Q., Huang, S., Guth, S., Zarubin, T., Motoyama, A., Chen, J., Padova, F. D., Lin, S., Gram, H., and Han, J. (2005) Involvement of microRNA in AU-rich element-mediated mRNA instability. *Cell,* **120**(5), 623–634.

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

[29] Smith, T. F. and Waterman, M. S. (1981) Identification of common molecular subsequences. *Journal of Molecular Biology,* **147**, 195–197.

[30] Wang, X. and El Naqa, I. M. (2008) Prediction of both conserved and nonconserved microRNA targets in animals. *Bioinformatics,* **24**(3), 325–332.

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

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

[33] Griffiths-Jones, S., Grocock, R. J., vanDongen, S., Bateman, A., and Enright, A. J. (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucl. Acids Res.,* **34**(suppl_1), D140–144.

[34] Davis, E., Caiment, F., Tordoir, X., Cavaillé, J., Ferguson-Smith, A., Cockett, N., Georges, M., and Charlier, C. (2005) RNAi-mediated allelic trans-interaction at the imprinted rtl1/peg11 locus. *Curr Biol,* **15**(8), 743–749.

[35] Duursma, A. M., Kedde, M., Schrier, M., leSage, C., and Agami, R. (2008) miR-148 targets human DNMT3b protein coding region. *RNA,* **14**(5), 872–877.

[36] Doench, J. G. and Sharp, P. A. (2004) Specificity of microRNA target selection in translational repression. *Genes Dev,* **18**(5), 504–511.

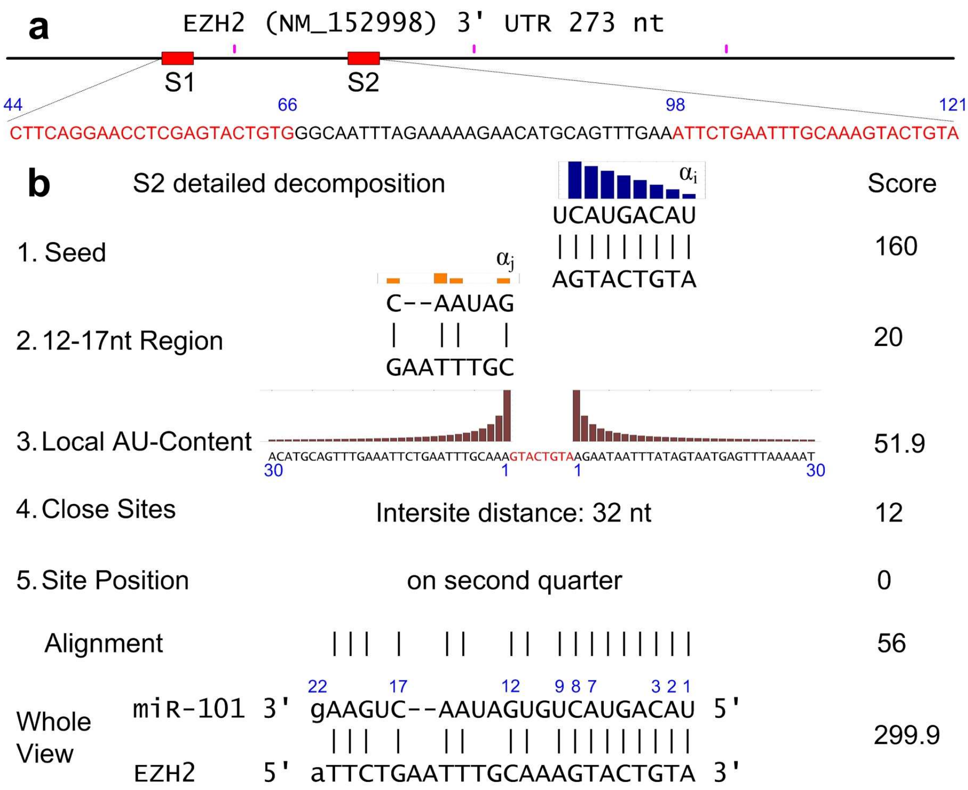


Fig. 1. A schematic view of sequence-specific determinants that affect hsa-miR-101 binding to the 3' UTR of EZH2 (NM_000609).

(a) The two predicted binding sites of hsa-miR-101, red boxes, in 3' UTR of EZH2 that is represented by the black solid line. The quarter points of the 3' UTR are indicated by the pink points above the 3' UTR. (b) Detailed decomposition of different determinants for site S2. With the values indicated with the bars, $\alpha_i$ and $\alpha_j$ above the seed and 12-17nt region are the numbers of continuous matches at that position that are defined in Equation (1) and Equation (2), respectively. For the local AU-content determinant, the weights of the position are represented by the heights of the bar above the nucleotides. The reward base for seed ($R$), 12-17nt region ($U$), local AU-content ($B$), close sites ($D$) and site position ($Q$) determinant are 4, 4, 8, 12 and 12 respectively.

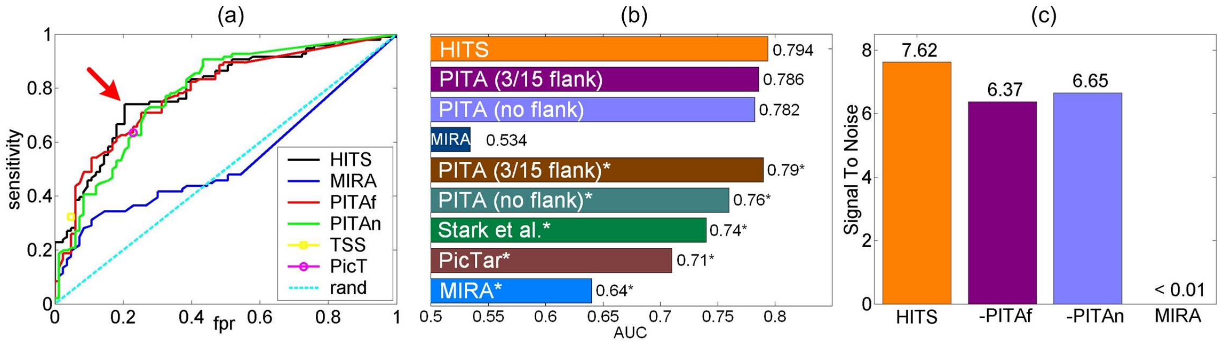


Fig. 2. The comparisons of different algorithms.

(a) The ROC curve, (b) AUC and (c) S2N of the compared algorithms for the training data set (dme96P+dme83N). HITS, MIRA, PITAf, PITAn, TSS and PicT stand for the Hitsensor, Miranda, PITA with flanking sequences, PITA without flanking sequences, TargetScanS, and PicTar algorithm, respectively. In part (a), the results obtained by a random scoring of the targets are shown by a dashed line. The point pointed by the red arrow was the best tradeoff between sensitivity and specificity reached by the Hitsensor algorithm. *In part (b), these results are obtained from [18] on 190 pairs.

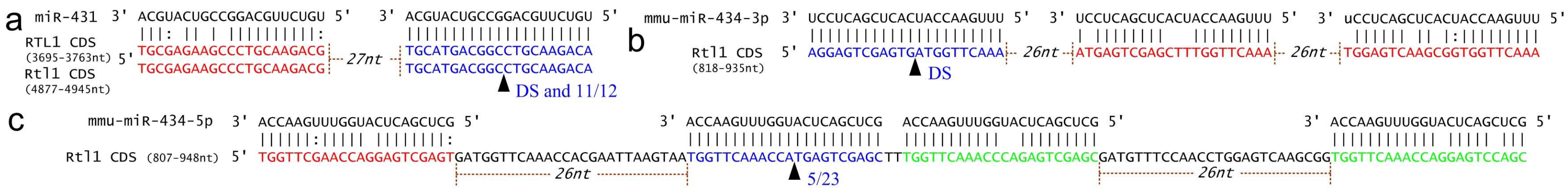
a
miR-431 3' ACGUACUGCCGGACGUUCUGU 5'
3' ACGUACUGCCGGACGUUCUGU 5'
RTL1 CDS (3695-3763nt) 5' TGCGAGAAGCCCTGCAAGACG 27nt TGCATGACGGCCTGCAAGACA
Rtl1 CDS (4877-4945nt) TGCGAGAAGCCCTGCAAGACG TGCATGACGGCCTGCAAGACA
DS and 11/12
b
mmu-miR-434-3p 3' UCCUCAGCUCACUACCAAGUUU 5'
3' UCCUCAGCUCACUACCAAGUUU 5'
3' uCCUCAGCUCACUACCAAGUUU 5'
Rtl1 CDS (818-935nt) 5' AGGAGTCGAGTGATGGTTCAAA 26nt ATGAGTCGAGCTTTGGTTCAAA 26nt TGGAGTCAAGCGGTGGTTCAAA
DS
c
mmu-miR-434-5p 3' ACCAAGUUUGGUACUCAGCUCG 5'
3' ACCAAGUUUGGUACUCAGCUCG
ACCAAGUUUGGUACUCAGCUCG 5'
3' ACCAAGUUUGGUACUCAGCUCG 5'
Rtl1 CDS (807-948nt) 5' TGGTTCGAACCAGGAGTCGAGTGATGGTTCAAACCACGAATTAAGTAATGGTTCAAACCATGAGTCGAGCTTTGGTTCAAACCCAGAGTCGAGCGATGTTTCCAACCTGGAGTCAAGCGGTGGTTCAAACCAGGAGTCCAGC
26nt
5/23
26nt

Fig. 3. Predicted putative synergistic miRNA binding sites on RTL1 (of *Homo sapiens*)/Rtl1 (of *Mus musculus*).

Blue sites were reported in [34]. Red and green sites are putative synergistic complementary sites predicted by the Hitsensor algorithm. Black arrows indicate cleavage sites reported in [34], which were identified by RLM 5' RACE either by direct sequencing of the PCR products (DS) or by sequencing of individual cloned products. The numbers indicate the fraction of clones that identify the blue cleavage site [34]. (a) conserved miR-331 complementary sites on RTL1/Rtl1. (b) mmu-miR-434-3p complementary sites on Rtl1. (c) mmu-miR-434-5p complementary sites on Rtl1.